International Journal of Technology Assessment in Health Care

cambridge.org/thc

Commentary

Ethics approval and consent to participate: Not applicable. *Availability of data and materials*: Not applicable. *Competing interests:* It should be noted that all three authors have a proprietary interest in Health Utilities Incorporated, Dundas, Ontario, Canada. HUInc. distributes copyrighted Health Utilities Index (HUI) materials and provides methodological advice on the use of the HUI. *Funding:* There was no funding source for this study. *Authors' contributions:* All three authors contributed to the conceptualization and writing of the study and have approved the final version of the manuscript. The authors acknowledge the constructive comments by the Editor, Deputy Editor, Associate Editor, and two reviewers, which have improved the study.

Cite this article: Feeny D, Furlong W, Torrance GW (2019). Commentary. In Praise of Studies That Use More Than One Generic Preference-Based Measure. *International Journal of Technology Assessment in Health Care* **35**, 257–262. https://doi.org/10.1017/S0266462319000412

Received: 19 March 2018
Revised: 11 May 2019
Accepted: 27 May 2019
First published online: 12 July 2019

Key words:
Comparative effectiveness research; Cost-effectiveness analysis; Cost-utility analysis; EuroQol EQ-5D; Generic preference-based measures; Health-related quality of life; Health technology assessment; Health Utilities Index; Multi-attribute utility measures; Outcome measures; Quality-adjusted life-years; Quality of Well Being Scale; Short-Form 6D

Author for correspondence:
David Feeny,
E-mail: feeny@mcmaster.ca

CAMBRIDGE UNIVERSITY PRESS

Commentary. In Praise of Studies That Use More Than One Generic Preference-Based Measure

David Feeny[1], William Furlong[2] and George W. Torrance[3]

[1]Department of Economics and Centre for Health Economics and Policy Analysis, McMaster University, Hamilton, ON Canada; Health Utilities Incorporated, Dundas ON Canada; [2]Centre for Health Economics and Policy Analysis, McMaster University, Hamilton, ON Canada; Health Utilities Incorporated, Dundas ON Canada and [3]Centre for Health Economics and Policy Analysis, McMaster University, Hamilton, ON Canada; Health Utilities Incorporated, Dundas ON Canada

Abstract

Objectives and Background. Generic preference-based (GPB) measures of health-related quality of life (HRQL) are widely used as outcome measures in cost-effectiveness and cost-utility analyses (CEA, CUA). Health technology assessment agencies favor GPB measures because they facilitate comparisons among conditions and because the scoring functions for these measures are based on community preferences. However, there is no gold standard HRQL measure, scores generated by GPB measures may differ importantly, and changes in scores may fail to detect important changes in HRQL. Therefore, to enhance the accumulation of empirical evidence on how well GPB measures perform, we advocate that investigators routinely use two (or more) GPB measures in each study.

Methods. We discuss key measurement properties and present examples to illustrate differences in responsiveness for several major GPB measures across a wide variety of health contexts. We highlight the contributions of longitudinal head-to-head studies.

Results. There is substantial evidence that the performance of GPB measures varies importantly among diseases and health conditions. Scores are often not interchangeable. There are numerous examples of studies in which one GPB measure was responsive while another was not.

Conclusions. Investigators should use two (or more) GPB measures. Study protocols should designate one measure as the primary outcome measure; the other measure(s) would be used in secondary analyses. As evidence accumulates it will better inform the relative strengths and weaknesses of alternative GPB measures in various clinical conditions. This will facilitate the selection and interpretation of GPB measures in future studies.

The key message of this Commentary is a recommendation that studies should use two or more generic preference-based (GPB) measures. This Commentary is aimed at investigators who design, execute, report, and interpret health technology assessments (HTA) and economic evaluations of healthcare technologies. This Commentary is also aimed at the policy makers who use evidence from HTA studies in decision making. Standardizing on a single GPB across studies is attractive because it seemingly enhances comparability of results. However, there is no gold standard GPB measure. Because a GPB measure may not be responsive in a particular context, relying on a single GPB can be perilous. Within studies, a single GPB measure provides an estimate of effectiveness that is precise but of unknown accuracy. Using more than one measure provides important additional information on accuracy, and contributes to the rapid accumulation of evidence on the contexts in which particular measures perform well and situations in which they do not. Additional rationales are outlined below.

