Available online at www.sciencedirect.com

ScienceDirect

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

Comparing 15D Valuation Studies in Norway and Finland—Challenges When Combining Information from Several Valuation Tasks

Yvonne Anne Michel, Dipl-Psych[1],*, *Liv Ariane Augestad, MD, PhD*[1], *Kim Rand, PhD*[1,2]

[1]*Department of Health Management and Health Economics, Medical Faculty, Institute of Health and Society, University of Oslo, Oslo, Norway;* [2]*Health Services Research Centre, Akershus University Hospital, Lørenskog, Norway*

ABSTRACT

Background: The 15D is a generic preference-based health-related quality-of-life instrument developed in Finland. Values for the 15D instrument are estimated by combining responses to three distinct valuation tasks. The impact of how these tasks are combined is relatively unexplored. **Objectives:** To compare 15D valuation studies conducted in Norway and Finland in terms of scores assigned in the valuation tasks and resulting value algorithms, and to discuss the contributions of each task and the algorithm estimation procedure to observed differences. **Methods:** Norwegian and Finnish scores from the three valuation tasks were compared using independent samples t tests and Lin concordance correlation coefficients. Covariance between tasks was assessed using Pearson product-moment correlations. Norwegian and Finnish value algorithms were compared using concordance correlation coefficients, total ranges, and ranges for individual dimensions. Observed differences were assessed using minimal important difference. **Results:** Mean scores in the main valuation task were strikingly similar between the two countries, whereas the final value algorithms were less similar. The largest differences between Norway and Finland were observed for depression, vision, and mental function. **Conclusions:** 15D algorithms are a product of combining scores from three valuation tasks by use of methods involving multiplication. This procedure used to combine scores from the three tasks by multiplication serves to amplify variance from each task. From relatively similar responses in Norway and Finland, diverging value algorithms are created. We propose to simplify the 15D algorithm estimation procedure by using only one of the valuation tasks.

Keywords: 15D, health-related quality of life, value algorithm, visual analogue scale.

Introduction

The 15D is a generic preference-based instrument used to measure health-related quality of life for estimating quality-adjusted life-years in health economic analyses [1]. The 15D was developed in Finland in the late 1970s. Values for 15D health states were derived in general population valuation studies using a set of valuation tasks based on the visual analogue scale (VAS) [2–4]. The 15D has been translated into 30 languages, including Norwegian, and more than 400 articles have been published using the instrument [5,6], 140 of which were in the last 5 years. The 15D is featured alongside the EuroQol five-dimensional questionnaire (EQ-5D-5L), the Assessment of Quality of Life (AQoL), the health utilities index (HUI), and the six-dimensional health state short form (SF-6D) in textbook presentations of health-related quality of life and instruments for measuring quality-adjusted life-years [7,8]. Because the 15D descriptive system covers many dimensions of health, it is often used in studies comparing such instruments. As such, it was recently part of a large multi-instrument comparison survey comparing five multi-attribute utility instruments [9].

Preferences for health states are assumed to vary between cultures. To capture differences in health state preferences between countries, country-specific algorithms are recommended by national guidelines [10–12]. Country-specific 15D value algorithms have been developed in Finland and Denmark [4,13]. The relevance and legitimacy of country-specific algorithms depend on their ability to adequately reflect the health state preferences of particular populations. Culture-dependent differences in health state preferences are still openly debated. Earlier research explores how country-specific differences in wealth, income, religion, health expenditure, and cultural factors such as power distance and individualism explain differences in preferences [14]. It is also possible that health state preferences change over time.

Respondents' health state preferences are not the only driver of differences in value algorithms. Valuation studies include

* *Address correspondence to:* Yvonne Anne Michel, Department of Health Management and Health Economics, Institute of Health and Society, University of Oslo, P.O. Box 1089, Blindern, Oslo 0318, Norway.

E-mail: y.a.michel@medisin.uio.no.

http://dx.doi.org/10.1016/j.jval.2017.09.018

choices about which methods to use and how to use them. Each valuation method raises questions about how to present the task, which visual aids to use, and which health states to value. There are also considerable differences in how case exclusion is handled [15]. The mode of administration can influence the data, and translation procedures could be a potential source of methodological variation [16]. Norman [17] highlights that "[t]he uncertain element in interpreting [different algorithms] is to identify whether the differences in models are a result of genuine differences in national attitudes toward ill health or whether they are the product of different study designs."

To which extent differences in health state preferences are driven by cultural or methodological variation remains unknown. Although there is a growing body of literature describing how methodological choices influence time trade-off–derived values for the EuroQol five-dimensional questionnaire [7,14,16,18], less is known about values derived for other instruments and other valuation methods. 15D algorithm values are derived by combining information from three VAS-based valuation tasks. Little is known about how country-specific 15D algorithms compare or about how the different valuation tasks contribute to the final algorithm values. Before being able to meaningfully interpret differences observed in 15D algorithm values, a better understanding of how the 15D valuation procedure shapes these values is necessary.

The aim of this study was to compare the results from 15D valuation studies conducted in Norway and Finland. Specifically, we compare the scores assigned in the valuation tasks, the value algorithms derived using the original valuation procedure, and discuss the contributions of each task and the algorithm estimation procedure to the observed differences in algorithm values.

Methods

15D Descriptive System

The 15D descriptive system consists of 15 dimensions, covering physical, mental, and social aspects of health [19]. Each dimension has five response options, with the first level corresponding to full functionality and the remaining levels describing declining levels of functionality.

