Internet-based treatment for eating disorders: bridging the treatment gap

Rohrbach, P.J.

Chapter 6

The ICEpop Capability Measure for Adults instrument for capabilities: development of a tariff for the Dutch general population

Published as: Rohrbach, P. J., Dingemans, A. E., Groothuis-Oudshoorn, C. G., Van Til, J. A., Essers, B. A., Van Furth, E. F., & Van den Akker-Van, M. E. (2022). The ICEpop Capability Measure for Adults Instrument for Capabilities: Development of a Tariff for the Dutch General Population. *Value in Health, 25*, 125-132. https://doi.org/10.1016/j.jval.2021.07.011

Abstract

Objectives: The ICEpop Capability Measure for Adults (ICECAP-A) assesses 5 capabilities (stability, attachment, autonomy, achievement, and enjoyment) that are important to one's quality of life and might be an important addition to generic health questionnaires currently used in economic evaluations. This study aimed to develop a Dutch tariff of the Dutch translation of the ICECAP-A.
Methods: The methods used are similar to those used in the development of the UK tariff. A profile case best–worst scaling task was presented to 1002 participants from the general Dutch population. A scale-adjusted latent class analysis was performed to test for preferences of ICECAP-A capabilities and scale heterogeneity.
Results: A 3-preference class 2-scale class model with worst choice as scale predictor was considered optimal and was used to calculate the resulting tariff. Results indicated that the capabilities stability, attachment, and enjoyment were considered more important aspects of quality of life than autonomy and achievement. Additionally, improving capabilities from low to moderate levels had a larger effect on quality of life than improving capabilities that were already at a higher level.
Conclusion: The ICECAP-A tariffs found in this study could be used in economic evaluations of healthcare interventions in The Netherlands.

Introduction

Efficient allocation of resources is becoming increasingly important when it comes to making decisions in healthcare and health policy. Cost-utility analysis is a central tool for judging the efficiency of interventions and can support decisions on healthcare funding. Generally, quality-adjusted life-years (QALYs) are the central outcome measure in cost-utility analyses. To assess quality of life, generic utility measures are often used, such as the EQ-5D (EuroQol Group, 1990) or the Short-Form 6 Dimensions (Brazier et al., 1998). Nevertheless, there is critique on the use of generic health questionnaires for economic evaluations, mainly that not all relevant domains of quality of life are captured by these instruments (Byford & Sefton, 2003; Carr-Hill, 1989; Coast, 2004). Indeed, Pietersma et al. (2013) analyzed several generic utility measures and found that they capture only a selective amount of domains of quality of life and use an almost exclusive focus on people's current functional abilities with little emphasis on coping capabilities and resources. Consequently, relevant benefits of interventions outside the area of physical health might be underestimated in current economic evaluations.

Accordingly, considering a different, broader approach not limited to health-related quality of life might be more appropriate for determining treatment outcomes, especially for patients with a psychiatric disorder (Mitchell et al., 2017) or chronic illness. One such approach is based on capabilities (Sen, 1992, 1993). Capabilities indicate the extent to which someone is able to do what one wishes to do. The ICEpop Capability Measure for Adults (ICECAP-A) (Al-Janabi et al., 2012) is an instrument that measures well-being based on capabilities and may be an appropriate addition to the established EQ-5D. The instrument is receiving increased international recognition (Flynn et al., 2015) and may be used for economic evaluations of treatments aimed at improving not only physical health but well-being in general. Indeed, regarding its construct, existing research suggests that the ICECAP-A correlates positively with concepts such as feelings of happiness and freedom (Al-Janabi et al., 2013) and that it can capture information beyond health-related quality of life (Afentou & Kinghorn, 2020; Engel et al., 2017; Keeley et al., 2016). Economic evaluations that have already been conducted with the ICECAP-A suggest that using capabilities might lead to different decisions on resource allocation (Al-Janabi et al., 2013).

