THE HEALTH UTILITIES INDEX
(George W. Torrance, David Feeny; 1986, 1990)

Purpose

The Health Utilities Index (HUI) system is a generic, preference-based measure of health status intended for use in clinical outcomes measurement, in health surveys, in program planning and resource allocation. The utility scores provide a summary index of health-related quality of life on a zero to 1.0 scale; they may be used in cost-utility analyses to calculate quality-adjusted life years and in population studies as quality weights for calculating quality-adjusted life expectancy (1, p250).

Conceptual Basis

The HUI separates the classification of a person s health state from the application of preference weights to that state. In classifying health states, the HUI records functional capacity rather than performance. The intent is to document the extent to which deficits in health status . . . inhibit or prohibit normal functioning, rather than to report the level at which an individual chooses to function (1, p241). Feeny et al. argued that performance measures confound three concepts: the person s health or physical capacity, their opportunities and choices, and their preferences. Thus, people with the same underlying capacity who face different opportunity sets or have different preferences may be assessed as having different health status in a system that relies on performance. . . (1, p243; 2, p496).

The HUI covers several aspects of health, which are termed attributes rather than dimensions to emphasize the link with multi-attribute utility theory (3). However, in keeping with the focus on capacity, the coverage is intentionally restricted to within the skin abilities (physical and emotional), and excludes role performance or social interaction. This corresponds to a focus on health-related quality of life, rather than quality of life in general (1, pp240-1; 2, p491; 3, p376). Feeny et al. argue that while it is important to measure social interaction as it interacts with health status, the two should be measured separately so that the relationships can be analysed in a non-tautological fashion (2, pp495-6). Similarly, the HUI omits positive aspects of emotional well-being, such as mastery or the sense of coherence (4).

The scaling system used to estimate utilities also influenced the content of the HUI. Within each attribute, several levels of function are specified, making a large number of possible permutations of levels across the attributes. The task of estimating utilities for large numbers of states is simplified if the health status classifications are chosen to be structurally independent. This means that a person s level on one attribute does not logically constrain their level on another, so that any combination of levels across the attributes should be biologically possible (3). This also means that there is little overlap in concepts across the attributes. The desire to achieve structural independence influenced the choice of attributes to be measured; working ability, for example, was not covered as it would be dependent on a certain level of physical function. The HUI3 system achieved full structural independence, whereas HUI2 has partial independence. For instance, level 5 of mobility (no use of arms or legs) precludes being in level 1 of self-care.

Description

The HUI has evolved over time; a summary is given by Patrick and Erickson (5, pp381-389). The original version (HUI Mark 1) was developed to assess the health of children who had been in neonatal intensive care; it was later extended for more general pediatric applications (6). The Mark 1 version is now seldom used. The Mark 2 was developed to measure treatment outcomes for childhood cancer (7). It assessed seven attributes: sensation (vision, hearing, speech), mobility, emotion, cognition, self-care, pain and fertility. Three to five levels of function were defined for each attribute. The most recent version, Mark 3, extends the range of the Mark 2 and alters the attributes included. In part this was due to an overlap between the self-care and other dimensions, which implied redundancy and complicated the scoring system. The Mark 3 covers eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain, each with five or six levels, described in Exhibit 10.35. The Mark 2 version is still used and may be the more appropriate instrument when it is necessary to assess self care or when fertility is an important issue. The Mark 2 also covers the type of analgesic required for pain relief, whereas the Mark 3 covers activity disruption due to pain (3, p378). The HUI2 emotion scale focuses on worry and anxiety; the HUI3 focuses on happiness versus depression. The standard questionnaires distributed by Health Utilities Incorporated include questions to classify people in both the HUI2 and 3 systems (address below).

The HUI3 classification system is shown in Exhibit 10.35. Unlike other health measures, there is no single questionnaire, but (as with a clinician rating scales), various approaches may be used to classify respondents into the functional levels shown in the Table. Alternative questionnaires are available from the HUI web site (www.fhs.mcmaster.ca/hug/index.htm). There is flexibility in the format of the questions and different versions cover self- or interviewer-administration; a proxy version is available for patients unable to respond themselves and different versions are available for adults and for children. For health surveys, questions may cover changes from the person s usual state; for use as an outcome measurement, current status may be assessed. Different recall periods may also be used (one, two, or four weeks) according to the purpose. Most versions include 31 questions, but skip patterns reduce this number in an actual interview, and administration takes an average of 2 minutes for a community population, ranging from 1.5 for people with no health problems to 5 or more minutes for people with multiple problems (8; 9).