15D Valuation System

The 15D allows the description of $5^{15} \approx 3.1 \times 10^{10}$ health states. 15D values are calculated using a predefined *value algorithm*, which is generated to reflect the preferences of the target population. The generation of a 15D value algorithm consists of two elements: the valuation tasks and the *value algorithm estimation procedure*. Because of the large number of dimensions and levels, the Finnish 15D value algorithm was derived using assumptions from the multi-attribute utility theory [20].

The Norwegian valuation study was based on the three valuation tasks developed by Sintonen [4]: 1) In the top task, respondents are asked to compare the top levels for all 15 dimensions, using a VAS anchored in "most important" (= 100) and "least important" (= 0, top task; see Appendix 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018); 2) In the bottom task, the bottom levels for all 15 dimensions, regarding the lowest levels of functioning, are rated on a VAS ranging from "best imaginable health state" (= 100) to "worst imaginable health state" (= 0, bottom task); 3) In the within-dimension task, the respondents are asked to place the five levels of one dimension, plus the state of "being dead," on a VAS anchored in "best imaginable health state" and "worst imaginable health state" (see Appendix 2 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018). We refer to the average scores derived from these tasks as top task scores, bottom task scores, and within-dimension scores, respectively. For brevity, L1, L2, L3, L4, and L5 are used to refer to levels 1 through 5 in the within-dimension task.

15D Algorithm Estimation Procedure

We use the term *algorithm estimation procedure* to refer to the steps taken to estimate an algorithm on the basis of scores from the valuation tasks averaged across all included respondents. Unless otherwise explicitly stated, all scores mentioned in this article are such averages. Briefly, the algorithm estimation procedure assigns each of the 15 dimensions a slot of the scale between "full health" (1) and "not being alive" (0), which reflects its relative importance. The levels of each dimension are assigned values within the respective slot. The dimensions are additive; summed up, they represent the full range of the 15D value algorithm. The valuation tasks were designed to provide input to the following value function described by Sintonen [4]:

$$V_H = \sum_j I_j(x_j) w_j(x_j),$$

where V_H is the social value of health state H and $I_j\ (x_j)$ is a set of positive constants for the jth dimension, representing the relative importance of the dimension at its various levels, constrained such that $\sum_j I_j = 1$ for any level. $w_j(x_j)$ is a numerical function of the jth dimension, representing the relative value of various levels of the dimension, such that the top level = 1 and being dead = 0.

The function was inspired by the multi-attribute utility theory [20] and builds on the idea of a two-stage valuation process in which levels within dimensions are valued in one task and the relative importance of the dimensions is determined separately. Nevertheless, the function developed by Sintonen assumes that the importance assigned to dimensions could vary by level.

We applied the algorithm estimation procedure developed by Sintonen [4]. An overview is presented here, and a more in-depth numerical example is given in Appendix 3 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018. The first step generates importance weights for each dimension on the basis of the top task. Averages for each dimension are calculated across respondents and are divided by the sum of all 15 such averages. The result is a set of 15 values (one for each dimension) that sum up to 1. The same procedure is used to generate importance weights from the bottom task (see Table 1 in Appendix 3 in Supplemental Materials for Norwegian top and bottom task scores and importance weights). The following steps are taken for each dimension separately: 1) Within-dimension scores are rescaled such that L1 is anchored in 1 and "being dead" is anchored in 0. These values are reserved for later. 2) L1 to L5 are rescaled, now such that L1 equals the top importance weight for the corresponding dimension and L5 equals the bottom importance weight for the corresponding dimension. 3) The results of step 1 are multiplied with the corresponding results from step 2. The resulting 15D value algorithm consists of 60 values, each referring to one of the five response options of the 15 dimensions. Algorithm values in this article are presented to indicate disutility; a positive value indicates a value loss associated with health problems and negative values indicate value gains.

Samples

The Finnish valuation study that was conducted in 1992 sampled 2500 members of the Finnish general population and is described in detail elsewhere [4]. The Finnish data collection differed from the Norwegian data collection in that there was no Web survey, and a single reminder was used. In addition to the five levels of functioning and "being dead," the Finnish within-dimension task included "being unconscious," which was dropped in the Norwegian study.

The Norwegian 15D data were collected in 2010, consisting of one postal and one Web sample, both surveyed by the market research firm TNS Gallup. For the postal study, a random sample of 5000 addresses was drawn from the Norwegian National Population Registry. Participants received the survey as letters with prepaid response envelopes, without follow-up reminders. The Web sample was recruited from an online panel of preregistered individuals maintained by TNS Gallup (the TNS-Gallup Panel). Individuals in the panel were recruited by email in waves so as to achieve a sample of more than 1000 individuals resembling the Norwegian general population in terms of age, sex, educational level, and geographic distribution.

Participants in the Norwegian study were randomized to four groups (A, B, C, and D). All answered a set of demographic questions and the 15D descriptive system. Three groups were administered the top task (A, C, and D), three groups the bottom task (B, C, and D), and two groups the within-dimension task (A and B). In the postal sample, respondents were asked to place the health states on a vertical VAS. In the within-dimension task, the first level was fixed to 100 (see Appendix 2 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018). The software used to conduct Web surveys by the market research company did not support vertical VAS presentation. Therefore, the within-dimension task was given a horizontal presentation. Also, because of technical limitations, L1 was not fixed at 100 in the Web task, meaning that respondents could assign a value lower than 100 for not having any problems on the dimension. To maintain comparability between the samples, we excluded cases in which the L1 was assigned a value lower than 90. For respondents assigning a value between 90 and 100, we rescaled the individual scores such that L1 was set to 100, before averaging across respondents.