To be able to use the ICECAP-A in economic evaluations tariffs are needed to translate answers of patients on the ICECAP-A to a capability value between "0" and "1," where "0" represents "not at all able to do what one wishes" and "1" represents "fully able to do what one wishes." These anchoring values are different to utility values where "0" represents "health as bad as death" and "1" represents "perfect health." Tariffs of the ICECAP-A of a certain population indicate how important the various capabilities are according to that population and they might differ between populations, cultures, and countries. A tariff already exists for the general population of the United Kingdom (Flynn et al., 2015), but to be able to reliably use the ICECAP-A in other countries, tariffs for those countries need to be developed. In The Netherlands, using the ICECAP in economic evaluations is recommended when benefits regarding well-being are expected (Zorginstituut Nederland, 2015), but no Dutch tariff is available. This study aimed to develop an ICECAP-A tariff for the Dutch general population.

Methods

Design, Participants, and Procedure

Methods used to establish the Dutch tariff of the ICECAP-A are similar to those used for the development of the UK tariff by Flynn et al. (2015). Participants were approached by a market research agency (Kantar Group). A sample of 1002 participants was recruited that was representative of the Dutch general population based on age, gender, region, and income. Because questionnaires were completed online with less possibility for guidance throughout the assessment compared with the interviews in the study by Flynn et al. (2015) and a larger study size was recommended by the UK research group, a sample size of 1000 was assumed to be adequate for establishing the Dutch tariff. Additionally, Yang et al. (2015) showed that in discrete choice experiments a sample size of 1000 provides sufficient power for study designs that were similar to that of the current study (type 2 best–worst scaling, conditional logit latent class model) in terms of estimator properties. Participants were first informed about the study and could only continue to the online questionnaire if they consented with participating. They were paid a small sum of money to complete the questionnaire. Only fully completed assessments were saved and no information on the amount or content of partially completed questionnaires was stored. Information the researchers received from the marketing bureau was anonymous and could not be traced back to individuals. An independent medical ethics committee evaluated the study and confirmed it did not fall within the Medical Research Act, waiving the need for ethical approval (Medisch Ethische Toetsingscommissie Leiden-The Hague-Delft, file number N19.119).

Measurements

Best-worst scaling task

The ICECAP-A comprises 1024 (4 levels for each capability) possible states. Using the orthogonal main-effect plan (OMEP) design created by Flynn et al. (2015), 16 profiles, each containing 1 possible ICECAP-A state, were presented to participants. Half of the participants were presented with the 16 profiles from the OMEP design and the other half with its 16 foldover profiles (e.g., capabilities presented at level 4, 3, 2, or 1 in the original OMEP design were presented at level 1, 2, 3, and 4, respectively, in this foldover). The OMEP design and its foldover can be found in Appendix E.1. For each of the 16 profiles, participants had to indicate which of the capabilities they valued as best and which as worst. This is known as a (profile case) best–worst scaling task (Potoglou et al., 2011). An example of a profile can be seen in Figure 1. A pilot questionnaire was completed in an in-person interview by a convenience sample of 10 people of different ages and educational level to confirm the task would be understood by participants.

In the final questionnaire, participants were first asked to complete questions on demographics and their health and the ICECAP-A (Flynn et al., 2010). Details on these questionnaires can be found in Appendix E.2. Here, the levels of capabilities (shown behind every statement) were presented. Participants rated the experienced difficulty of completing the ICECAP-A on a 4-point scale (ranging from 1 "very easy" to 4 "very difficult"). Then, based on experiences from the pilot, an explanation of the best–worst scaling task was given

Figure 1. Example of a completed best–worst profile.

	Best	Worst
I am able to feel settled and secure in a few areas of my life [2]	○	○
I can have quite a lot of love, friendship and support [3]	○	○
I am able to be completely independent [4]	●	○
I can achieve and progress in all aspects of my life [4]	○	○
I cannot have any enjoyment and pleasure [1]	○	●

Note. Sixteen such profiles were completed in Dutch by participants. The number in straight brackets [#] indicates the level of the corresponding statement, ranging from [1], the lowest level, to [4], the highest level. In the example, the participant evaluated statement 3 "completely independent" to be the best (i.e., adds the most to a valuable life) and statement 5 "cannot have any enjoyment and pleasure" to be the worst (i.e., obstructs having a valuable life the most).

with an example of one completed profile. The explanation and best–worst scaling task can be found in Appendix E.3.