Each different combination of levels from each attribute forms a unique health state. Utility weights for these were derived in two main stages. First, utility weights were estimated for each level of each attribute (shown in Exhibit 10.36) and second, a scoring function translates the resulting profile of weights into an overall index score, as outlined in Exhibit 10.37.

The attribute weights were generated in a preference scaling task employing a general population sample of 504 adults from Hamilton, Ontario, Canada (10). For simplicity, most of the weighting used visual analog scales (termed a feeling thermometer), and the resulting value scores were converted to utilities using a study-specific power function (1; 10). The power function was derived from a comparison between visual analog scores and utility scores derived from a separate standard gamble scaling task undertaken for a selection of health states (11, p714). This produced the eight single-attribute utility functions shown in Exhibit 10.36. These are scaled on a 0.0 to 1.0 scale, which does not reflect the differing relative weights of each attribute in forming a multi-attribute health profile. These weights are only used when comparing single attributes (e.g., pain levels before and after treatment).

Because of the very large number of possible permutations in the HUI, it was not feasible to score each multi-attribute health state directly. Instead, a subset of the possible states was scored using a standard gamble scaling procedure to generate weights, and then weights for the remaining permutations were estimated using a multi-attribute scaling analysis. Multi-attribute utility theory extends the von Neumann-Morgenstern utility theory to cover the utility weighting of combinations of states. Based on empirical analyses, Feeny et al. used the multiplicative form which allows for interactions between the individual attributes in forming the overall score. Thus, for example, the effect of being both blind and deaf might be rated as more severe than either alone, but less severe than the sum of the two (10, pp116-117). The multiplicative multi-attribute scoring function was used to estimate scores for all permutations of the health states; the calculation is shown below Exhibit 10.37. The mildest departure from perfect health (wearing glasses) has a utility of 0.973, while the utility for the lowest level on all eight attributes is -0.36, judged to be markedly worse than being dead. It will be noted that the weights in Exhibit 10.36 for Cognition levels 2 and 3 appear to be reversed; this is correct, and arose because the categories were first defined on a conceptual basis and only later were the empirical weights derived. Indeed, in a French study, the magnitude of the weights follows the rank ordering shown in Table 1 (12, Table 5).

The utility weights were cross-validated by comparing the estimated weights against weights measured directly in a separate scaling task. For the Mark 2, the highest discrepancy was 0.08 and the mean difference for three health states was 0.002 (13, Table 2; 11, Table 10). For the Mark 3, the mean difference for three marker states was 0.01, and the intraclass correlation (ICC) between them was 0.91 (10, p124). The ICC agreement for 73 states across two separate surveys was 0.88 (10, p125).

Exhibit 10.35 The HUI3 Classification System (see Exhibit for questions from which the classification is derived)

Attribute	Level	Description
Vision	1	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses.
	2	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses.
	3	Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses.
	4	Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses.
	5	Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses.
	6	Unable to see at all.
Hearing	1	Able to hear what is said in a group conversation with at least three other people, without a hearing aid.
	2	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people.
	3	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid.
	4	Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.
	5	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.

	6	Unable to hear at all.
Speech	1	Able to be understood completely when speaking with strangers or friends.
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well.
	3	Able to be understood partially when speaking with strangers or people who know me well.
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well.
	5	Unable to be understood when speaking to other people (or unable to speak at all).
Ambulation	1	Able to walk around the neighbourhood without difficulty, and without walking equipment.
	2	Able to walk around the neighbourhood with difficulty; but does not require walking equipment or the help of another person.
	3	Able to walk around the neighbourhood with walking equipment, but without the help of another person.
	4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood.
	5	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood.
	6	Cannot walk at all.
Dexterity	1	Full use of two hands and ten fingers.
	2	Limitations in the use of hands or fingers, but does not require special tools or help of another person.
	3	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person).
	4	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools).