Case Selection

In the Norwegian study, we excluded individuals with more than two dimensions missing in the top and bottom tasks. In the within-dimension task, we excluded individuals if more than one of L2 to L5 were missing. Missing on "being dead" was allowed, because values for death are qualitatively different [21]. We excluded respondents assigning the same score to L2, L3, L4, and L5. No cases were excluded in the Finnish study [4].

Comparing Norwegian and Finnish Task Scores

Weights were applied before all analyses to adjust for differences between the demographic makeup of the samples and the populations (Norway in 2010 and Finland in 1992). In the Norwegian sample, exclusions were performed at the level of individual tasks, and weights were calculated separately for each task. For the Finnish sample, a single set of weights was used for all analyses. We compared Norwegian and Finnish scores provided in the valuation tasks using independent samples t tests, applying an α level of 0.05. We applied Bonferroni correction [22] for performing 90 tests (60 for the within-dimension tasks and 15 each for the top and bottom tasks). We calculated the Lin concordance correlation coefficient (CCC) [23] to compare the Norwegian and Finnish within-dimension scores for L2 to L5 for all dimensions (60 values) and the top and bottom task scores. The top and bottom tasks provide information on the importance/severity of each dimension. Similarly, the score for L5 from the within-dimension task can also be used as an indicator of overall severity for each dimension.

Covariance between the Tasks

We calculated Pearson product-moment correlations between the dimension means from the top task, the bottom task, and the score of L5 from the within-dimension task. We applied an α level of 0.05 on the basis of a two-sided test and reported the correlation strength according to the Cohen classification [24].

Comparing Algorithms

The Finnish algorithm was taken from the scoring sheet provided by Sintonen [5]. We compared the Finnish and the Norwegian algorithms by total ranges and by differences between the Norwegian and Finnish values. We estimated the CCC to compare the concordance of all L2 to L5 disutilities between the samples [23]. Because the algorithms are based on combinations of averages, there is no variance that allowed statistical tests of mean differences between Norway and Finland. Therefore, another criterion was needed to define what is considered to be a meaningful difference. King [25] discusses approaches to identify a meaningful difference for patient-reported outcome measures, her main point being that a minimal important difference (MID) is sample-specific [25]. Consequently, we used an estimate derived from our data to quantify differences between samples and dimensions, rather than using a previously published MID for the 15D [26]. Assuming that patients would find the deterioration or improvement moving from one level to the next in the 15D descriptive system to be of importance, we chose the smallest value change in the Norwegian algorithm as a conservative MID threshold.

Results

Sample and Case Selection

The Norwegian data set includes 2256 members of the Norwegian general population aged between 19 and 101 years (Table 1).

Table 1 – Characteristics of the Norwegian sample and the Norwegian general population in 2010.

Characteristic	Unweighted Norwegian sample (n = 2,256)	Norwegian population in 2010 (n = 3,937,847)
Sex		
Men	1,089 (48%)	1,956,835 (50%)
Women	1,167 (52%)	1,981,012 (50%)
Age (y)		
18–24*	140 (6%)	575,921 (15%)
25–39	433 (19%)	985,937 (25%)
40–59	895 (40%)	1,331,512 (34%)
60–66	360 (16%)	398,529 (10%)
>67	428 (19%)	645,948 (16%)
Education		
Elementary school	472 (20%)	1,111,379 (28%)
High school	960 (43%)	1,625,640 (41%)
University bachelor's degree	449 (20%)	811,360 (21%)
University master's degree	318 (14%)	269,627 (7%)
No formal education/ no response	57 (3%)	119,841 (3%)

* The Norwegian sample data include individuals from the age of 19 y, whereas the Norwegian population data include individuals from the age of 16 y.

We excluded 218 cases out of 1729 respondents who received the top task, and 290 cases out of 1714 respondents who received the bottom task. Case exclusion for the within-dimension task is reported in Appendix 4 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018.

Comparing Norwegian and Finnish Task Scores

L5 disutility scores display small differences between the dimensions and between the country-specific samples (Fig. 1A). There appears to be a pattern in which L2 and L3 disutility scores are