Statistical Analyses

Best-worst pairs table

Firstly, a table was constructed with all possible best–worst pairs. In other words, a count was made of how often, for example, stability at level 1 was chosen as best, whereas attachment at level 1 was chosen as worst, which resembled 1 of the 320 possible best–worst pairs. The margins of the table provided an initial understanding of the perceived importance to quality of life of the 20 capability levels. Moreover, the table allowed inspection of the frequencies of unlikely choices (e.g., attributes presented at level 4 chosen as worst or attributes presented at level 1 chosen as best), providing insight into the quality of the data.

Best-minus-worst scores

Second, best-minus-worst scores for participants showed individual preferences for capability levels and were used to estimate choice consistency. Within the OMEP design (and its foldover), each capability level was presented 4 times. The best-minus-worst score for 1 capability level, then, equaled the times that a participant picked that capability level as best minus the times it was picked as worst. This resulted in 20 best-minus-worst scores ranging from $+4$ (0 times picked as best and 4 times picked as worst) to -4 (4 times picked as best and 0 times picked as worst). Next, for each individual, the sum of squares for each capability was used to calculate the empirical scale parameter (ESP), which gave an indication of the consistency with which a participant made choices. An ESP (ranging from 0 to 8) of approximately 4 was considered normal for a participant who understood the task and made consistent choices (Flynn et al., 2015). Participants with a suspicious answering pattern on the best–worst scaling task, identified by differing more than 2 standard deviations (SDs) from the average ESP, were excluded from analyses concerning the tariff development. Table 1 depicts a set of best-minus-worst scores of a participant to illustrate the calculations.

Table 1. Best-minus-worst scores for one of the participants.

Capability	Level	Best-minus-worst score	Normalized ($*1/4$) and squared	Sum of squares
Stability	1	-3	0.56	1.38
	2	0	0	
	3	2	0.25	
	4	3	0.56	
Attachment	1	-2	0.25	0.88
	2	-1	0.06	
	3	0	0	
	4	3	0.56	
Autonomy	1	-1	0.06	0.38
	2	-2	0.25	
	3	1	0.06	
	4	0	0	
Achievement	1	-1	0.06	0.19
	2	0	0	
	3	1	0.06	
	4	1	0.06	
Enjoyment	1	-3	0.56	1.44
	2	-2	0.25	
	3	1	0.06	
	4	3	0.56	
ESP				4.25

ESP empirical scale parameter

Scale-adjusted latent class analysis

Latent Gold 5.1 software was used for scale-adjusted latent class (SALC) analysis. These analyses can distinguish individuals with different preferences (i.e., preference heterogeneity) by adding preference classes and also individuals with similar preferences but with different choice consistency (i.e., scale heterogeneity) by adding scale classes (Magidson & Vermunt, 2007). Although SALC models are not the only option to model both preference and scale heterogeneity, they are widely used and unique in estimating separate classes with differing preferences (Groothuis-Oudshoorn et al., 2018). As new preference classes are added to the model, the software uses the data to predict the probability for an individual to fall within a certain class. Each class has its own parameters (comparable with regression coefficients) for each of the 20 capability levels of the ICECAP-A, where parameters further away from 0 signify greater importance (i.e., are more often chosen as best or worst than other capability levels). Effects coding was used with level 4 of enjoyment as reference level. Adding more classes to a model will often improve the fit, but a balance between fit and interpretability is warranted. Nevertheless, there are no clear guidelines for choosing one model over another. Therefore, we chose to follow a pragmatic approach by, on one hand, minimizing the Bayesian information criterion (BIC) and, on the other hand, looking for a solution with classes that were clearly separable. Apart from adding classes, it is possible to add scale classes to separately target scale heterogeneity (Vass et al., 2018). For people in the same class but in a different scale class, parameters of capability levels showed a similar pattern, but were scaled. The scaling factor was smaller than 1 if they were less consistent or larger than 1 if they were more consistent in making best–worst choices. Additionally, to account for possible heteroscedasticity (i.e., allow a different scale factor) between best and worst choices, a dummy variable indicating a worst choice was added as scale predictor to the estimated models. Finally, multiple starting seeds were used when estimating the SALC model to verify the stability of the solution.