Background

Establishing confidence in a result often demands more than one criterion. GPB measures generate overall summary scores that are widely used to estimate health outcomes in clinical trials and observational studies; see, for instance, Feeny and colleagues (1). GPB measures are also used in clinical practice for quality improvement and to assist in the management of individual patients; see, for instance, Santana and colleagues (2). In addition, GPB measures are included in population health surveys; see for instance, Fryback and colleagues (3).

A major use of GPB measures is in economic evaluation: cost-effectiveness and cost-utility analyses (CEA, CUA). CEA and CUA are important components of HTA. Evidence from CEA/CUA studies often plays an important role in decision making about the adoption and usage of healthcare technologies. Indeed, CUA submissions are required by agencies in

several countries, including England (4), Canada (5), Australia (6;7), and Japan (8). The measure of health effects favored by each of these agencies is the quality-adjusted life-year, the QALY. QALYs combine the effects of an intervention on mortality and morbidity and are estimated by multiplying the utility score for a health state, its desirability in terms of health-related quality of life (HRQL), by the duration of that health state.

Each of these HTA agencies recommends the use of generic preference-based (GPB) measures as the source of utility scores for computing QALYs. The Canadian Agency for Drugs and Technologies in Health (CADTH) and the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia recommend the use of GPB measures but do not identify a particular preferred GPB measure. Similarly, the 2nd Panel on Cost Effectiveness in Health and Medicine endorses the use of GPB for the reference case analyses (9;10) The EQ-5D is preferred by The National Institute for Health and Care Excellence (NICE) (4, p 39). Examples of widely used GPB measures are EQ-5D (11), Health Utilities Index (HUI) Mark 2 (HUI2) and HUI Mark 3 (HUI3) (12;13), the Quality of Well Being Scale (QWB) (14), and the Short-Form 6D (SF-6D) (15;16). Table 1 provides a brief summary of the characteristics of nine prominent GPB measures.

GPB measures are favored for numerous reasons including their broad coverage of components of physical and mental health, capturing comorbidity and unintended effects of interventions; being brief with little burden on respondents and research staff; being applicable across a wide variety of diseases and conditions, allowing for comparisons of the comprehensive burdens of disease and effects of interventions across conditions; and using scoring systems based on community preferences. GPB measures are also highly relevant for investigators when selecting a health-related quality of life (HRQL) outcome measure even if they do not contemplate doing a CEA or CUA.

All major GPB measures produce scores on the standard health scale where dead has a score of zero and perfect/full health as a score of one. Despite this, it is widely understood that scores from different major GPB measures are often not interchangeable (17;18). Gamst-Klaussen and colleagues (17) review important reasons for this, including the attributes (dimensions, domains) of health status included in the measure, different methods for obtaining preference scores with which to estimate multi-attribute utility (scoring) functions, and the choice of different functional forms and estimation methods for creating scoring functions. Gamst-Klaussen and colleagues discuss various "mapping" or "cross-walk" approaches that provide evidence of relationships between scores from different GPB measures and the importance of head-to-head comparisons of instruments. While the mapping and cross-walk approaches focus on the relationships among measures and how to attempt to make scores from different measures more commensurate, the head-to-head comparisons approach is more inductive. In particular, focusing on longitudinal studies, this Commentary investigates whether measures perform well, or not, depending upon clinical context or population. Are major GPB measures equally responsive? Does responsiveness vary among measures within contexts? Should investigators rely on a single GPB measure for primary data collection in longitudinal studies? Below we provide a non-systematic review of several longitudinal head-to-head comparison studies that used two or more GPB measures and thus provide important illustrative information on the performance of GPB measures in a wide variety of clinical areas and contexts. Another criteria influencing the choice of case studies for review is the inclusion of health conditions and diseases that affect a wide range of attributes with varying degrees of severity. The illustrative studies are not intended to represent the universe of published results or be a random selection of such results. It should be noted that, given that the authors of the Commentary are among the developers of the HUI2 and HUI3, there is a tendency to focus on head-to-head studies that included HUI.

