



A Comparison of PROPr and EQ-5D-5L Value Sets

Tianxin Pan^{1,2} · Brendan Mulhern³ · Rosalie Viney³ · Richard Norman² · Janel Hanmer⁴ · Nancy Devlin¹

Accepted: 31 October 2021 / Published online: 17 November 2021
© The Author(s) 2021

Abstract

Objectives The EQ-5D-5L and its value sets are widely used internationally. However, in the US and elsewhere, there is growing use of PROMIS, which has a value set (PROPr) based on the stated preferences of the US population. This paper aims to compare the characteristics of EQ-5D-5L and PROPr value sets and to highlight potential implications for users.

Methods US, Australian and English value sets were used for EQ-5D-5L. PROPr utilities were calculated based on PROMIS-29 + 2. We examined, in each case, (i) the characteristics (e.g. range of values, number of unique values) and distribution of all possible ‘theoretical’ utilities; (ii) dimension/domain importance ranking by the utility of corner states (i.e. health states with the worst level in one domain and the best in all others); (iii) comparisons of utilities for health states hypothesised to be comparable in terms of severity across EQ-5D-5L descriptive systems and PROMIS-29 + 2 domain scores; (iv) the changes in values of adjacent states (i.e. a one-level change in one dimension for EQ-5D-5L and a four-point change in raw scores for PROMIS-29 + 2, with the other dimensions held constant) for dimensions hypothesised to overlap conceptually or be correlated between the two instruments.

Results EQ-5D-5L and PROPr utilities differ systematically. First, the US EQ-5D-5L utilities range from – 0.573 to 1, whereas PROPr values for PROMIS-29 + 2 range from – 0.022 to 0.954. Second, in the US (and English) EQ-5D-5L value sets, pain is the most important dimension whereas in PROPr pain is one of the least important (apart from sleep disturbance). Third, classified based on severity across EQ-5D-5L descriptive systems and PROMIS-29 + 2 domain scores, PROPr has substantially lower values than EQ-5D-5L values for comparable ‘mild’ health states, but higher values for more ‘severe’ health states. Last, when one dimension is considered across its best to worst levels and all other dimensions are held constant at their best or moderate level, in EQ-5D-5L value sets, the greatest changes in utility occur between levels 3 and 4 (moderate and severe) problems; in PROPr that occurred between the most severe states and their descriptively adjacent health states.

Conclusion There are very considerable differences between US EQ-5D-5L and PROPr utilities, despite both in principle representing utility on the same scale anchored at 0 and 1 and both representing the preferences of the US general public. It is important for decision makers and clinical trialists to be aware of these differences. Further work is needed to assess the impact of these differences in value sets using population and patient data, and in longitudinal settings.

✉ Tianxin Pan
tianxin.pan1@unimelb.edu.au

Brendan Mulhern
Brendan.Mulhern@chere.uts.edu.au

Rosalie Viney
Rosalie.Viney@chere.uts.edu.au

Richard Norman
richard.norman@curtin.edu.au

Janel Hanmer
hanmerjz@upmc.edu

Nancy Devlin
nancy.devlin@unimelb.edu.au

¹ Health Economics Unit, Centre for Health Policy, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie Street, Carlton, VIC 3053 Melbourne, Australia

² School of Population Health, Curtin University, Perth, WA, Australia

³ Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

⁴ Department of General Internal Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Digital Features for this article can be found at <https://doi.org/10.6084/m9.figshare.16910563>

A Comparison of PROPr and EQ-5D-5L Value Sets: A Video (MP4 21913 kb)

Key Points for Decision Makers

Both EQ-5D-5L and PROMIS are important preference-accompanied patient-reported outcomes for use in economic evaluation, but they differ in how they describe and value health.

PROPr (the values accompanying PROMIS) has a narrower range of theoretical values compared with EQ-5D-5L. PROPr has substantially lower values than EQ-5D-5L for comparable ‘mild’ health states, and higher values for more ‘severe’ health states. The patterns in utility decrements within the dimensions of each instrument, and the relative importance of dimensions, are very different, despite both representing the preferences of the US general public.

It is important for decision makers and clinical trialists to be aware of these differences and their potential implications for assessing the impact of healthcare interventions.