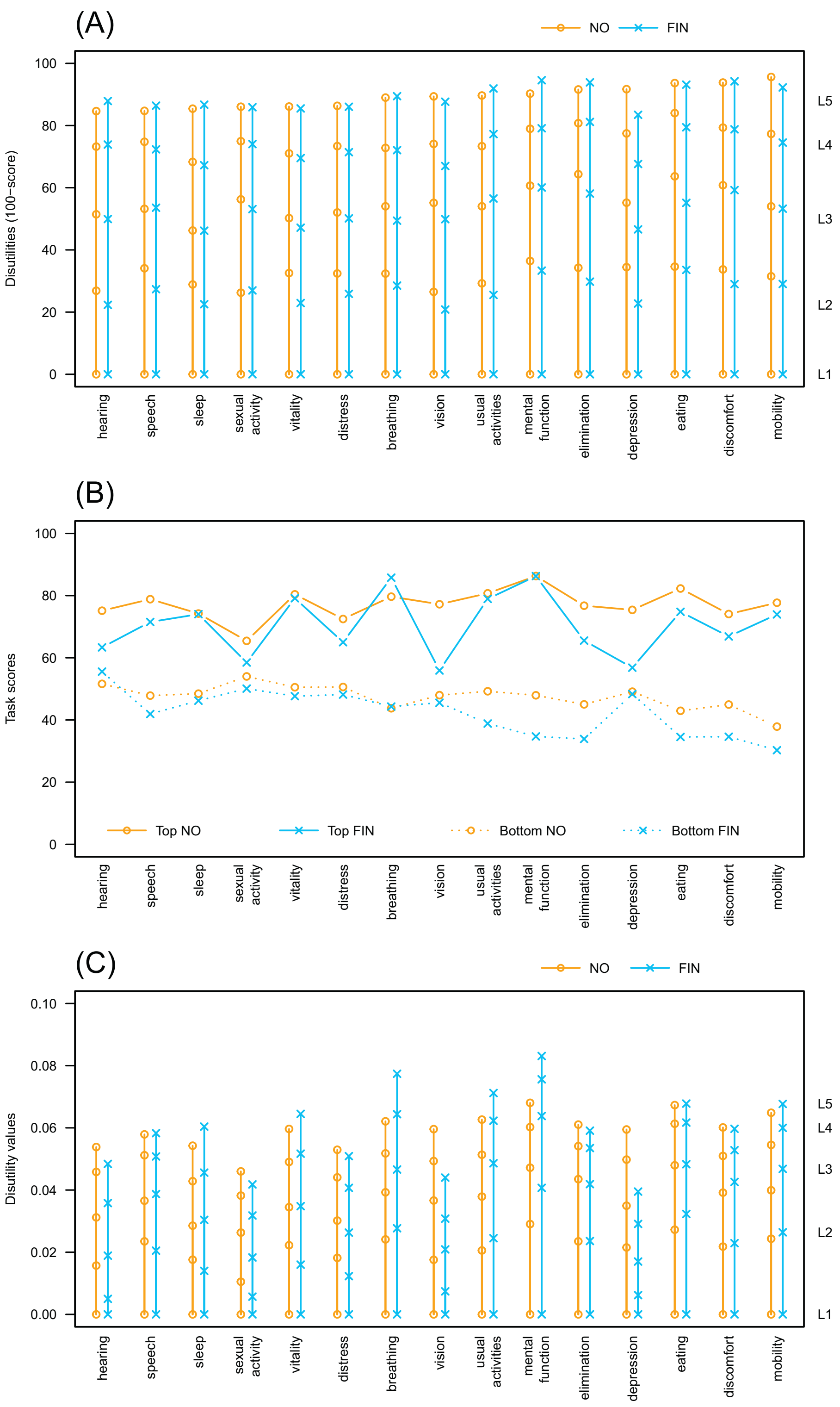

Fig. 1 – Norwegian and Finnish task scores and algorithm values: (A) within-dimension task scores, (B) top and bottom task scores, and (C) algorithm disutility values.

smaller in the Finnish sample than in the Norwegian sample (Fig. 1A). The difference is significant for six dimensions on L2 and three dimensions on L3 (Table 2; see also Appendix 5 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018). There were statistically significant differences between the Norwegian and Finnish scores for all levels of the depression dimension.

The top task scores for both samples display greater variation between the dimensions (Fig. 1B) compared with the L5 disutility scores from the within-dimension task (Fig. 1A). Top task scores range from 55.9 to 86.3 in the Finnish sample and from 65.4 to 86.3 in the Norwegian sample. Norwegian top task scores were significantly higher for the dimensions of hearing, speech, sexual activity, distress, vision, elimination, depression, eating, and discomfort, and were significantly lower for breathing compared with the Finnish top task scores (Table 2).

The Finnish bottom task scores (30.3, 55.6) have a wider range than the corresponding Norwegian bottom task scores (37.9, 54.0; Fig. 1B). The independent samples *t* test indicated statistically significant lower Finnish bottom task scores for the dimensions of usual activities, mental function, elimination, eating, and discomfort (Table 2; see also Appendix 6 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2017.09.018).

The CCC between the countries was the highest for the within-dimension scores ($p_{\text{c within}} = 0.98$) and considerably lower for top and bottom task scores ($p_{\text{c top}} = 0.41$ and $p_{\text{c bot}} = 0.47$).

Covariance between the Tasks

There were medium, negative, statistically nonsignificant correlations between the top task scores and the bottom task scores in both samples ($r_{\text{NO}} = -0.42$, $P = 0.122$; $r_{\text{FIN}} = -0.41$, $P = 0.131$). The correlations between top task scores and the within-dimension scores for L5 were also statistically nonsignificant and negatively correlated ($r_{\text{NO}} = -0.29$, $P = 0.291$; $r_{\text{FIN}} = -0.43$, $P = 0.105$). Nevertheless, the bottom task scores display a high, statistically significant, and positive correlation with the within-dimension scores for L5 ($r_{\text{NO}} = 0.78$, $P < 0.001$; $r_{\text{FIN}} = 0.84$, $P < 0.001$). The correlational pattern was similar for the two country-specific samples.

Comparing Algorithms

The largest disutility values were assigned to the dimension of mental function in both the Norwegian (0.068) and the Finnish (0.083) algorithms (Table 3). Sexual activity had the lowest Norwegian disutility value for L5 (0.0460), and depression had the lowest Finnish L5 disutility (0.0395; Table 3). The range from full health to the worst possible 15D health state was 1 to 0.1062 for the Finnish algorithm and 1 to 0.1102 for the Norwegian value algorithm.