A Brief Primer on Measurement Properties

Context specific evidence on measurement properties is a key criterion that should influence the choices of GPB measures for a HTA study (19;20). Because many HTA studies extract GPB measure scores from the existing literature, evidence on cross-sectional construct validity of the GPB measure in that context is important. Does the measure capture the construct that it is supposed to capture? Does it distinguish among known groups? Does it identify the level of severity? For clinical trials and other longitudinal studies, there is an additional a key criterion: responsiveness (or longitudinal construct validity). Does the measure capture meaningful change when it occurs?

In a particular context a GPB measure may not be responsive because it omits an attribute for which there is important change in the context, it is subject to floor or ceiling effects, the responsiveness of a measure is attenuated due to the limited number of intermediate levels within its attributes, or other measurement issues. Floor effects occur when the range of a measure is insufficient for capturing higher degrees of impairment. Ceiling effects occur when the range of the measure is insufficient for capturing lower degrees of impairment. Examples discussed below illustrate the importance of these (and other) criteria.

Effectiveness in CEA/CUA is assessed by estimating the effects of the intervention both on the HRQL and longevity of those affected by the intervention. We define effectiveness with respect to assessing the effects on HRQL as follows. An intervention is deemed to be effective if it produces a clinically important difference (CID) in HRQL relative to its comparator. The concepts of CID, that is, minimal CID, patient-important difference, and the methods for developing empirical guidance on the threshold for CID, are discussed in Guyatt and colleagues (21) and Schünemann and Guyatt (22). An important component of this approach is that for the change to be important it must be both noticeable and important to patients, a patient-centric viewpoint.

Examples of Contributions by Studies That Use Multiple GPB Measures

Lack of Interchangeability

An example of studies using more than one GPB measure is found in two papers by Marra and colleagues (23;24). A study of 313 rheumatoid arthritis (RA) patients being treated by one of eight rheumatologists examined the construct validity, reliability and responsiveness of four major GPB measures: EQ-5D-3L, HUI2, HUI3, and SF-6D. Numerous disease-specific measures were also used (23). The authors note that overall scores differed substantially among the GPB measures, lack of interchangeability, and conclude that "each of the instruments were well-accepted, the overall scores are all able to distinguish between groups defined by measures of RA severity" (23, pp 1580–1581).

The longitudinal component of the study by Marra and colleagues (24) examined test-retest reliability and responsiveness of the GPB measures in the same cohort of RA patients. The authors

https://doi.org/10.1017/S0266462319000412 Published online by Cambridge University Press

Table 1. Brief Descriptions of Nine Prominent Generic Preference-Based Measures

Name of generic preference-based measure	Number of attributes	Number of levels per attribute	Preference scoring type	Range of scores[a]
AQoL-8D [36; 37; 38]	8	4 to 6	TTO	-0.04 to 1.00
EQ-5D-3L [11; 38]	5	3	TTO	-0.59 to 1.00[b]
EQ-5D-5L [38; 39]	5	5	TTO	-0.285 to 1.00[c]
HUI2 [12; 38]	7	3 to 5	VAS and SG	-0.03 to 1.00
HUI3 [13; 38]	8	5 or 6	VAS and SG	-0.36 to 1.00
PROMIS [40]	7	Many	SG	-0.022 to 1.000
QWB-SA [14; 37; 38]	4	3 to 27	VAS	0.32 to 1.00
SF-6D [15; 16; 37]	6	4 to 6	SG	0.20 to 1.00
15D [38; 41]	15	4 or 5	VAS	0.11 to 1.00