1 Introduction

A number of generic and condition-specific patient-reported outcome measures (PROMs) are accompanied by value sets (also referred to as utilities or preference weightings) that enable the estimation of quality-adjusted life years (QALYs). These PROMs and their accompanying utilities differ in how they describe and value health, potentially yielding different estimates of QALYs and resulting incremental cost-effectiveness ratios used in decision making [1, 2].

Both EQ-5D and PROMIS are important preference-accompanied PROMs [1]. In addition to its use in monitoring population health, the EQ-5D is widely used in cost-effectiveness analysis for Health Technology Assessment (HTA) around the world [3]. It includes five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression; and each dimension has one item. The initial version (EQ-5D-3L) has three levels of problems for each dimension; the EQ-5D-5L retains the same five dimensions but increases the number of levels of problems on each to five [4]. Utilities for the EQ-5D-5L have been produced in

many countries and are typically obtained using a standard protocol (the EuroQol Valuation Technology, EQ-VT) that includes both discrete choice experiments (DCE) and the composite time trade-off (cTTO), with resulting value sets based either on cTTO or a hybrid of cTTO and DCE [5].

PROMIS was developed in the US using item response theory (IRT) to calibrate a large number of questions for each health domain to create an item bank [6]. PROMIS has over 100 health domains with a calibrated ‘item bank’ of questions that can be administered via standard short forms, custom short forms, or computer adaptive testing [7]. There are three PROMIS profile measures comprising fixed sets of short forms: PROMIS-29, PROMIS-43 and PROMIS-57. Profile measures comprise 4-, 6-, or 8-item short forms for each of seven domains: Anxiety, Depression, Fatigue, Pain Interference, Physical Function, Sleep Disturbance, Ability to Participate in Social Roles and activities (from here, known as social roles), plus a 0–10 Pain Intensity numeric rating item [8]. Among the three profile measures, PROMIS-29 is the most widely used. A generic societal preference-based scoring system for PROMIS (PROPr) has been produced based on multi-attribute utility theory and eliciting stated preferences using standard gamble (SG) [9]. Seven PROMIS health domains are needed to obtain PROPr utilities: Cognitive Function—Ability (from here, described as Cognitive Function), Depression, Fatigue, Pain Interference, Physical Function, Sleep Disturbance, Social Roles. While computerised adaptive testing (CAT) administration is the recommended method to obtain the domain scores necessary to calculate PROPr, the use of PROMIS-29 and two items from Cognitive Function v2.0 (PROMIS-29 + 2 Profile v2.1, from here, described as PROMIS-29 + 2) is the recommended measure to obtain PROPr if a fixed form is required for the project [10].

Both PROPr and EQ-5D-5L are potential candidates for use in economic evaluation. However, a recent systematic review of studies comparing EQ-5D and PROMIS-29 instruments noted the lack of research comparing the measurement properties and utilities of the two, and highlighted the importance of comparing the characteristics of their respective value sets [11]. It is important for decision makers and those using the measures in clinical trials to understand the differences between instruments and their implications for assessing the impact of healthcare interventions. A full account of the differences requires an understanding of both the characteristics and measurement properties of the items and their preference weights.

The utilities that accompany each are, in principle, measured on the same scale (i.e. anchored at full health and dead). This is the assumption underpinning the use of these utilities in QALY estimation, that is, that utilities ‘mean’ the same thing in terms of representing, on average, a society’s preferences about health-related quality of life (HRQoL),

Table 1 Comparing the measurement of health in EQ-5D-5L and PROPr

	EQ-5D-5L	PROPr (PROMIS-29 + 2)
Dimension/health domains	Mobility Self-care Usual activity Pain/discomfort Anxiety/depression	Cognitive Function Depression Fatigue Pain Interference Physical Function Sleep Disturbance Ability to Participate in Social Roles
Number of items in each dimension/domain	One question per dimension	Four questions per domain (except for Cognitive Function which has two questions)
Item response descriptions	5 levels: no problem; slight; moderate; severe; unable to/extreme	5 levels (different wording for different items); values ranging from 1 to 5
Dimension/domain level descriptions	The same as the item response level	Domain sum raw score Standardised T-score
Descriptions of health state	A five-number string, representing the level of each dimension in the order in which they appear in the questionnaire	A vector with 7 elements, each representing a level of T-score on one domain in the order mentioned above
Display of health state in the valuation study	Describe health states as five sentences (each sentence representing one level in each dimension)	Display health state by domain, with two representative items from an item bank in a domain presented together