In the final model, the relative attribute importance within each class gave an indication of the preferences of participants in that class. Attribute importance was calculated for the five attributes in all classes by dividing the parameter range of one ICECAP-A attribute (i.e., the difference between level 1 and level 4 parameters of an attribute) by the sum of five attribute parameter ranges.

ICECAP-A tariff

After identifying the preferred model, the parameters of each class and scale class were weighted by the size of the class (i.e., the probability that a participant falls into that particular class) by calculating the product of the raw parameters and the group probability. Finally, adding the weighted parameters for every capability level across groups resulted in 20 parameters that, when linearly transformed to range from 0 (i.e., level [1] for all capabilities) to 1 (i.e., level [4] for all capabilities), constituted the final tariff.

Results

Participants

In total, 1002 participants completed the online questionnaire. The distribution of the ESP can be found in Appendix E.4. The ESP differed 2 SDs from the mean (4.04, $SD = 1.18$) for 69 participants (40 below and 29 above the mean). Visual inspection confirmed that these participants had suspicious answering patterns (e.g., always choosing stability as best and enjoyment as worst, regardless of the level on which they were presented) suggesting they did not understand the task or did not take it seriously. These participants were excluded, leaving 933 participants for analyses. Excluding these participants did not influence representativeness of the sample (see Appendix E.5) or the balance between randomization to version 1 and 2 of the best–worst scaling task (50.1% vs 49.9%) and had a small effect on quality of the data (see Appendix E.6). The questionnaire took on average 14.2 minutes ($SD = 28.9$, range $3.8 - 618.4$) to complete. One participant for whom completion time was 5692 minutes was not included in this calculation. There were no missing data. Table 2 presents participant characteristics. The sample was highly representative of the general Dutch population in terms of age, gender, region, and income (see Appendix E.5). Most participants found the ICECAP-A very easy or easy to complete (93.9%).

Table 2. Frequencies (%) and means (standard deviations) of participant characteristics.

Variable	Category	Sample mean ($N = 933$)
Age		48.9 (17.1)
Gender	Female (%)	479 (51.3)
	Male (%)	453 (48.6)
	Other (%)	1 (0.1)
ICECAP-A	Capability value*	0.88 (0.13)
ICECAP-A difficulty	Very easy (%)	469 (50.5)
	Easy (%)	407 (43.6)
	Hard (%)	55 (5.9)
	Very hard (%)	2 (0.2)
EQ-5D-5L	Index scores*	0.86 (0.20)
ESP		4.07 (0.95)

Note. Values represent mean values with standard deviations in parentheses unless indicated otherwise.
ESP Empirical scale parameter.
*Values reflect scores based on the Dutch population tariff.

Best-Worst Pairs Table

The number of times each of the 320 best–worst pairs was chosen across all participants is presented in Table 3. The last column indicates how often a capability at a certain level is chosen as best, whereas the last row indicates how often a capability at a certain level is chosen as worst. For example, the capability attachment presented at level 4 ("I can have a lot of love, friendship and support") was chosen 1772 times (11.9% of best choices) as best and 229 times (1.5% of worst choices) as worst across all profiles that participants completed. The table suggests that high levels of stability, attachment, and, to a lesser extent, autonomy and enjoyment were often chosen as best, whereas high levels of achievement were infrequently chosen as best (9.7%, 11.9%, 7.2%, and 8.7%, respectively, vs 3.5%). For worst choices, preferences appeared less explicit, with low levels of stability, attachment, autonomy, achievement, and enjoyment all frequently chosen as worst (10.5%, 9.8%, 9.3%, 8.3%, and 10.1%, respectively).

Table 3. Best–worst pairs frequencies.