Notes.
[a]The range of scores reports the minimum and maximum scores generated by the measure on a scale in which the score for dead = 0.00 and the score for perfect/full health = 1.00.
[b]United Kingdom.
[c]England.

conclude that test-retest reliability was acceptable except for EQ-5D-3L (24, p 1341), that EQ-5D-3L was the most responsive measure in detecting worsening while HUI3 and SF-6D "were more superior in detecting improvement" (24, p 1342), and that HUI3 yield the largest change score while SF-6D yields the smallest (24, p 1342). The authors recommend HUI3 and SF-6D for GPB measures in clinical trials of RA (24, pp 1342–1343).

Fryback and colleagues (25) compare the performance of major GPB measures in a population health survey. They note that while there appears to be a common core of physical and mental health reflected in all of the measures, the measures are also clearly not interchangeable.

Differences in Responsiveness

A study of teenagers with sub-threshold depression or depression as compared to a reference group of teens without depression provides useful information on the cross-sectional construct validity and responsiveness of several GPB measures including EQ-5D-3L, HUI2, HUI3, SF-6D, and the QWB-SA (26;27). With respect to known-groups construct validity all of the GPB measures performed well, although Lynch and colleagues (26) noted that the duration of the interviews to complete the QWB-SA was substantially greater than for the other GPB measures. With respect to responsiveness, again all of the measures performed well. Furthermore, Dickerson and colleagues (27, p 452) report that HUI3, SF-6D, and the QWB-SA were "among the most responsive measures while EQ-5D-3L was among the least responsive".

Differences in Responsiveness

Langfitt and colleagues (28) compared four GPB measures (EQ-5D-3L with the U.K. and U.S. scoring systems; HUI2; HUI3; and SF-6D) in a longitudinal study of patients with chronic epilepsy. They report a substantial ceiling effect, a rate of 34 percent, for EQ-5D-3L, compared with < 10 percent for the other measures. Results for responsiveness across measures were mixed. Only changes in SF-6D and HUI3 were associated with improvements in seizure control. Langfitt and colleagues (28) conclude that SF-6D showed advantages in the study because SF-6D captured both the physical health consequences of seizures as well as the effects of seizures on social functioning.

Differences in Responsiveness

In a natural history study of recovery after stroke, Pickard and colleagues (29) found that EQ-5D-3L and HUI3 were much more responsive than HUI2 and SF-6D. Among survivors, the mean gain in HRQL score registered during the six-month follow-up period was 0.31 for EQ-5D-3L (U.K. scoring system), 0.24 for EQ-5D-3L (U.S. scoring system), 0.25 for HUI3, 0.13 for HUI2, and 0.13 for SF-6D.

Lack of Coverage of an Important Attribute

In a study comparing a cohort of patients before and after cataract surgery, there were clinically important increases in vision-specific measures and in overall HUI2 and HUI3 scores, while there were no important changes in scores for the EQ-5D-3L, SF-6D, and QWB-SA (30).

Lack of Responsiveness

EQ-5D-3L, HUI2, HUI3, QWB-SA, and SF-6D were used in a prospective cohort study of patients with congestive heart failure referred to a specialty clinic for care. According to a condition-specific measure, the Minnesota Living With Heart Failure instrument, clinically important improvement occurred in the cohort over the 6-month period from referral to follow-up. However, only three of the five GPB measures, that is, HUI3, QWB-SA, and SF-6D, showed clinically important improvement at the cohort level (30;31).