Considering the systematic difference between the two instruments, we summarised the terminology and descriptions used in each instrument. For detailed terminology/glossary, please see:

EQ-5D Terminology (<https://euroqol.org/support/terminology/>)

PROMIS Glossary (<http://www.healthmeasures.net/resource-center/measurement-science/glossary>)

regardless of which descriptive system they accompany. This is important in, for example, the use of these utilities to estimate QALYs and incremental QALY gains in cost effectiveness analysis. When decision makers use such evidence, they rely on utilities having broad comparability. For example, various PROMs (e.g. PROMIS-29, EQ-5D-5L, HUI3, SF-6Dv2) have values < 0, and this should be capable of being interpreted as meaning that these states, in whatever PROMs they relate to, are worse than being dead. The characteristics of these utilities are important because they introduce an exogenous source of variance into the analysis of the data generated by the respective descriptive systems [12]. Examining the characteristics of the utilities that accompany different PROMs is therefore an essential precursor to a full understanding of the way that measurement and valuation properties interact when analysing patient data.

This paper compares EQ-5D-5L value sets and PROPr to better understand the differences between the characteristics of the utilities generated by each. The primary focus of comparison is between the EQ-5D-5L US value set and PROPr value set, both of which were generated from the stated preferences of the US general public. Differences between any two value sets may result from differences in sample population, differences in the classification system of the instruments and differences in various aspects of the valuation method used [1]. Therefore, in order to test whether results regarding the nature of PROPr and EQ-5D-5L utilities are specific to the characteristics of the US EQ-5D-5L value set,

we additionally included two further EQ-5D-5L value sets from English-speaking countries: England and Australia. The English value set was selected as it was developed using the same valuation protocol (i.e. EQ-VT) as the US value set, and the Australian value set because it was developed using a different valuation method (DCE with duration).

2 Method

For each instrument, we provide an overview of (1) the descriptive system and (2) the valuation process. We then describe the approach we used to compare the utilities for all possible health states described by each instrument.

2.1 The Two Instruments

2.1.1 EQ-5D-5L Questionnaire and Value Sets

The development of the EQ-5D was motivated in part by the aim to facilitate QALY estimates for use in economic evaluation [13]. In the EQ-5D-5L, there are five response levels for each of the five dimensions (see Table 1): no, slight, moderate, severe and unable to/extreme problems. EQ-5D-5L health states can be described as a five-number string, representing the level of each dimension in the order in which they appear in the questionnaire (as described above) [14]. The best health state is described as 11111 and the worst as 55555.

The EuroQol Group has developed a protocol (EQ-VT) to create value sets for the EQ-5D-5L, which has been used in more than 25 countries [5, 15], including in the US [16] and England [17]. The Australian value set was developed using DCE with duration [18]. It includes an interaction (described as the N5 term), capturing an additional reduction in value for any health state with a level 5 problem on at least one of the dimensions.

2.1.2 PROMIS Measures and PROPr Scores

The PROPr builds on PROMIS domains and is based on multi-attribute utility theory. A PROMIS domain's item bank is an IRT-calibrated set of items for that domain. Domains are calibrated to the US general population and presented as T-scores with a mean of 50 and standard deviation (SD) of 10. A higher T-score represents more of the concept being measured [19].¹ As shown in Table 1, the items used in PROPr each have five response levels. For PROMIS-29 + 2, a fixed four-item short form, the sum raw score for each domain except for Cognitive Function (which has 2 items) ranges from 4 to 20; the Cognitive Function domain ranges from 2 to 10.

The PROPr algorithm allows preference-based scores to be estimated from health states described by the seven PROMIS domains mentioned in the Introduction. Detailed information on domain selection and methods used to create PROPr have been described elsewhere [9, 20, 21]. There are different options to get the domain scores to calculate the PROPr scores, as described in Appendix A in electronic supplementary material (ESM) 1.