Best		Worst																				Total	% (best choices)
		Stability				Attachment				Autonomy				Achievement				Enjoyment					
	Level	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4		
Stability	1	×	×	×	×	15	14	9	12	14	11	9	8	16	15	13	14	22	8	10	8	198	1.33
	2	×	×	×	×	64	19	18	25	56	22	26	18	33	28	37	29	55	27	27	13	497	3.33
	3	×	×	×	×	207	47	19	17	134	73	34	56	171	114	43	55	193	78	24	30	1295	8.67
	4	×	×	×	×	179	61	27	15	159	151	46	33	134	66	104	82	184	151	25	26	1443	9.67
Attachment	1	5	5	9	10	×	×	×	×	9	8	13	12	28	22	23	17	14	7	10	5	197	1.32
	2	80	29	18	7	×	×	×	×	39	28	22	30	55	60	42	42	107	24	14	14	611	4.09
	3	165	76	18	33	×	×	×	×	252	98	51	48	191	158	81	77	149	78	18	21	1514	10.14
	4	233	143	30	24	×	×	×	×	169	121	34	47	170	174	113	91	236	143	24	20	1772	11.87
Autonomy	1	7	9	10	9	13	5	7	10	×	×	×	×	9	12	13	10	9	6	11	5	145	0.97
	2	72	13	12	17	49	24	15	16	×	×	×	×	35	24	27	25	97	22	18	19	485	3.25
	3	199	87	19	39	148	40	18	15	×	×	×	×	83	60	61	36	121	101	23	27	1077	7.21
	4	141	56	25	21	101	119	9	13	×	×	×	×	79	109	54	47	133	111	31	27	1076	7.21
Achievement	1	9	6	9	3	17	11	18	12	6	5	8	11	×	×	×	×	4	6	5	4	134	0.90
	2	60	14	7	19	31	18	10	19	18	14	10	15	×	×	×	×	44	10	11	8	308	2.06
	3	87	45	7	17	132	38	20	18	49	18	19	16	×	×	×	×	105	32	12	8	623	4.17
	4	65	60	11	14	64	54	12	14	59	51	14	17	×	×	×	×	40	29	10	12	526	3.52
Enjoyment	1	14	5	5	6	8	5	8	6	7	10	5	4	11	4	5	7	×	×	×	×	110	0.74
	2	56	23	15	18	59	14	7	10	50	22	22	23	34	16	20	33	×	×	×	×	422	2.83
	3	194	60	23	16	219	58	5	14	211	86	24	37	91	70	41	43	×	×	×	×	1192	7.98
	4	180	134	29	33	162	60	29	13	162	82	36	42	91	133	63	54	×	×	×	×	1303	8.73
Total		1567	765	247	286	1468	587	231	229	1394	800	373	417	1231	1065	740	662	1513	833	273	247	29856	100.0
% (worst choices)		10.5	5.1	1.7	1.9	9.8	3.9	1.6	1.5	9.3	5.4	2.5	2.8	8.3	7.1	5.0	4.4	10.1	5.6	1.8	1.7	100.0	

Note. Based on $N = 933$. Row margins indicate best choice frequencies and column margins indicate worst choice frequencies.

SALC Estimates

A 3-preference class 2-scale class model with worst choice as a scale predictor was considered optimal ($df = 871$, $BIC = 68992$, $R^2 = 0.25$). A 3-preference class was chosen because a third class added a substantial group with interpretable differences compared with a 2-preference class model ($df = 894$, $BIC = 71166$, $R^2 = 0.19$). Adding a fourth class resulted in one relatively small group that did not provide clear discrimination between already existing preference classes ($df = 854$, $BIC = 69224$, $R^2 = 0.25$). Two scale classes were added because they improved the fit of the model considerably. Adding a third scale class reduced both the fit and the interpretability of the model. All attribute parameters for participants in the second scale class were estimated to be 0.29 times those of participants in the first scale class, with most participants (58.1%) predicted to be in the first scale class. Finally, adding worst choice as a scale predictor increased the fit of the model and seemed relevant to control for the questionnaire design (where participants could pick the best and worst choice in whatever order they preferred). Indeed, the scaling factor for worst choices compared with best choices was 0.68 ($p < .001$). This suggests that participants switched the order of making best and worst choices throughout the best–worst scaling task, strengthening the choice to correct for questionnaire design by adding worst choice as a predictor in the model. Relatedly, a strong linear relation between the amount of best choices and the inverse of worst choices across each of the 20 capability levels was found ($r = 0.97$, $R^2 = 0.95$), indicating that best and worst data were proportional and can likely be pooled for analyses. A summary of the results on all estimated models can be found in Appendix E.7.