Floor Effects/Change Scores

HUI2, HUI3, SF-6D, the standard gamble (SG), and several specific measures were used in a longitudinal study of patients waiting for and undergoing elective total hip arthroplasty (THA) (1). Among the four utility-based measures, SF-6D, HUI2, and HUI3 had the same order of magnitude of responsiveness, while the SG was less responsive. However, at the cohort level the four utility-based measures provided importantly different estimates of the gain in HRQL associated with THA, the mean difference between pre- and postsurgery scores. The difference was 0.10 for SF-6D; 0.16 for the SG; 0.22 for HUI2; and 0.23 for HUI3. Although patients reported improvements in physical functioning and reductions in bodily pain, floor effects associated with SF-6D (20) led to a much lower estimate of overall improvement.

https://doi.org/10.1017/S0266462319000412 Published online by Cambridge University Press

Specific Versus Generic Measures of HRQL

Generic measures are applicable across a wide variety of populations and health conditions and thus enable broad comparisons. GPB measures are a class of generic measures with scoring systems based on preferences for health states. Furthermore, there is guidance in the literature on CIDs for the major GPB (see, for instance, Feeny and colleagues) (30). But it is often the case that specific measures are more responsive than generic measures (24;32). Thus, there is the potential for disagreement between results based on GPB and specific measures. In the absence of a gold standard measure of HRQL, how should such cases be resolved? If both GPB and specific measures indicate that the intervention is not effective, one would conclude that the intervention is not effective. Similarly, if both indicate that it is effective, one would conclude that the intervention is effective.

It could be the case that the GPB indicates that the intervention is harmful while the specific measure indicates that the intervention is effective. This could result because of side-effects of the intervention that are not included in the specific measure (in arthritis an anti-inflammatory drug may reduce pain and swelling but result in gastro-intestinal distress) or an interaction with a comorbidity. The GPB provides overall information on the outcome: on net are patients better or worse off? It could also be the case that, while the specific measure indicates that the intervention is effective, the GPB does not. This could be the result of a lack of responsiveness of the GPB chosen. The use of two or more GPB measures reduces the risk of such an outcome.

Discussion

Inductive Propositions

It is useful to distill some generalizations from the case studies described above. For instance, virtually all GPB measures include an attribute assessing emotional health. For teen depression, Lynch and colleagues and Dickerson and colleagues found that all of the measures included performed well with respect to cross-sectional construct validity (26, 27). However, with respect to responsiveness, HUI3, SF-6D, and the QWB-SA were the most responsive.

Given that HUI3 includes vision while EQ-5D and SF-6D do not, it is not surprising that HUI3 is more responsive in the cataract surgery study. An implication of these results is that to obtain valid preference-based scores it is necessary to use a measure that includes the key attributes important in that context. Another implication is that investigators need to consider the potential for floor and/or ceiling effects. SF-6D was responsive in the elective total hip arthroplasty study but under estimated the gain in overall HRQL.

In the context of chronic epilepsy, Langfitt and colleagues reported substantial ceiling effect issues for EQ-5D-3L and noted that only changes in SF-6D and HUI3 were associated with improvements in seizure control. Langfitt and colleagues recommend the use of SF-6D (28).

The attributes included in the GPB measures and the range of health status covered by those measures affects their performance. Results from the non-systematic review described above highlights the lack of interchangeability among these measures and illustrate contexts in which particular measures perform well or do not perform well.

These results have important policy implications in the context of HTA. There were substantial differences in the estimated change in HRQL among measures in the elective total hip arthroplasty and stroke studies. Results based on some of the measures would be quite favorable to the adoption and usage of elective total hip arthroplasty, while results based on other measures would be much less favorable. Measures that suggest only small or modest improvements in HRQL might result in incremental cost to incremental QALYs gained ratios above the "adoption" threshold. Policy decisions are subject to the choice of GPB measure.

Reprise: Rationale for Using Two or More GPB Measures to Assess Health Outcomes, Cost-Effectiveness and Cost-Utility

The key argument of this Commentary is a recommendation that investigators in their primary data collection use two or more, not just one, GPB measures per study. Similarly, when investigators systematically extract utility scores from the literature for modelling purposes, they should gather scores based on two or more, not just one, GPB measure. Estimates of QALYs would then be prepared separately for each GPB measure for which there are scores. Such a practice would benefit the CUA studies themselves as well as the fields of HTA and outcomes research.