	5	Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools).
	6	Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools).
Emotion	1	Happy and interested in life.
	2	Somewhat happy.
	3	Somewhat unhappy.
	4	Very unhappy.
	5	So unhappy that life is not worthwhile.
Cognition	1	Able to remember most things, think clearly and solve day to day problems.
	2	Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.
	3	Somewhat forgetful, but able to think clearly and solve day to day problems.
	4	Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.
	5	Very forgetful, and have great difficulty when trying to think or solve day to day problems.
	6	Unable to remember anything at all, and unable to think or solve day to day problems.
Pain	1	Free of pain and discomfort.
	2	Mild to moderate pain that prevents no activities.
	3	Moderate pain that prevents a few activities.
	4	Moderate to severe pain that prevents some activities.
	5	Severe pain that prevents most activities.

Exhibit 10.36 HUI3 Single Attribute Utility Functions

Level	Vision	Hearing	Speech	Ambulation	Dexterity	Emotion	Cognition	Pain
1	1	1	1	1	1	1	1	1
2	0.95	0.86	0.82	0.83	0.88	0.91	0.86	0.92
3	0.73	0.71	0.67	0.67	0.73	0.73	0.92	0.77
4	0.59	0.48	0.41	0.36	0.45	0.33	0.7	0.48
5	0.38	0.32	0	0.16	0.2	0	0.32	0
6	0	0		0	0		0	

Exhibit 10.37 HUI3 Multi-Attribute Utility Function on Dead to Healthy Scale

Level	Vision b1	Hearing b2	Speech b3	Ambulation b4	Dexterity b5	Emotion b6	Cognition b7	Pain b8
1	1	1	1	1	1	1	1	1
2	0.98	0.95	0.94	0.93	0.95	0.95	0.92	0.96
3	0.89	0.89	0.89	0.86	0.88	0.85	0.95	0.9
4	0.84	0.8	0.81	0.73	0.76	0.64	0.83	0.77
5	0.75	0.74	0.68	0.65	0.65	0.46	0.6	0.55
6	0.61	0.61		0.58	0.56		0.42	

The multi-attribute utilities are assumed to refer to a chronic condition that will last a lifetime, judged on a scale running from dead (value 0.0) to healthy (value 1.0). To calculate the multi-attribute utility, take the value in Exhibit 10.37 for each attribute (b1 through b8) that corresponds to the person s level on that attribute. The formula is:

$$\text{Utility} = 1.371 (b1 * b2 * b3 * b4 * b5 * b6 * b7 * b8) - 0.371$$

As an example, for someone with a level 2 hearing rating, and a level 3 ambulation problem and level 3 pain, but otherwise healthy:

$$\text{Utility} = 1.371 (0.95 * 0.86 * 0.90) - 0.371 = 0.637$$

Reliability

One-month retest reliability for HUI Mark 3 scores in a population survey application (n = 506) gave an ICC of 0.77 (1, p240). This overall figure masked variations in estimates for individual questions, from kappa = 0.18 to 0.77, and estimates for the eight attribute scores from 0.14 to 0.73 (14). In a study of mothers reporting on their child s health after a 2 to 4-week delay, weighted kappa test-retest reliability ranged from 0.7 to 1.0 for the attribute scores of the Mark 2 instrument (15, Table 3). In a three-day retest study on hospital patients, kappa figures ranged from 0.61 to 0.94, except for Pain, which was 0.48 (16, Table 10).

Grootendorst et al. analysed the impact of proxy versus self-report, and of interviewer versus self-administration on responses using 24,000 survey responses to the HUI3 pain and emotion scales (17). A question on employment status was used as a comparator. Weighted kappa indices for inter-rater agreement were 0.29 and 0.31 for emotion and pain, respectively, compared to 0.82 for employment (17, Table 5). The proxy respondents tended to report less dysfunction than the self-reports. Agreement for mode of administration was 0.53 for emotion and 0.47 for pain, compared to 0.91 for employment status (Table 9). The observed disagreement for the emotion scale was sub-divided into 58% due to random error, 35% due to rater and 7% due to mode of administration. For the pain scale, 31% was due to random error, 10% to rater type and 59% to the mode of administration (17, Table 11). Mathias et al. obtained higher inter-rater agreement in a comparison of ratings by stroke patients and by their caregivers. Kappa values for selected individual questions ranged from 0.37 to 0.80 (18, Table 1), while ICC values for attribute utility scores ranged from 0.39 to 0.81 (Table 2).