A table with attribute importance, based on the parameters from Table 4, can be found in Appendix E.8. Participants in preference class 1, containing 40.2% of the sample, showed little variation in attribute importance with stability, attachment, autonomy, achievement, and enjoyment accounting for 0.23, 0.20, 0.21, 0.17, and 0.20 of the space, respectively. Participants in class 2, containing 30.3% of the sample, were characterized by a very low preference for achievement ($.02$) with high preferences for the other four capabilities. Class 3 contained 29.5% of the participants and was distinguished by a high preference for attachment (0.30) and enjoyment (0.27) while indicating low importance of autonomy (0.14) and especially achievement ($.09$). For the total sample, the attribute importance for stability, attachment, autonomy, achievement, and enjoyment weighted by class size was 0.22, 0.24, 0.19, 0.13, and 0.22, respectively.

Table 4. Final model parameters and Dutch general population ICECAP-A tariffs.

	Class 1 sClass 1	Class 1 sClass 2	Class 2 sClass 1	Class 2 sClass 2	Class 3 sClass 1	Class 3 sClass 2	Final Dutch tariff
Class probability	0.2337	0.1686	0.1761	0.1270	0.1712	0.1234	
	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)	Coef (SE)	
Stability [1]	25.84 (.33)	21.71 (.04)	20.68 (.14)	20.20 (.20)	23.86 (.13)	21.13 (.08)	−0.0073

Stability [2]	21.02	20.30	0.52	0.15	20.63	20.19	0.1061
	(.12)	(.03)	(.04)	(.12)	(.10)	(.04)	
Stability [3]	3.17	0.93	1.37	0.40	1.97	0.58	0.2007
	(.18)	(.03)	(.07)	(.13)	(.10)	(.05)	
Stability [4]	4.22	1.23	1.34	0.39	2.09	0.61	0.2163
	(.20)	(.03)	(.09)	(.14)	(.10)	(.05)	
Attachment [1]	25.11	21.50	20.71	20.21	24.45	21.30	−0.0035
	(.31)	(.04)	(.13)	(.20)	(.13)	(.09)	
Attachment [2]	20.83	20.24	0.76	0.22	0.43	0.13	0.1223
	(.11)	(.03)	(.04)	(.11)	(.10)	(.03)	
Attachment [3]	2.71	0.79	1.56	0.46	3.49	1.02	0.2118
	(.18)	(.03)	(.07)	(.13)	(.11)	(.06)	
Attachment [4]	3.80	1.11	1.59	0.47	4.17	1.22	0.2344
	(.19)	(.03)	(.09)	(.15)	(.12)	(.08)	
Autonomy [1]	25.30	21.55	21.21	20.35	23.07	20.90	0.0027
	(.32)	(.04)	(.13)	(.18)	(.12)	(.08)	
Autonomy [2]	20.90	20.26	0.33	0.10	20.78	20.23	0.1043
	(.11)	(.03)	(.04)	(.11)	(.10)	(.03)	
Autonomy [3]	2.69	0.79	0.97	0.29	0.87	0.26	0.1784
	(.16)	(.03)	(.06)	(.12)	(.11)	(.04)	
Autonomy [4]	3.88	1.14	0.69	0.20	0.86	0.25	0.1920
	(.19)	(.04)	(.09)	(.17)	(.13)	(.05)	
Achievement	24.41	21.29	21.80	20.53	22.56	20.75	0.0143
[1]	(.29)	(.04)	(.11)	(.17)	(.13)	(.06)	
Achievement	20.90	20.26	21.69	20.49	20.94	20.28	0.0813
[2]	(.12)	(.04)	(.04)	(.12)	(.14)	(.04)	
Achievement	1.76	0.52	21.54	20.45	0.08	0.02	0.1308
[3]	(.14)	(.04)	(.05)	(.12)	(.13)	(.03)	
Achievement	2.90	0.85	21.63	20.48	0.02	0.00	0.1451
[4]	(.19)	(.04)	(.07)	(.14)	(.13)	(.04)	
Enjoyment [1]	25.25	21.53	21.10	20.32	24.14	21.21	−0.0063
	(.31)	(.04)	(.13)	(.19)	(.12)	(.08)	
Enjoyment [2]	21.64	20.48	0.09	0.03	0.07	0.02	0.1001
	(.14)	(.03)	(.05)	(.12)	(.12)	(.04)	
Enjoyment [3]	2.57	0.75	0.59	0.17	2.87	0.84	0.1932
	(.17)	(.04)	(.07)	(.12)	(.13)	(.06)	
Enjoyment [4]*	3.48	1.02	0.54	0.16	3.53	1.03	0.2122
	(.19)	(.04)	(.08)	(.14)	(.13)	(.07)	