In choosing among GPB measures investigators should take into account the potential range of effects and side-effects of the intervention as well as the nature of HRQL burden associated with the condition being studied. Does the measure being considered include all of the potentially relevant attributes and levels of severity? As Yang and colleagues note, "an inadequate measure may result in a misallocation of resources" (33, p 42).

In a multi-GPB measure approach, the research protocol could *ex ante* designate one of the GPB measures as the primary measure and classify the other(s) as secondary. If, however, the designated primary GPB outcome measure is not responsive while the secondary one is, the HTA agency may give little weight to the results based on the secondary measure. Alternatively, the protocol could weight each measure equally, with multiple testing adjustment of statistical significance for study-wide comparisons, or use Bayesian analyses of existing evidence for specifying statistical significance. In any case, the investigators will have generated evidence to enhance the design of future studies, including studies for submission to HTA agencies.

An important additional benefit is the increase in scientific evidence on comparisons among measures, or in the language of economics, a positive externality. Information is a public good. Such evidence will better inform the selection and interpretation of measures for future studies. As noted above, studies that provide head-to-head comparisons of the performance of measures are especially valuable (34;35). The fields of outcome measurement, CEA/CUA and HTA would benefit from an increased understanding of the advantages, disadvantages, and limitations of existing GPB measures in a variety of clinical areas. This is especially important in contexts in which there is uncertainty about the performance of GPB measures.

A potential disadvantage of using two measures is that investigators may "game" the system by choosing the GPB measure for which there is evidence on responsiveness in that context or through the selective reporting of results for the GPB that favors the intervention while omitting results for other GPB measures that do not. With respect to the first point, given the lack of a gold standard, selecting a GPB measure with a strong track record with respect to cross-sectional construct validity and responsiveness in previous studies in that context is appropriate and justifiable. Furthermore, it is important to note that the "gaming" issue

https://doi.org/10.1017/S0266462319000412 Published online by Cambridge University Press

also applies to the choice of one from the many GPB measures for use in the study. With modest revision, current standards for submissions to HTA agencies would probably be adequate to handle the selective reporting issue. For instance, requiring the submission to an HTA agency of study protocols before the study is conducted could guard against selective reporting. In addition, some journals publish study protocols.

Another disadvantage of using two measures is the increase in the burden to respondents and to research staff, and the increase in the cost of conducting the study. However, the commonly used GPB measures typically involve minimal respondent burden and low cost, in particular compared with the overall burden and cost of most studies. Note that licensing fees are associated with the use of some GPB measures. If a study uses multiple GPB measures, it may be advisable to randomize the order of their administration. In general, using two measures will represent a very small increase in overall burden and cost.

In conclusion, the routine use of two (or more) GPB measures enhances the rigor and accuracy of CUA studies and HTA. This argument also applies to HRQL studies that include GPB measures even if CUA and submission to an HTA agency is not among the objectives of the study. If both measures indicate that the intervention is effective or ineffective, the analysts can be more confident in that conclusion. If the two measures disagree and there is substantial empirical evidence that one of the measures performs poorly in that context, the analyst can avoid a false negative conclusion. The use of two (or more) GPB measures is feasible, justifiable and would add value.

Conflicts of interest. The authors report competing interests: It should be noted that David Feeny, William Furlong, and George Torrance have a proprietary interest in Health Utilities Incorporated, Dundas, Ontario, Canada. HUInc. distributes copyrighted Health Utilities Index (HUI) materials and provides methodological advice on the use of the HUI.