Agreement between different sets of raters for the utility scoring task has been reported. A study of 144 ratings made by parents of children with cancer, compared to other parents, showed an ICC of 0.99 between their judgments; the mean difference was only 0.018 (19). In another study, correlations among three observers were 0.8 or above (20). A cautionary note has been raised, however, by Feeny et al., who report that using HUI scores to reflect utilities as measured directly by the standard gamble is appropriate at a group level, but that considerable variability occurs at the individual level so that HUI scores are not a good substitute for directly measured utility scores at the individual level (21).

Validity

Face validity. Empirically, all levels of each attribute have been represented by responses in population surveys (3; 13, p515). The independence of the attributes in the HUI3 was supported in analyses of a community health survey. The highest Kendall correlation between attributes was 0.35; only 4 of 28 correlations were 0.20 or greater, compared to 10 of 10 equivalent correlations for the EuroQol EQ-5D) (9). In a small qualitative study of content validity, the HUI3 omitted many aspects of quality of life noted as being important to patients with Alzheimer s disease, and had a less comprehensive coverage than the Quality of Well-Being scale (22).

Discriminant evidence. Three of the six HUI2 subscales, and the overall utility score, discriminated significantly between children on active cancer therapy and those on follow-up care (15, Table 5). Using self-report data from a population survey, and after adjustment for demographic factors and number of chronic conditions, mean HUI3 scores were 0.538 for people who had had a stroke, 0.765 for people with arthritis, and 0.925 for people with neither (23, p294). HUI2 scores distinguished between children aged eight years who had been born with a birth-weight and normal birth-weight children (24).

HUI2 and 3 scores discriminate between severity levels of patients with Alzheimer s disease. Mark 2 scores ranged from 0.73 for borderline cases to 0.14 for those with advanced disease (25); Mark 3 scores declined from 0.47 for borderline to -0.23 for terminal Alzheimer s disease (26). Similarly, HUI scores varied across severity levels of Parkinson s disease (27, Table 2). The HUI2 and HUI3 were compared in a study of diabetes; the HUI3 provided better discrimination of more severe states than did the HUI2 (28; 29, Tables 4-6).

Mark 3 scores showed significant differences between children before, during and after receiving chemotherapy; the effect size was 2.8 (3, pp380-381). However, in a study of rheumatology patients who rated their own change in health over a 12-month period, the HUI showed a substantially smaller effect size than the EQ-5D, and slightly smaller than that of the SF-6D (30, Table 4). In a study of effect sizes for hip replacements, the effect size for the HUI2 was 1.00; that for the HUI3 was 1.19. Effect sizes for the pain scales were 1.0 and 1.3 for the HUI2 and HUI3. These findings were lower than the corresponding figures for the SF-36 (31, Table 2).

Concurrent validity. Gold et al. derived scores that approximated the Mark 1 HUI from data collected in the 1982 National Health Examination Survey (n = 10,163) (32). The HUI scores correlated 0.52 with a simple, 3-point self-rating of health. As expected, both measures were associated with age, educational level, socioeconomic status and the presence of chronic health conditions. However, despite the much finer gradation of the HUI, the statistical significance of the associations between the self-rated health question and most chronic conditions was stronger than that for the HUI (32, Table 3). This close correspondence between the HUI and an overall health rating has been replicated (33, Table 1). Using data from a Canadian national survey, Kopec et al. reported associations between HUI scores and self-rated health, activity restrictions, drug use, the presence of chronic conditions, age, and income level (34).