2.2 Analysis of Theoretical Values

As mentioned, the primary focus of analysis is comparing the EQ-5D-5L US value set and PROPr value set, because both were generated from the stated preferences of the US general public. We additionally included the EQ-5D-5L value sets for England and Australia.²

We chose to calculate PROPr utilities using PROMIS-29 + 2 given its wide use [10]. All the possible T-scores were obtained from the scoring manual³ for PROMIS-29 + 2 [19]. In this section and the results section, PROPr utilities refer to the utility scores generated based

¹ For function domains like Physical Function, Social Role and Cognitive Function, a T-score of 60 is one SD *better* than average. However, for symptom domains like Depression, Fatigue, Pain and Sleep Disturbance, a T-score of 60 is one SD *worse* than average.

² See Appendix B in ESM 1 for more information on the three EQ-5D-5L value sets and their model characteristics.

³ The scoring manual is a look-up table to convert domain raw scores to expected T-scores, instead of being IRT scored.

on PROMIS-29 + 2 (rather than the PROPr scoring system). We generated all theoretically possible health states for both instruments using Python v3.8. We obtained the PROPr utilities based on code written by Janel Hanmer in September 2017 using SAS 9.4 and transformed this to Stata code (see ESM 2). We used Stata 14.2 for the analyses. In order to check whether our results are specific to PROMIS-29 + 2, we also repeated analyses on utilities calculated for PROMIS-29 (results provided in Appendix C in ESM 1).

2.2.1 Value Set Characteristics

We produced descriptive statistics for all 'theoretically possible' values for EQ-5D-5L using the three value sets and for PROPr. In each case this included the range of values, the percentage of health states having values < 0 and the number of unique values. We plotted the overall distribution of the 'theoretical' values using a kernel density histogram.

We assessed dimension/domain importance by their ranking produced by the utility of corner states, that is, the utility of health states with the worst level in one domain and the best in all others.

2.2.2 Value Set Characteristics for Matched States

We carried out a comparative analysis on the health states hypothesised to be comparable in terms of severity across EQ-5D-5L and PROMIS-29 + 2. We considered 'full health', 'mild', 'moderate', 'severe' and 'worst' health states. We considered comparable states based on item response descriptions given that both instruments use five-level scales.⁴ In EQ-5D-5L, that corresponds to 11111, 22222, 33333, 44444 and 55555. In PROMIS-29 + 2, we used

⁴ It is noteworthy that EQ-5D-5L and PROMIS use different approaches to define the severity of health states for their respective classification systems. In EQ-5D-5L it is defined as the five levels of problem. PROMIS scientists provided guidelines for interpreting PROMIS scores based on T-scores; with 0.5, 1.0 and 2.0 standard deviations used as thresholds to categorise as 'within normal limit', 'mild', 'moderate' and 'severe' states [22]. However, it is unknown where the EQ-5D-5L category fits on the same construct. Therefore, we considered the 'comparable' severity of states in a more straightforward way, on item level descriptions and domain raw scores. In sensitivity analysis, we calculated corresponding PROPr utilities for the five 'comparable' health states defined based on domain T-scores, with 'full health' equaling best possible T-score on PROMIS-29+2, 'mild' = T-score of 57.5 or 42.5 (the former for symptom concept domains and latter for function concept domains), 'moderate' = T-score of 65 or 35, 'severe' = T-score of 75 or 25, 'worst' = worst possible T-score. Because the cut points are general guidelines for PROMIS domains, the thresholds for each PROMIS domain may differ slightly. More detail and the results are reported in Appendix D in ESM 1.

five cut points on sum raw scores. The corresponding sum raw scores for full health, mild, moderate, severe and worst health states for symptom domains (e.g. Depression) are 4, 8, 12, 16 and 20, and for function domains (e.g. Physical Function) are 20, 16, 12, 8 and 4. The matched health states are presented in Sect. 3.2.