References

1. **Feeny D, Wu L, Eng K** (2004) Comparing short form 6D, standard gamble, and Health Utilities Index Mark 2 and Mark 3 utility scores: Results from total hip arthroplasty patients. *Qual Life Res* **13**, 1659–1670.
2. **Santana MJ, Feeny D, Johnson JA, *et al.*** (2010) Assessing the use of health-related quality-of-life measures in the routine care of lung-transplant patients. *Qual Life Res* **19**, 371–379.
3. **Fryback D, Dunham NC, Palta M, *et al.*** (2007) U.S. norms for six generic health-related quality of life indexes from the National Health Measurement Study. *Med Care* **45**, 1162–1170.
4. **Guide to the methods of technology appraisal** 2013. Process and methods [PMG9]. National Institute for Health and Care Excellence; April, 2013. https://www.nice.org.uk/process/pmg9/ (accessed January 17, 2019).
5. **Guidelines for the economic evaluation of health technologies: Canada** (4th ed). Canadian Agency for Drugs and Technologies in Health; March, 2017. https://www.cadth.ca/dv/guidelines-economic-evaluation-health-technologies-canada-4th-edition (accessed January 17, 2019).
6. **The Pharmaceutical Benefits Advisory Committee guidelines: version 5.0.** Australian Government Department of Health; September, 2016. https://pbac.pbs.gov.au/.
7. **Henry D** (1992) Economic analysis as an aid to subsidisation decisions: The development of Australian guidelines for pharmaceuticals [review]. *Pharmacoeconomics* **1**, 54–67.
8. **Shiroiwa T, Fukuda T, Ikeda S, Takura T, Moriwaki K** (2017) Development of an official guideline for economic evaluation of drugs/medical devices in Japan. *Value Health* **20**, 372–378.
9. **Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG**, eds (2017) *Cost-effectiveness in health and medicine*. 2nd ed. New York: Oxford University Press.
10. **Feeny D, Krahn M, Prosser LA, Salomon JA** (2017) Valuing health outcomes. In: Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG, eds. *Cost-effectiveness in health and medicine*. 2nd ed. New York: Oxford University Press; p. 167–199.
11. **Rabin R, de Charro F** (2001) EQ-5D: A measure of health status from the EuroQol Group. *Ann Med* **33**, 337–343.
12. **Torrance GW, Feeny DH, Furlong WJ, *et al.*** (1996) Multi-attribute preference functions for a comprehensive health status classification system: Health Utilities Index Mark 2. *Med Care* **34**, 702–722.
13. **Furlong W, Feeny DH, Torrance GW, Barr RD** (2001) The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Ann Med* **33**, 375–384.
14. **Kaplan RM, Anderson JP** (1996) The general health policy model: An integrated approach. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2nd ed. Philadelphia: Lippincott-Raven Press; p. 309–322.
15. **Brazier J, Roberts J, Deverill M** (2002) The estimation of a preference-based measure of health status from the SF-36. *J Health Econ* **21**, 271–292.
16. **Brazier JE, Roberts J** (2004) The estimation of a preference-based measure of health from the SF-12. *Med Care* **42**, 851–859.
17. **Gamst-Klaussen T, Chen G, Lamu AN, Olsen JA** (2016) Health state utility instruments compared: Inquiring into nonlinearity across EQ-5D-5L, SF-6D, HUI-3 and 15D. *Qual Life Res* **25**, 1667–1678.
18. **Chen G, Khan MA, Iezzi A, Ratcliffe J, Richardson J** (2016) Mapping between 6 multiattribute utility instruments. *Med Decis Making* **36**, 160–175.
19. **Fayers PM, Machin D** (2016) *Quality of life: Assessment, analysis and reporting of patient-reported outcomes*. 3rd ed. Chichester: John Wiley & Sons.
20. **Feeny DH, Eckstrom E, Whitlock EP, Perdue LA** (2013) A primer for systematic reviewers on the measurement of functional status and health-related quality of life in older adults. (Prepared by the Kaiser Permanente Research Affiliates Evidence-based Practice Center under Contract No. 290-2007-10057-I.) AHRQ Publication No. 13-EHC128-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2013. https://effectivehealthcare.ahrq6.gov/topics/quality-of-life-functional-status-measurement/white-paper (accessed January 19, 2019).
21. **Guyatt GH, Osoba D, Wu AW, Wyrwich K, Norman GR; Clinical Significance Consensus Meeting Group** (2002) Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* **77**, 371–383.
22. **Schünemann HJ, Guyatt GH** (2005) Commentary - Goodbye M(C)ID! Hello MID, where do you come from? *Health Serv Res* **40**, 593–597.
23. **Marra CA, Woolcott JC, Kopec JA, *et al.*** (2005) A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Soc Sci Med* **60**, 1571–1582.
24. **Marra CA, Rashidi AR, Guy D, *et al.*** (2005) Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? *Qual Life Res* **14**, 1333–1344.
25. **Fryback DG, Palta M, Cherepanov D, Bolt D, Kim J-S** (2010) Comparison of five health-related quality-of-life indexes using item response theory analysis. *Med Decis Making* **30**, 5–15.
26. **Lynch FL, Dickerson JF, Feeny DH, Clarke GN, MacMillan AL** (2016) Measuring health-related quality of life in teens with and without depression." *Med Care* **54**, 1089–1097.
27. **Dickerson JF, Feeny DH, Clarke GN, MacMillan AL, Lynch FL** (2018) Evidence on the longitudinal construct validity of major generic and utility measures of health-related quality of life in teens with depression. *Qual Life Res* **27**, 447–454.
28. **Langfitt JT, Vickery BG, McDermott MP, *et al.*** (2006) Validity and responsiveness of generic preference-based HRQOL instruments in chronic epilepsy. *Qual Life Res* **15**, 899–914.
29. **Pickard AS, Johnson JA, Feeny DH** (2005) Responsiveness of generic health-related quality of life measures in stroke. *Qual Life Res* **14**, 207–219.
30. **Feeny D, Spritzer K, Hays RD, *et al.*** (2012) Agreement about identifying patients who change over time: Cautionary results in cataract and heart failure patients. *Med Decis Making* **32**, 273–286.
31. **Feeny D** (2013) Standardization and regulatory guidelines may inhibit science and reduce the usefulness of analyses based on the application of