Because of their similarity, several studies have compared the HUI3 to the EuroQol EQ-5D. In a community sample of 1,477 respondents, Bélanger et al. compared the two scales to a visual analog scale (VAS) rating your health today (8). The HUI overall scores correlated 0.69 with the EQ-5D scores, and 0.63 with the VAS (8, Table 8). Correlations between scores for comparable dimension on the HUI and EQ-5D included 0.56 for pain; 0.42 for emotion; and 0.47 for mobility (8, Table 10). Kendall correlations confirmed the independence of the HUI attributes; only two of 28 correlations exceeded 0.25 (Table 12). The area under the ROC curve for the HUI in distinguishing respondents who had been hospitalized in the past year was 0.66, virtually identical to the 0.64 for the EQ-5D and 0.66 for the VAS (Table 16). Areas under the ROC curve for detecting the presence of chronic conditions were 0.70 for the HUI, 0.71 for the EQ-5D and 0.69 for the VAS; for activity restrictions the values were 0.83 for the HUI, 0.84 for the EQ-5D and 0.81 for the VAS (8, Table 16). A statistical comparison of HUI and EQ-5D scores for a range of sub-populations suggested that the two scales provided virtually identical ratings for all but the most healthy groups (8, pp43-46). A Dutch study obtained a rank correlation of 0.90 between the scores on the EQ-5D and the HUI, although the mean scores on the HUI were 0.16 points lower than those for the EQ-5D for six case descriptions (35, Table 2). Also remarkable in that study was the contrast in preference scores between the HUI-2 and the HUI-3 (Table 2). By contrast, an American study found that mean scores for the HUI were 0.11 points *higher* than the EQ-5D scores. In this study, the intraclass correlation between scores was 0.49 before treatment, 0.66 one month after treatment and 0.78 at 3 months; the ICC between change scores for the two scales over time was only 0.30 (36, p597). In a study of Parkinson patients, the correlation between the HUI and the EQ5-D was 0.74; the correlations between the EQ5-D and measures of disease severity and quality of life were in general higher than those for the HUI (27, p105). For rheumatology patients, the correlation between HUI and EQ-5D was 0.68 (30, Table 3).

Spearman coefficients comparing HUI2 scores to the Pediatric Oncology Quality of Life (POQOL) scale for children with cancer ranged from -0.41 (self-care dimension), to -0.51 (pain) and -0.53 (emotion). Equivalent correlations with the Child Behavior Checklist were -0.48 (emotion) and -0.29 (pain) (15, Table 4). In the same study, the VAS score for the HUI showed a stronger correlation with the POQOL overall score, at -0.70, than did the HUI utility score, at -0.44 (15, p427). Correlations between HUI2 overall scores and the SF-12 were 0.60 for the SF-12 physical score, and 0.59 for the mental score; equivalent figures for the HUI3 were 0.64 and 0.60 (29, Tables 7 and 8).

Predictive validity. In Gold s study, HUI1 scores were compared to health outcomes recorded five years later. Baseline HUI1 scores were associated with subsequent declines in health, with hospital admissions and mortality. However, the gradient in scores across these outcome categories was steeper for the self-rating of health than for the HUI scores (32, Table 4). Gold concluded that the two measures were complementary in predicting the outcomes, but that they did not fully cover the waterfront of domains of HRQOL (32, p174).

The responsiveness of the HUI3 to changes in health status over two years was tested in a population survey (n = 81,804) (33). Baseline HUI3 scores discriminated between those who remained healthy and those who were subsequently hospitalized, or those who became restricted in activities; it also identified those who developed a severe chronic illness, but not those who developed a moderate condition (33, p566). For each of these outcomes, the standardized response means for HUI3 were comparable to, but slightly lower than, those of a single-item overall health rating (33, Tables 5-7). A before- and after-study of hip replacement patients showed the HUI3 to be more responsive to improvement in health status than the HUI2 or the SF-36, but less so than disease-specific instruments such as the Harris Hip Scale, the WOMAC or the MACTAR (31).

Alternative Forms

A questionnaire that incorporates the Mark 3 version is shown in the Appendix of the book by Patrick and Erickson (5, pp385-389).

The HUI web site provides information on translations of questionnaires, which are available in English, French, Spanish, Italian, German, Dutch and Japanese, with Polish, Swedish, Norwegian, Finnish and Danish versions planned. Validity information is available on French translations (16; 37-39), and Canadian French (15). A Spanish version has a reported alpha of 0.76 and a correlation of 0.79 with the EQ-5D (40).

Reference Standards

The HUI3 has been used in several Canadian national surveys, and reference scores are available for community and institutionalized populations by age and sex (3, p381).

Furlong et al. quote mean scores for various patient groups (3, Table 2), and reproduce the frequency distributions for each attribute from a French study (Table 3).

Commentary

The HUI has received widespread attention and frequent use in survey research, especially in Canada. The basic articles describing the instrument give a clear indication of the steps taken in developing the scale, especially the utility scaling component. Indeed, the impression is that the authors are more interested in the utility scaling procedures than in question wording and content validity. The web site (www.fhs.mcmaster.ca/hug/index.htm) contains basic information on the HUI, an extensive bibliography, and order forms for unpublished technical reports which are available at cost. The HUI scoring function has been applied to various other data collection methods, such as the Minimum Data Set (41) and the SF-12 (42).