2.2.3 Differences in Utility Between Adjacent States

We carried out an analysis to understand how changes in utility reflect changes in severity on the descriptive system and domain scores in each instrument. We followed the method described in Mulhern et al. [23] to assess the values of adjacent states within the descriptive system, and comparing the differences across the value sets.⁵ An adjacent state pair was defined as having one dimension with a one-level difference for EQ-5D-5L or with a four-point change in sum raw score for PROMIS-29 + 2, with the other dimensions held at the best health level, moderate health level and worst health level, respectively. We compared dimension pairs that we hypothesised to overlap conceptually or those found to be highly correlated using data from the general population and patients [26, 27]. For example, for mobility in EQ-5D-5L, we report the increase in utility between 51111, 41111, 31111, 21111 and 11111; between 53333, 43333, 33333, 23333 and 13333; and between 55555, 45555, 35555, 25555 and 15555. For physical function in PROPr, we compare this with the increase in utility using domain raw score at 4, 8, 12, 16 and 20 for physical function with best, moderate and worst health domain scores for other domains, respectively.⁶

3 Results

3.1 Comparing Value Set Characteristics

Table 2 compares the value set characteristics. The theoretical value of the US EQ-5D-5L value set ranges from -0.573 to 1 , whereas PROPr values range from -0.022 to 0.954 . Compared with all three EQ-5D-5L value sets, PROPr has a higher minimum value (close to 0) for the worst state. The decrement from the best to next-best health state (i.e. from full health to the mildest state) is also larger for PROPr (a change of 0.065) compared with that of US EQ-5D-5L (0.057). In the EQ-5D-5L value sets, pain/discomfort

is either the most or the second most important dimension (proxied by the overall size of the decrement) while usual activities is the least important. However, in PROPr (based on PROMIS-29 + 2), pain is the least important (apart from sleep disturbance). Appendix E (in ESM 1) reports the utilities of each disutility corner state for each instrument.

Figure 1 demonstrates that the distribution of theoretical values for PROPr differ substantially from the distributions apparent for the three value sets for EQ-5D-5L. EQ-5D-5L US values are symmetric, with a wider range of values, whereas PROPr values are skewed, with a narrower and lower range of values. Given the narrower range of values, the number of unique values that exist within PROPr as explored by this exercise are fewer than that for the EQ-5D-5L value sets (as shown in Table 2).

3.2 Value Set Characteristics for Matched States

Table 3 shows that PROPr utilities are consistently lower across the less severe health states compared with EQ-5D-5L value sets but PROPr values are consistently higher for more severe health states. The pattern remained the same whether we classified the health states in PROMIS-29 + 2 based on domain raw scores or domain T-scores (as shown in Appendix D in ESM 1).

3.3 Differences in Utility Between Adjacent States

Figure 2 presents the changes in utilities between adjacent states in the EQ-5D-5L US value set and PROPr in comparable dimensions. When holding all other dimensions at the best health level, the decrement in utility associated with any move away from full health to a mild health state was larger in all comparable dimensions in PROPr than in EQ-5D-5L value sets. Consequently, EQ-5D-5L utilities are higher for such mild states. Notably, the mildest problem in mobility on EQ-5D-5L, with a utility of 0.904 in the US value set, corresponds with a PROPr utility of 0.729 . In EQ-5D-5L, the largest change in utility occurs in the move between severe (level 4) and moderate (level 3) problems, whether holding other dimensions at their best (level 1), moderate (level 3) or worst level (level 5). Utility associated with any given dimension, considered across its best to worst levels, and holding all other dimensions constant at either best or moderate levels, is always *lower* in PROPr than EQ-5D-5L. However, this pattern ‘flips’ when any one dimension is considered across its best to worst levels and all other dimensions are held constant at their most severe level; in these cases, there is *no* (or close to no) change in utility on PROPr, and PROPr utilities are *higher* than EQ-5D-5L for such states. The changes in utilities between adjacent states in all EQ-5D-5L dimension and PROPr domains are reported in Appendix F in ESM 1.

⁵ We also considered comparing the mean transition values between EQ-5D-5L and PROPr [24, 25]. However, given that there will be 25,000 possible transitions for each of the 3125 EQ-5D-5L health states and many more transitions for the 217 million possible PROMIS-29+2 health states, we did not pursue this analysis.

⁶ See Appendix D in ESM 1 for sensitivity analysis by categorising adjacent health states based on domain T-scores.