https://doi.org/10.1017/S0266462319000412 Published online by Cambridge University Press

preference-based measures for policy decisions. *Med Decis Making* **33**, 316–319.
32. **Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C** (2003) Comparative responsiveness of generic and specific quality-of-life instruments. *J Clin Epidemiol* **56**, 52–60.
33. **Yang Y, Brazier J, Tsuchiya A** (2014) Effect of adding a sleep dimension to the EQ-5D descriptive system: A 'bolt-on' experiment. *Med Decis Making* **34**, 42–53.
34. **Drummond MF, Sculpher MJ, Torrance GW, O'Brien B, Stoddart GL** (2005) *Methods for the economic evaluation of health care programmes*. 3rd ed. Oxford: Oxford University Press.
35. **Gold MR, Patrick DL, Torrance GW, *et al.*** (1996) Identifying and valuing outcomes. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine*. New York: Oxford University Press.
36. **Richardson J, Khan MA, Iezzi A, Maxwell A** (2015) Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI3, 15D, QWB, and AQoL-8D multiattribute utility instruments. *Med Decis Making* **35**, 276–291.
37. **Richardson J, McKie J, Bariola E** (2014) Multiattribute Utility Instruments and their use. In: Culyer AJ, ed., *Encyclopedia of health economics*. vol **2**. San Diego: Elsevier; pp. 341–357.
38. **Brazier J, Ara R, Rowen Donna, Chevrou-Severac H** (2017) A review of generic preference-based measures for use in cost-effectiveness models. *Pharmacoeconomics* **35**, S21-S31.
39. **Devlin N, Shah KS, Feng Y *et al.*** (2018) Valuing health-related quality of life: An EQ-5D-3L value set for England. *Health Econ* **27**, 7–22.
40. **Dewitt B, Feeny D, Cella D, *et al.*** (2018) Estimation of a single preference-based summary score for the patient reported outcomes measurement information system: The PROMIS-Preference (PROPr) Score. *Med Decis Making* **38**, 683–698.
41. **Sintonen H** (2001) The 15D instrument of health-related quality of life: Properties and applications. *Ann Med* **33**, 328–336.

https://doi.org/10.1017/S0266462319000412 Published online by Cambridge University Press