The wide range of states of poor health make it unlikely that the HUI will suffer floor problems. However, the HUI deliberately focuses on capacity; the arguments in support of this are cogent, but users should recognize that this gives it a narrower scope than instruments such as the SF-36. It may also lead to ceiling problems in that the HUI may not be sensitive to variations at the positive end of the health continuum. This may restrict its application in prognostic studies and in measuring outcomes where performance above mere competence is relevant. , although many other instruments also suffer ceiling problems (1, p243). For example, comparisons with the EQ-5D suggest similar performance on validity tests, although the HUI3 is more sensitive to variation at the healthy end of the spectrum and so may be more suitable for use in community surveys. This derives from a limitation of the EQ-5D scoring, in which it is not possible to achieve a score between 0.88 and 1.00; of respondents rated 1.0 by the EQ-5D, the HUI3 rated many as falling between 0.9 and 0.99 (8, p24). On a technical note, Austin et al. described a modified regression approach, Tobit analysis, which can improve estimates in the presence of ceiling effects (43; 44).

As with other health measures, it seems probable that the various modes of administering the HUI (personal interviews, telephone, or mail) may not provide equivalent results (45). Reliability results suggest that variations may be greatest for the pain and emotion scales (17). Studies making more than one administration of the HUI should probably use the same mode of administration on each occasion.

A limitation of the EuroQol EQ-5D may be that it has relatively few response steps for each item, making it insensitive to minor variations in health. By contrast, descriptions of the HUI3 often cite as an advantage the fact that it can rate 972,000 unique health states. However, this may have limited practical relevance, as very many fewer actually occur. In the 1990 Ontario Health Survey (n = 68,394), a total of 1,755 states were used, but 129 states accounted for almost 95% of responses. In the 1991 Canadian General Social Survey (n = 11,567), only 12 states had a prevalence greater than 1% of respondents; together these 12 accounted for 75% of all responses. Nonetheless, Feeny et al. note that several states were identified that would not have been picked up by the EuroQol or by Rosser s Disability Index (1, p244). It also remains somewhat disconcerting that several studies have indicated that a single summary question or a visual analogue scale can provide discrimination results as good as those of the HUI3 and the EQ-5D.

The utility scoring system has applications beyond assessing individual health status. In a modelling study, for example, typical health outcomes for a therapy can be taken from published reports, mapped onto the HUI classification system and then scored, permitting cost-utility analyses. The preference scores can also be used in calculating quality-adjusted life expectancy (13, p517).

Different health indexes take different approaches to estimating preferences, and the debate is active. Feeny et al. recommend utilities rather than values, arguing that as future health is uncertain, utilities are the appropriate measure of preference. By contrast, the Quality of Well-Being scale used values derived from visual analog scaling tasks, while the EuroQol used VAS and time trade-off methods. Feeny et al. shows empirically how the approaches differ in their relative weighting of major and minor disabilities: values pay relatively greater attention to minor disabilities compared to utilities (1, Table 7). In terms of the multi-attribute scoring function, the HUI uses a multiplicative model while the Quality of Well-Being scale uses the simpler additive model. Torrance et al. argue that the additive model tends to underestimate the impact of that disability in situations where a person has a single, severe disability (13, p513). There remains disagreement over methodological details of how best to estimate utilities, but the team developing the HUI have contributed significantly to pushing this field forward.

In the context of the debate over health indexes versus health profile measures, some information has been collected in a comparison between the HUI and the SF-6D. The latter is a re-scored version of the Short-Form-36 that provides utility scores. Although both used variants of a standard gamble approach to derive utility weights, the agreement between the two methods is low. This may in part reflect the contrast in content: the HUI focuses on within a the skin definition of health, whereas the SF questions cover handicap and function. It may also reflect methodological differences in the way weights were derived (46, p980). The implication is that we should not yet consider utility scores to be equivalent across instruments; further work is required to standardize the utility metric and also to indicate which health index provides superior results.