Note. Scale factor sClass 2 compared with sClass $1 = 0.2925$.
Coef Coefficient, *sClass* Scale class
*Used as reference level

ICECAP-A Tariff for the General Dutch Population

Table 4 shows the coefficients for the different preference classes and scale classes, together with the tariff. The capability value can be deduced from the tariff by adding the values for the corresponding score. For example, a change in an ICECAP-A score of [12211] to [44323] would result in a change in capability value of 0.6762: from 0.2274 ($-0.0073 + 0.1223 + 0.1043 + 0.0143 - 0.0063$) to 0.9036 ($0.2163 + 0.2344 + 0.1784 + 0.0813 + 0.1932$). The capability value was scaled to range from 0 [11111] to 1 [44444].

In the chosen model, the capability attachment on level 4 was valued as most desirable (parameter $= 2.28$, tariff $= 0.2344$) and capability stability on level 1 as least desirable (parameter $= -2.60$, tariff $= -0.0073$) to one's quality of life. The largest increase in capability equals 0.1258 and is obtained when going from attachment level 1 ("cannot have any love, friendship, and support") to level 2 ("can have a little love, friendship, and support"). The average difference between capability levels was 0.0667. The largest relative importance was ascribed to attachment, accounting for 22.3% of the possible improvement, whereas achievement received the lowest preference, accounting for 13.1% of the possible improvement. In general, the capabilities stability, attachment, and enjoyment seem to be somewhat more important to quality of life than autonomy and achievement. In addition, improvements within a capability from a low level to a higher level (e.g., going from level 1 to 2) yielded larger increases in capability value than improving attributes that were already moderate to high (e.g., going from level 3 to 4).

Explorative analyses were conducted after developing the tariff to investigate what aspects of quality of life are important for different people. Details on these explorative analyses can be found in Appendix E.9.

Discussion

This study aimed to develop a tariff for the ICECAP-A based on a large representative sample from the general Dutch population ($N = 933$). The tariff shows that the five capabilities described in the ICECAP-A all contribute to quality of life. The capabilities stability, attachment, and enjoyment were somewhat more important than autonomy, and achievement contributed the least to quality of life. Going from one level to the next within an attribute does not have a linear effect on the tariff. Indeed, improving capabilities from low to moderate levels rather than from moderate to high is more valuable according to the current sample. Consequently, prioritizing to help people with low capabilities might result in larger well-being gains for society as a whole. This relates to the concept of "sufficient capability," an approach with the aim to maximize the number of people above a level of sufficient capability (Goranitis et al., 2017; Mitchell et al., 2015).

Most study findings are similar to those reported for the UK tariff (Flynn et al., 2015). It is to be expected that Dutch and UK populations have comparable preferences. Nevertheless, it is interesting to note that the Dutch sample seems to value high levels of enjoyment more and high levels of achievement less compared with the UK sample. This difference in preferences was also apparent in the European Values Study (European Values Study, 2017), where 95.6% of Dutch respondents indicated that leisure time is important in their lives compared with 91.9% of their UK counterparts. More strikingly, 81.0% of UK respondents indicated that the feeling to achieve something is an important aspect of a job, whereas this was only the case for 62.4% of the Dutch respondents. Consequently, interventions that increase the ability to enjoy life might have a slightly greater impact on quality of life in The Netherlands than the United Kingdom. Capturing these differences between countries in tariffs is important because they might ultimately influence funding decisions (Kiadaliri et al., 2015).