Table 2 Overall descriptive characteristics of the four value sets (modelled values)

	EQ-5D-5L			PROPr
	US	England	AU	PROMIS-29 + 2
Total number of ‘theoretical’ possible health states ^a	3125	3125	3125	217,238,121
Range	− 0.573 to 1	− 0.285 to 1	− 0.676 to 1	− 0.022 to 0.954
Median (interquartile range)	0.242 (0.044, 0.429)	0.388 (0.255, 0.546)	0.123 (− 0.076, 0.331)	0.130 (0.072, 0.207)
Number of health states worse than dead	626 ^d	159	1043	1,683,741
Percentage of health states worse than dead	20.03%	5.09%	33.38%	0.78%
Number of unique values	1504	1319	1443	909
Dimension importance order ^b	PD	PD	AD	PF
	MO	AD	PD	D
	AD	MO	MO	Fatigue
	SC	SC	SC	CF
	UA	UA	UA	SR
				Pain
				SD
‘Mildest’ state	0.943 (11112)	0.950 (12111, 11211)	0.928 (21111, 12111)	0.935 (SD)

^aThe number of theoretical possible health states in PROPr (based on PROMIS-29 + 2) is calculated based on the possible combination of domain raw scores in each domain. For example, in each of the seven domains except for Cognitive Function, there are 17 possible domain raw scores, and for Cognitive Function, there are 9 raw scores. The number of possible health states described in raw score in PROMIS-29 + 2 is 217,238,121 ($=17^6 \times 9$)

^bMO mobility, SC self-care, UA usual activity, PD pain/discomfort, AD anxiety/depression, CF Cognitive Function, D Depression, Pain Pain Interference, PF Physical Function, SD Sleep Disturbance, SR Social Roles

^cImportance is judged by the size of the coefficient for disutility corner state in each dimension (e.g. 11115 for anxiety/depression dimension for EQ-5D-5L, and [10; 20; 4; 4; 20; 4; 20] for depression domain for PROPr—order of the domain: CF, D, Fatigue, Pain, PF, SD, SR). The disutility corner state for a domain corresponds to the state described by the unhealthiest level on that domain, and the healthiest on all others

^dThis contrasts with Pickard et al., who stated 624 health states were worse than dead. The distinction is that Pickard et al. rounded two health states valued at $-7.45e-09$ to zero. [16]

We adapted the table from Mulhern et al. [23]

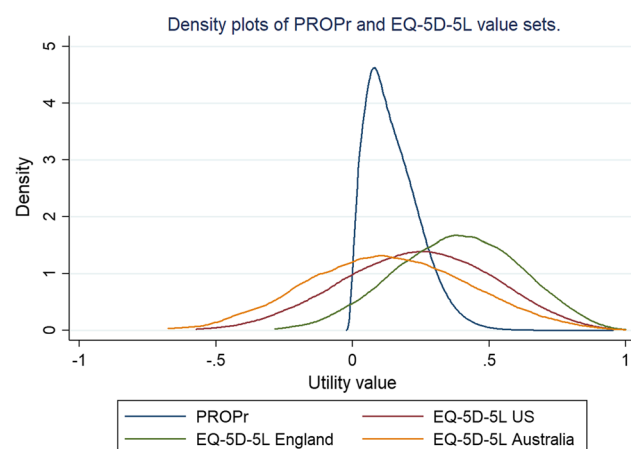


Fig. 1 Density plot of PROPr and EQ-5D-5L value sets. This figure shows the density plots of theoretical values for PROPr and EQ-5D-5L (utility value as X-axis and density as Y-axis)

4 Discussion

We have compared ‘theoretical’ values for three EQ-5D-5L value sets (US, England and Australia) and the US PROPr utilities based on PROMIS-29 + 2. The results show that EQ-5D-5L and PROPr utilities differ systematically, in important ways. We begin by summarising the key differences and then discuss each in more detail. First, PROPr has a narrower range of values compared with EQ-5D-5L. Second, the relative importance of dimensions and the utility decrements across dimensions for each instrument are very different. The relative distance between full health and the disutility corner states in PROPr differentiates less between domains: the utility decrements across PROPr are similar between domains in contrast to EQ-5D-5L values. Third, PROPr has substantially lower values for comparable relatively mild health states while having higher values for descriptively more severe health states, compared with the