Address

William Furlong, Health Utilities Inc., 88 Sydenham Street, Dundas, Ontario L9H 2V3, Canada.
E mail: huinfo@healthutilities.com

References

(1) Feeny D, Torrance GW, Furlong WJ. Health Utilities Index. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. Philadelphia: Lippincott-Raven, 1996: 239-252.

(2) Feeny D, Furlong WJ, Boyle M, Torrance GW. Multi-attribute health status classification systems: Health Utilities Index. Pharmacoecon 1995; 7:490-502.

(3) Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. Ann Med 2001; 33:375-384.

(4) Richardson CG, Zumbo BD. A statistical examination of the Health Utility Index-Mark III as a summary measure of health status for a general population health survey. Soc Indicat Res 2000.

(5) Patrick DL, Erickson P. Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford University Press, 1993.

(6) Cadman D, Goldsmith C. Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. J Chronic Dis 1986; 39:643-651.

(7) Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. J Clin Oncol 1992; 10:923-928.

(8) Bélanger A, Berthelot J-M, Guimond E, Houle C. A head-to-head comparison of two generic health status measures in the household population: McMaster Health Utilities Index (Mark3) and the EQ-5D. Statistics Canada Report , 1-62. 2002. Ottawa, Canada.

(9) Houle C, Berthelot J-M. A head-to-head comparison of the Health Utilities Index Mark 3 and the EQ-5D for the population living in private households in Canada. Quality of Life Newsletter 2000; 24:5-6.

(10) Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S et al. Multiattribute and single-attribute utility functions for the Health Utilitiws Mark 3 system. Med Care 2002; 40:113-128.

(11) Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.

Med Care 1996; 34:702-722.

(12) Le Galès C, Buron C, Costet N, Rosman S, Slama G. Development of a preference-weighted health status classification system in France: the Health Utilities Index 3. Health Care Manag Sci 2002; 5:41-51.

(13) Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions: Health Utilities Index. Pharmacoecon 1995; 7:503-520.

(14) Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index--Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995; 4:249-257.

(15) Trudel JG, Rivard M, Dobkin PL, Leclerc J-M, Robaey P. Psychometric properties of the Health Utilities Index Mark 2 system in paediatric oncology patients. Qual Life Res 1998; 7:421-432.

(16) Costet N, Le Gales C, Buron C, Kinkor F, Mesbah M, Chwalow J et al. French cross-cultural adaptation of the health utilities indexes Mark 2 (HUI2) and 3 (HUI3) classification systems. Qual Life Res 1998; 7(3):245-256.

(17) Grootendorst PV, Feeny DH, Furlong W. Does it matter whom and how you ask? Inter- and intra-rater agreement in the Ontario Health Survey. J Clin Epidemiol 1997; 50:127-135.

(18) Mathias SD, Bates MM, Pasta DJ, Cisternas MG, Feeny D, Patrick DL. Use of the Health Utilities Index with stroke patients and their caregivers. Stroke 1997; 28:1888-1894.

(19) Wang Q, Furlong W, Feeny D, Torrance G, Barr R. How robust is the Health Utilities Index Mark 2 utility function? Med Decis Making 2002; 22(4):350-358.

(20) Gemke RJBJ, Bonsel GJ. Reliability and validity of a comprehensive health status measure in a heterogeneous population of children admitted to intensive care. J Clin Epidemiol 1995.

(21) Feeny D, Blanchard C, Mahon JL, Bourne R, Rorabeck C, Stitt L et al. Comparing community-preference-based and direct standard gamble utility scores: Evidence from elective total hip arthroplasty. International Journal of Technology Assessment In Health Care 2003; 19(2):362-372.

(22) Silberfeld M, Rueda S, Krahn M, Naglie G. Content validity for dementia of three generic preference based health related quality of life instruments. Qual Life Res 2002; 11:71-79.

(23) Grootendorst P, Feeny D, Furlong W. Health Utilities Mark 3. Evidence of construct validity for stroke and arthritis in a population health survey. Med Care 2000; 38:290-299.

(24) Saigal S, Feeny D, Furlong WJ, Rosenbaum P, Burrows E, Torrance GW. Comparison of the health-related quality of life of extremely low birth weight children and a reference group of

children at age eight years. J Pediatrics 1994; 125:418-425.

(25) Neumann PJ, Kuntz KM, Leon J, Araki SS, Hermann RC, Hsu MA et al. Health utilities in Alzheimer's disease: a cross-sectional study of patients and caregivers. Med Care 1999; 37:27-32.