Strengths and Limitations

The SALC model used to find clusters of participants with similar answer patterns is a flexible model that enables the modeling of both preference and scale heterogeneity, resulting in a parsimonious model. The BIC was used to determine the final model. Nevertheless, this measure tends to overstate the number of preference classes (Groothuis-Oudshoorn et al., 2018), so the final model was also based on interpretability and face validity, inevitably introducing subjective judgment. Another choice was to use case 2 (profile) best–worst scaling to establish participant preferences on the ICECAP-A. It must be noted that although best–worst scaling tasks might be more statistically efficient than discrete choice experiments, estimates of preferences seem to be similar across methods (Whitty & Gonçalves, 2018) and evidence on the burden on participants is mixed (Flynn et al., 2007; Himmler et al., 2021; Mühlbacher et al., 2016). A strength of the study was the recruitment of a large sample to develop the tariff.

Several limitations were also present. First, people with lower education were somewhat underrepresented because the assessment was online and education was not included in the quotations. Additionally, the sample was slightly under representative of the 75- to 99-year age group. Possibly, this is related to a difficulty of finding participants in this age group with access to the internet. These differences between the sample and the Dutch population might have influenced the tariff slightly. Second, a pilot was conducted to identify problems and to assess the difficulty of the best–worst task, which led to significant improvements in explanations in the questionnaire. Nevertheless, the final questionnaire was completed online with no guidance making it impossible to check how participants interpreted the questions. At least 69 participants did not understand the task or take it seriously and were excluded from analyses, but it is realistic to assume that more participants struggled with the questionnaire. Indeed, the margins of best–worst pairs table reveal that in the remaining sample 12% of worst choices were a capability presented at level 4 and 5% of best choices were a capability presented at level 1. This is strange considering all profiles presented to participants had balanced capability levels with some capabilities presented at a high level and others at a low level. Nevertheless, because the conducted analyses could account for scale heterogeneity and the sample was large with the majority seeming to understand the task, it is expected that the current results still reflect preferences on quality of life of the Dutch general population accurately.

Use in Economic Analyses

To be able to compare (economic) benefits across interventions, it is necessary to consider both the effectiveness (i.e., quality of life) and life extension (i.e., quantity of life). Conceptually, it is difficult to interpret the capability value derived from tariffs of the ICECAP-A in the context of health economics and cost-utility analyses and in comparison with QALYs (Coast, Smith, et al., 2008; Cookson, 2005). The capability value is not a QALY because the lowest value is not anchored to "death," but to "no capability." Nevertheless, death is accounted for in the sense that death is associated with no capability even though the reverse is not necessarily true (e.g., consider a state in which capabilities are nonexistent or a state of unconsciousness) (Coast, Flynn, et al., 2008). Consequently, capability values have a meaningful anchor (i.e., no capability) and can be adjusted for time, by estimating gains in years lived with full capability (Flynn et al., 2015). Therefore, they can be used in economic evaluations in a similar way as QALYs. Although applied similarly, the ICECAP-A measures a related but distinct concept compared with generic health questionnaires (Afentou & Kinghorn, 2020). This suggests that the ICECAP-A is not a substitute, but rather a complement to generic health questionnaires (Engel et al., 2017; Keeley et al., 2016), as is also advocated by the National Institute for Health and Care Excellence social care guidelines (National Institute for Health and Care Excellence, 2016). Accordingly, the instrument seems to be especially suitable and valuable in contexts outside the traditional healthcare model, such as general well-being, social care, mental health (Goranitis et al., 2016; Mitchell et al., 2017), public health, and chronic illness. Indeed, the Dutch guidelines for conducting economic evaluations in healthcare recommend the use of the ICECAP when considering interventions aimed at improving general well-being (Zorginstituut Nederland, 2015).

Conclusion

This study developed a tariff for the ICECAP-A based on a large Dutch general population. This makes the ICECAP-A ready for use in economic evaluations in The Netherlands. The instrument is expected to be a valuable addition to other generic health questionnaires, especially when evaluating interventions outside the traditional health intervention model.