We calculated the differences between the 60 algorithm values by subtracting the Finnish values from the corresponding Norwegian values. The average difference between the two value

Table 2 – Significant independent samples t test results per valuation task.

Valuation task	M_{NO}	SD_{NO}	M_{FIN}	SD_{FIN}	t	*df*	Probability
			Within-dimension task				
Speech$_{\text{L2}}$	34.079	25.118	27.348	23.113	4.138	861.167	0.000*
Sleep$_{\text{L2}}$	28.875	24.356	22.526	21.375	4.116	855.424	0.000*
Vitality$_{\text{L2}}$	32.549	24.691	22.912	20.577	6.361	853.508	0.000*
Distress$_{\text{L2}}$	32.420	24.292	25.870	22.176	4.299	925.502	0.000*
Vision$_{\text{L2}}$	26.509	24.016	20.822	19.830	3.845	822.518	0.000*
Vision$_{\text{L4}}$	74.093	30.130	66.976	22.247	3.981	787.982	0.000*
Elimination$_{\text{L3}}$	64.359	27.042	58.110	23.948	3.564	812.862	0.000*
Depression$_{\text{L2}}$	34.445	27.040	22.755	20.436	7.261	799.967	0.000*
Depression$_{\text{L3}}$	55.179	27.096	46.603	22.347	5.091	817.728	0.000*
Depression$_{\text{L4}}$	77.465	26.045	67.637	19.906	6.281	805.153	0.000*
Depression$_{\text{L5}}$	91.701	27.078	83.463	17.009	5.378	716.416	0.000*
Eating$_{\text{L3}}$	63.651	25.295	55.144	22.667	5.293	877.827	0.000*
			Top task				
Hearing$_{\text{Top}}$	75.141	24.693	63.343	24.640	7.967	503.513	0.000*
Speech$_{\text{Top}}$	78.844	23.762	71.516	23.228	5.219	511.153	0.000*
Sexual activity$_{\text{Top}}$	65.443	30.408	58.480	28.745	3.942	511.384	0.000*
Distress$_{\text{Top}}$	72.474	25.485	65.012	26.089	4.777	495.637	0.000*
Breathing$_{\text{Top}}$	79.658	21.291	85.776	17.299	–5.655	607.448	0.000*
Vision$_{\text{Top}}$	77.217	24.034	55.914	29.571	12.349	441.001	0.000*
Elimination$_{\text{Top}}$	76.756	22.717	65.534	23.921	7.897	490.020	0.000*
Depression$_{\text{Top}}$	75.403	24.903	56.822	29.610	10.741	454.066	0.000*
Eating$_{\text{Top}}$	82.289	22.338	74.792	23.218	5.411	491.230	0.000*
Discomfort$_{\text{Top}}$	74.069	23.852	66.863	26.778	4.551	462.849	0.000*
			Bottom task				
Usual activities$_{\text{Bot}}$	49.230	34.745	38.845	31.353	5.449	588.654	0.000*
Mental function$_{\text{Bot}}$	47.934	36.358	34.697	39.247	5.677	496.270	0.000*
Elimination$_{\text{Bot}}$	45.015	36.686	33.881	34.997	5.281	557.813	0.000*
Eating$_{\text{Bot}}$	42.944	40.285	34.548	38.085	3.638	554.603	0.000*
Discomfort$_{\text{Bot}}$	44.957	34.073	34.606	34.153	5.075	534.772	0.000*

Note. Independent samples *t* tests between the means of the Norwegian and the Finnish scores from the valuation tasks were conducted with an α level of 0.05, applying Bonferroni correction for 90 tests. Results for all levels and dimensions can be found in Appendices 5 and 6 in Supplemental Materials.

df, degrees of freedom; M, mean; SD, standard deviation; *t*, *t* test value.

* Significant mean differences judged against the Bonferroni corrected α level (0.00055).

algorithms was 0.0006, whereas the absolute average difference was 0.0055. The concordance between the Norwegian and the Finnish algorithm values was smaller ($p_{c\ \mathrm{L2L5\ alg}} = 0.87$) than the concordance between the corresponding scores from the within-dimension task ($p_{c\ \mathrm{L2L5\ VAS}} = 0.98$). The observed pattern from the within-dimension task in which Norwegian L2 and L3 disutility scores were larger than the corresponding Finnish scores was not reflected in systematically larger Norwegian algorithm values on L2 and L3. Using the MID margin corresponding to the smallest value change in the Norwegian algorithm (0.0060) as a margin of importance, the Norwegian algorithm disutility values exceeded the Finnish values on 14 levels and were lower on 12 levels. The Finnish algorithm disutility values exceeded the Norwegian values on the dimensions of mobility (L3), breathing (L3, L4, and L5), sleeping (L5), usual activities (L3, L4, and L5), and mental function (L2, L3, L4, and L5) (Fig. 1C). The Norwegian algorithm disutility values exceeded the Finnish values for vision (L2, L3, L4, and L5), hearing (L2, L3, and L4), depression (L2, L3, L4, and L5), vitality (L2), and sexual activity (L3 and L4) (Fig. 1C).