(26) Neumann PJ, Sandberg EA, Araki SS, Kuntz KM, Feeny D, Weinstein MC. A comparison of HUI2 and HUI3 utility scores in Alzheimer's disease. Med Decis Making 2000; 20:413-422.

(27) Siderowf A, Ravina B, Glick HA. Preference-based quality-of-life in patients with Parkinson's disease. Neurology 2002; 59:103-108.

(28) Maddigan SL, Feeny DH, Johnson JA. A comparison of the health utilities indices mark 2 and mark 3 in type 2 diabetes. Med Decis Making 2003; 23(6):489-501.

(29) Maddigan SL, Feeny DH, Johnson JA. Construct validity of the RAND-12 and Health Utilities Index Mark 2 and 3 in type 2 diabetes. Qual Life Res 2004; 13:435-448.

(30) Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. Med Care 2003; 41:791-801.

(31) Blanchard C, Feeny D, Mahon JL, Bourne R, Rorabeck C, Stitt L et al. Is the Health Utilities Index responsive in total hip arthroplasty patients? J Clin Epidemiol 2003; 56:1046-1054.

(32) Gold M, Franks P, Erickson P. Assessing the health of the nation. The predictive validity of a preference-based measure and self-rated health. Med Care 1996; 34:163-177.

(33) Kopec JA, Schultz SE, Goel V, Williams JI. Can the Health Utilities Index measure change? Med Care 2001; 39:562-574.

(34) Kopec JA, Williams JI, To T, Austin PC. Measuring population health: correlates of the Health Utilities Index among English and French Canadians. Can J Public Health 2000; 91:465-470.

(35) Oostenbrink R, Moll HA, Essink-Bot M-L. The EQ-5D and the Health Utilities Index for permanent sequelae after meningitis: a head-to-head comparison. J Clin Epidemiol 2002; 55:791-799.

(36) Bosch JL, Hunink MGM. Comparison of the Health Utilities Index Mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudication. Qual Life Res 2000; 9:591-601.

(37) Costet N, Le Galès C, Buron C, Kinkor F, Mesbah M, Chwalow J et al. French cross-cultural adaptation of the Health Utilities Indexes Mark 2 (HUI2) and 3 (HUI3) classification

systems. Clinical and Economic Working Groups. Qual Life Res 1998; 7:245-256.

(38) Le GC, Buron C, Costet N, Rosman S, Slama PR. Development of a preference-weighted health status classification system in France: the Health Utilities Index 3. Health Care Manag Sci 2002; 5:41-51.

(39) Le Galès C, Costet N, Gentet JC, Kalifa C, Frappaz D, Edan C et al. Cross-cultural adaptation of a health status classification system in children with cancer. First results of the French adaptation of the Health Utilities Index Marks 2 and 3. Int J Cancer Suppl 1999; 12:112-118.

(40) Ruiz M, Rejas J, Soto J, Pardo A, Rebollo I. Adaptation and validation of the Health Utilities Index Mark 3 into Spanish and correction norms for Spanish population. Medicina Clinica 2003; 120:89-96.

(41) Wodchis WP, Hirdes JP, Feeny DH. Health-related quality of life measure based on the minimum data set. International Journal of Technology Assessment In Health Care 2003; 19(3):490-506.

(42) Franks P, Lubetkin EI, Gold MR, Tancredi DJ. Mapping the SF-12 to preference-based instruments - Convergent validity in a low-income, minority population. Med Care 2003; 41(11):1277-1283.

(43) Austin PC, Escobar M, Kopec JA. The use of the Tobit model for analyzing measures of health status. Qual Life Res 2000; 9:901-910.

(44) Grootendorst P. Censoring in statistical models of health status: what happens when one can do better than '1'? Qual Life Res 2000; 9:911-914.

(45) Verrips GH, Stuifbergen MC, den Ouden AL, Bonsel GJ, Gemke RJ, Paneth N et al. Measuring health status using the Health Utilities Index: agreement between raters and between modalities of administration. J Clin Epidemiol 2001; 54:475-481.

(46) O'Brien BJ, Spath M, Blackhouse G, Severens JL, Dorian P, Brazier J. A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index. Health Econ 2003; 12(11):975-981.