Discussion

Summarizing Results from the Task Scores

Compared with the scores from the top and bottom tasks, the scores of the within-dimension task display very similar patterns across the dimensions and the two samples, which is reflected by only a few statistically significant differences and the highest observed CCCs. These findings are in line with the early VAS-based valuation studies of the EuroQol group that summarized several country-specific VAS-based valuation studies to one value set [27]. Nevertheless, the scores for L2 and L3 from the within-dimension task were systematically smaller in the Finnish sample than in the Norwegian sample. The top and bottom task scores displayed lower concordance between the countries, and the ranges of the Finnish top and bottom task scores were wider.

Summarizing Value Algorithm Results

On average, the two algorithms are very similar with regard to their ranges and the average differences between their algorithm values. Nevertheless, there are substantial differences in the ranges of individual dimensions. Furthermore, the concordance between the Finnish and Norwegian algorithm values were substantially lower than the corresponding concordance between the scores from the within-dimension task. On the dimension level, there are differences in algorithm values between samples and dimensions. The largest differences in the algorithms of the two samples were observed in the dimensions of depression, vision (larger disutility values in the Norwegian algorithm), and mental function (larger Finnish disutility values). The L5 algorithm disutility value for mental function was the largest in both samples. To sum up, the comparison of the Norwegian and Finnish 15D algorithms displays substantial differences that are not reflected in the corresponding scores from the within-dimension task.

Associations between Task Scores and Resulting Algorithm Values

Scores from all three tasks are combined in the procedure developed by Sintonen [4]. The similarity in Norwegian and Finnish within-dimension scores (Fig. 1A) contrasts with considerable differences observed between the final algorithm values from the two countries. This contrast can be demonstrated by the example of the breathing dimension. As can be seen in Figure 1A, the Norwegian and Finnish scores from the within-dimension

Table 3 – Norwegian and Finnish 15D algorithm disutility values.

Level	Hearing	Speech	Sleep	Sexual activity	Vitality	Distress	Breathing	Vision	Usual activities	Mental function	Elimination	Depression	Eating	Discomfort	Mobility
								Norwegian							
Level 2	0.0157	0.0235	0.0176	0.0105	0.0223	0.0182	0.0241	0.0176	0.0206	0.0291	0.0235	0.0215	0.0272	0.0218	0.0243
Level 3	0.0312	0.0366	0.0285	0.0263	0.0345	0.0302	0.0393	0.0366	0.0379	0.0472	0.0435	0.0349	0.0480	0.0392	0.0399
Level 4	0.0458	0.0512	0.0428	0.0382	0.0490	0.0441	0.0518	0.0493	0.0513	0.0602	0.0541	0.0498	0.0613	0.0510	0.0545
Level 5	0.0538	0.0579	0.0543	0.0460	0.0597	0.0529	0.0621	0.0596	0.0627	0.0680	0.0610	0.0595	0.0673	0.0601	0.0649
								Finnish							
Level 2	0.005	0.0205	0.014	0.0057	0.016	0.0123	0.0277	0.0074	0.0245	0.0407	0.0236	0.0062	0.0323	0.0229	0.0264
Level 3	0.0189	0.0387	0.0304	0.0183	0.0348	0.0263	0.0466	0.0209	0.0486	0.0638	0.0419	0.017	0.0483	0.0426	0.0468
Level 4	0.0358	0.0508	0.0456	0.0318	0.0517	0.0407	0.0644	0.0308	0.0623	0.0756	0.0535	0.0291	0.0617	0.0528	0.06
Level 5	0.0484	0.0583	0.0604	0.0418	0.0645	0.0509	0.0774	0.044	0.0712	0.0831	0.0591	0.0395	0.0678	0.0597	0.0677

Note. The value of a health state is 1 minus the sum of the disutilities associated with the level of each dimension.

task are very similar, whereas the Finnish top task score for breathing is clearly higher than the corresponding Norwegian score (Fig. 1B). The resulting algorithm values for breathing dimension differ considerably between Norway and Finland (Fig. 1C; Table 3). This raises the question as to where the observed differences in algorithm values originate.

In the algorithm estimation procedure, the scores from the within-dimension task are rescaled using the scores from the top and bottom tasks. This procedure stretches or compresses the range of the scores from the within-dimension task, depending on the distance between the top task and the bottom task scores for the relevant dimension. In the case of the breathing dimension, the higher Finnish top task score implies a large difference between the top and the bottom task scores, which results in a wider range for the Finnish algorithm values for breathing than the corresponding Norwegian algorithm values. The rescaling with the top task and bottom task scores also explains why the dimensions of vision, sexual activity, and depression have much smaller algorithm disutility values than observed for scores from the within-dimension task: for these three dimensions the distance between the top and bottom task scores is small.

Covariance between the Valuation Tasks

The within-dimension task provides valuations of each level of all dimensions directly related to death. If we assume that the prospect of “being dead” has a fixed level of attractiveness for each respondent, each level of each dimension is valued in relation to the same anchor. Note that this does not suggest that “being dead” should be assigned the same score on the VAS across the dimensions. “Being dead” is listed as the last item to be valued in the within-dimension task, below the five levels of functioning. If respondents value the levels of each dimension from top to bottom, they are likely to focus on the relative distance between levels and then consider the overall severity or importance of the dimension when they reach “being dead” (see Appendix 2 in Supplemental Materials for a visual example). Thus, information is provided about the relative distances between levels as well as their absolute distances to the “being dead” anchor. If one accepts that valuing health states against death is a valid method, then this task alone should arguably provide sufficient information to estimate a 15D algorithm. We also note that these assumptions are, essentially, required for the rescaling conducted in the estimation methods developed by Sintonen, because scores from the within-dimension task are rescaled using death as an anchor.

The top task asks respondents to place 15 descriptions of full function on a VAS anchored in “most important” and “least important.” The bottom task asks respondents to put the 15 descriptions of the lowest function on a VAS anchored in “best imaginable health state” and “worst imaginable health state.” The scores of these tasks are used in the algorithm estimation procedure to allocate a proportion of the 15D algorithm range to each dimension according to their importance.

It is not obvious from the publication introducing the 15D algorithm estimation procedure as to how the valuation tasks are supposed to be related [4]. The bottom task and the within-dimension task use the same descriptions of L5 and the same scale. On the basis of these observations, we should expect that the scores from the bottom task and the corresponding L5 scores from the within-dimension task should capture similar information about dimension severity. The high, positive correlation observed between these scores indicates that the two tasks capture common variance. In contrast, we did not observe statistically significant correlations between the top task with any of the other tasks. One possible explanation for lacking common variance might be that the top task elicits scores on a scale with anchors that are different from those of the other two tasks. Nevertheless, the question of how the anchor concepts of “imaginable health state” and “importance” overlap remains open.

Problems Arising from Combining Valuation Tasks

Because the relationships between the valuation tasks remain unclear, it is not obvious how information from these tasks should ideally be combined. The original 15D algorithm estimation procedure uses the scores of the top and bottom tasks to rescale the scores of the within-dimension task. These rescaled scores are then multiplied with the scores of the within-dimension task rescaled to 1 and 0. This approach raises at least two questions: Why are the top and bottom tasks performed in addition to the within-dimension task? and Why are the scores of the three valuation tasks combined by multiplication?

Collecting conceptually overlapping measures from different tasks can be driven by the idea of method triangulation or to increase reliability. Reliability is generally increased by taking a weighted average of scores from several measures or tasks, the result of which will be less susceptible to variation and error than the scores from each constituent measure. Unlike a weighted average, multiplication of information from different sources will tend to increase the risk of error because of random variation or bias. Which approach to choose so as to combine different sources of information is relevant beyond the 15D and beyond the context of valuation methods. For example, two recently proposed valuation methods both involve multiplication of scores from different tasks: the analytic hierarchy process [28,29] and the modeling of personal utility function-based value sets [30]. The analytic hierarchy process involves weighing of attributes within wider dimensions and global weighing between dimensions. The final weights for the attributes are derived by multiplying the global weights with the local ones. The methods being developed to measure personal utility functions for health use swing-weighing and visual props reminiscent of VAS and also use weighing at local and global levels multiplied to derive final weights.

A Simplified 15D Algorithm Estimation Procedure

Given these considerations, we propose to use only the within-dimension task alone to estimate future 15D algorithms. The within-dimension scores are all on the same scale, anchored in death, and provide all information theoretically required to calculate a 15D algorithm. In contrast to the top and bottom task scores, the interpretation of the scores from the within-dimension task appears straightforward. Furthermore, we assume that the within-dimension task is the easiest to understand for respondents, because few and related levels are valued. Limiting the number of tasks has the added bonus of reducing survey costs and respondent burden.

Study Limitations

There are several differences between the two samples that may limit comparability. First, we used Finnish and Norwegian 15D translations. It is possible that differences between the algorithms are due to words that differed in meaning [11]. Second, 18 years had passed between the Finnish data collection in 1992 and the Norwegian collection in 2010. Although it is possible that health state preferences had changed over time, a Finnish replication study conducted in 2001 found no considerable changes in Finnish values [31]. Third, the Norwegian and the Finnish valuation studies varied in the mode of administration (Web and postal vs. postal) and in how the within-dimension task was presented in the Norwegian postal and the Web sample (vertical vs. horizontal). Earlier studies showed that a horizontal

VAS produces scores with a more uniform distribution than a vertical VAS [32]. Fourth, in the Finnish valuation study, the within-dimension task included five levels of functioning plus "unconscious" and "being dead," whereas "unconscious" was dropped in the Norwegian study. If one assumes that respondents will display a tendency to distribute scores on the VAS [32], valuing six levels instead of seven might tend to lower values for L5 in the Norwegian sample, because there is more space available at the lower end of the scale. Nevertheless, Figure 1A shows that Norwegian L5 scores were not systematically lower than corresponding Finnish scores.

Conclusions

Comparing a Norwegian to a Finnish 15D algorithm illustrates several differences, the largest of which were observed between values for depression, vision, and mental function. Although we found very similar patterns in the Norwegian and Finnish samples for the valuation task comparing levels of each dimension to death, larger differences were found in the two remaining tasks. 15D algorithms are a product of combining three VAS-based valuation tasks by use of methods involving multiplication. This procedure serves to amplify variance from each of the tasks and appears to produce a result that is less similar between Norway and Finland than the results of the constituent tasks. We propose to use the within-dimension task alone in 15D value algorithm estimation to simplify interpretation, reduce risk of random error and bias, reduce survey costs, and reduce the burden on participants.

Acknowledgments

We thank Harri Sintonen for valuable feedback to numerous 15D-related questions. We are grateful to Mathias Barra for his generosity in helping us disentangle the mathematical niceties of the multi-attribute utility theory and how this applies to valuing the 15D instrument. Finally, we greatly appreciate the constructive and insightful feedback provided by anonymous reviewers, which has helped substantially in improving the article.

Source of financial support: Y.A. Michel received funding from the Norwegian Research Council to conduct this study (project number 229101). L.A. Augestad and K. Rand received funding from the Southeastern Norway Regional Health Authority to conduct the Norwegian 15D valuation study (project number 2008015) and from the Norwegian Research Council (project number 186500).

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2017.09.018 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Weinstein MC, Torrance G, McGuire A. QALYs: the basics. Value Health 2009;12:S5–9.

[2] Sintonen H. An approach to measuring and valuing health states. Soc Sci Med Med Econ 1981;15:55–65.

[3] Sintonen H, Richardson J. The 15-D measure of health-related quality of life: reliability, validity and sensitivity of its health state descriptive system. Working Paper, National Centre for Health Program Evaluation, Melbourne, Australia, 1994.

[4] Sintonen H. The 15D-measure of health-related quality of life, II: feasibility, reliability and validity of its valuation system. Melbourne, Australia: Working Paper, National Centre for Health Program Evaluation, 1995.

[5] Sintonen H. 15D instrument homepage. Available from: http://www.15d-instrument.net/15d/. [Accessed June 29, 2017].

[6] Stavem K. Quality of life in epilepsy: comparison of four preference measures. Epilepsy Res 1998;29:201–9.

[7] Drummond MF, Sculpher MJ, Claxton K, et al. Methods for the Economic Evaluation of Health Care Programmes. Oxford, UK: Oxford University Press, 2015.

[8] Brazier J, Ratcliffe J, Saloman J, et al. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford, UK: Oxford University Press, 2017.

[9] Richardson J, Khan MA, Iezzi A, Maxwell A. Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. Med Decis Making 2015;35:276–91.

[10] Canadian Agency for Drugs and Technologies in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada. Ottawa, ON: Canadian Agency for Drugs and Technologies in Health, 2006.

[11] National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. 2013. Available from: https://www.nice.org.uk/process/pmg9/chapter/the-reference-case. [Accessed May 23, 2017].

[12] National Health Care Institute. Guideline for economic evaluations in healthcare. 2016. Available from: https://english.zorginstituutnederland.nl/publications/reports/2016/06/16/guideline-for-economic-evaluations-in-healthcare. [Accessed June 29, 2017].

[13] Wittrup-Jensen K, Pedersen KM. Modelling Danish weights for the 15D quality of life questionnaire by applying multi-attribute utility theory (MAUT). Health Economics Papers 7, University of Southern Denmark, Odense M, Denmark, 2008.

[14] Bailey H, Kind P. Preliminary findings of an investigation into the relationship between national culture and EQ-5D value sets. Qual Life Res 2010;19:1145–54.

[15] Engel L, Bansback N, Bryan S, et al. Exclusion criteria in national health state valuation studies: a systematic review. Med Decis Making 2016;36:798–810.

[16] Lamers LM, McDonnell J, Stalmeier PFM, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 2006;15:1121–32.

[17] Norman R, Cronin P, Viney R, et al. International comparisons in valuing EQ-5D health states: a review and analysis. Value Health 2009;12:1194–200.

[18] Arnesen T, Trommald M. Are QALYs based on time trade-off comparable? A systematic review of TTO methodologies. Health Econ 2005;14:39–53.

[19] Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med 2001;33:328–36.

[20] Raiffa H. Decision Analysis with Multiple Conflicting Objectives. New York, NY: Wiley, 1976.

[21] Augestad LA, Rand-Hendriksen K, Stavem K, Kristiansen IS. Time trade-off and attitudes toward euthanasia: implications of using "death" as an anchor in health state valuation. Qual Life Res 2013;22:705–14.

[22] Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961;56:52–64.

[23] Lin L, Hedayat AS, Sinha B, Yang M. Statistical methods in assessing agreement: models, issues, and tools. J Am Stat Assoc 2002;97:257–70.

[24] Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd ed). Hillsdale, NJ: Routledge, 1988.

[25] King MT. A point of minimal important difference (MID): a critique of terminology and methods. Expert Rev Pharmacoecon Outcomes Res 2011;11:171–84.

[26] Alanne S, Roine RP, Räsänen P, et al. Estimating the minimum important change in the 15D scores. Qual Life Res 2015;24:599.

[27] Greiner W, Weijnen T, Nieuwenhuizen M, et al. A single European currency for EQ-5D health states. Eur J Health Econ 2003;4:222–31.

[28] Reddy BP, Adams R, Walsh C, et al. Using the analytic hierarchy process to derive health state utilities from ordinal preference data. Value Health 2015;18:841–5.

[29] Danner M, Hummel JM, Volz F, et al. Integrating patients' views into health technology assessment: analytic hierarchy process (AHP) as a method to elicit patient preferences. Int J Technol Assess Health Care 2011;27:369–75.

[30] Devlin N, Shah K, Mulhern B, et al. Modelling personal utility function-based value sets for EQ-5D. Presented at: The 33rd EuroQol Plenary Meeting, Berlin, September 15–16, 2016.

[31] Kotomäki T, Sintonen H. Are the valuations for the 15D HRQoL instrument stable over time? Ital J Public Health 2005;2:42–53.

[32] Scott J, Huskisson EC. Vertical or horizontal visual analogue scales. Ann Rheum Dis 1979;38:560.