Table 3 Value set characteristics for matched states

	EQ-5D-5L health states	PROMIS-29 + 2 raw score ^a	EQ-5D-5L			PROPr (PROMIS-29 + 2)
			US	England	AU	
Full health	11111	10; 4; 4; 4; 20; 4; 20	1	1	1	0.954
Mild	22222	8; 7; 7; 7; 17; 7; 17	0.630	0.701	0.521	0.437
		8; 8; 8; 8; 16; 8; 16				0.391
		8; 9; 9; 9; 15; 9; 15				0.358
Moderate	33333	6; 11; 11; 11; 13; 11; 13	0.449	0.593	0.375	0.244
		6; 12; 12; 12; 12; 12; 12				0.189
		6; 13; 13; 13; 11; 13; 11				0.149
Severe	44444	4; 15; 15; 15; 9; 15; 9	- 0.329	- 0.094	- 0.434	0.082
		4; 16; 16; 16; 8; 16; 8				0.067
		4; 17; 17; 17; 7; 17; 7				0.054
Worst health	55555	2; 20; 20; 20; 4; 20; 4	- 0.573	- 0.285	- 0.676	- 0.022

^aDomain order: Cognitive Function, Depression, Fatigue, Pain, Physical Function, Sleep Disturbance, Social Roles

For symptom domains (with a higher score indicating better health) like Depression, Fatigue, Pain, Sleep Disturbance, the corresponding sum raw scores for full health, mild, moderate, severe and worst health states are 4, 8, 12, 16 and 20. For function domains like Physical Function and Social Role, the corresponding sum raw scores are 20, 16, 12, 8 and 4. For the Cognitive Function domain, these are 10, 8, 6, 4 and 2. In addition, we conducted a sensitivity analysis for the comparisons at the mild, moderate and severe health states by examining three domain sum raw scores (i.e. one point lower than the intermediate sum raw score, the intermediate score and one point higher). For example, for the intermediate score at 8, we also examined sum raw scores at 7 and 9

EQ-5D-5L values. Taken together, this means PROPr has a narrower range of values than the EQ-5D-5L values. Further, whereas the most commonly arising values in the EQ-5D-5L value sets lie in the mid-range of values, in PROPr the most frequently arising values are much lower. Lastly, the patterns in utility decrements within the dimensions of each instrument are different. When one dimension is considered across its best to worst levels and all other dimensions are held constant at their best or moderate level, in EQ-5D-5L value sets the greatest changes in utility occurred between levels 3 and 4 (i.e. the change from moderate to severe problems); in PROPr that occurred between the most severe ‘corner’ states and their adjacent states. However, when all other dimensions are held constant at their most severe level, there is no (or close to no) change in utility on PROPr, and PROPr utilities are higher than EQ-5D-5L for such states.

The narrower range of values in PROPr is a product both of its higher minimum value and the fact that the ‘best’ state (based on PROMIS-29 + 2) has a utility < 1. It is worth noting that PROPr utility can only reach 1 when using some PROMIS measures. PROPr was designed for the wider PROMIS system, and not only for PROMIS-29 + 2. PROPr focuses on a subset of the seven domains, and the items displayed in the valuation study for each domain were chosen from the domain item banks, in order to span the health state space. As a consequence, the chosen items are not necessarily included in every PROMIS measure (as shown in Appendix G in ESM 1), so

the domain score range can be reduced.⁷ This leads to the truncated PROPr utilities.

The relative importance of dimensions is also very different between the EQ-5D-5L US value set and PROPr—despite both being based on the stated preferences of the US general public. For example, pain is the most important dimension in the EQ-5D-5L US value set, whereas in PROPr it is the *least* important, apart from sleep disturbance. It is worth noting that pain ranks third in the whole PROPr scoring system, which is the main difference in the relative importance of domains using the PROMIS-29 + 2 instrument and PROPr utilities. This is because pain is the only domain for which the worst possible T-score on PROMIS-29 + 2 cannot be achieved (as shown in Appendix G in ESM 1). More generally, it is important to note that the importance of domains in PROPr is intertwined with the domain T-score ranges determined by the specific item chosen for that domain in the valuation study.

The utility decrements across domains in PROPr differentiate less between domains than is the case for EQ-5D-5L. PROPr utilities are strongly affected by the level of problem,

⁷ Similarly, when we calculated PROPr using PROMIS-29, the best health state described by PROMIS-29 has a utility of 0.905. This is largely because the Cognitive Function domain (imputed by other PROMIS-29 domains) cannot reach the best possible theta score [28].

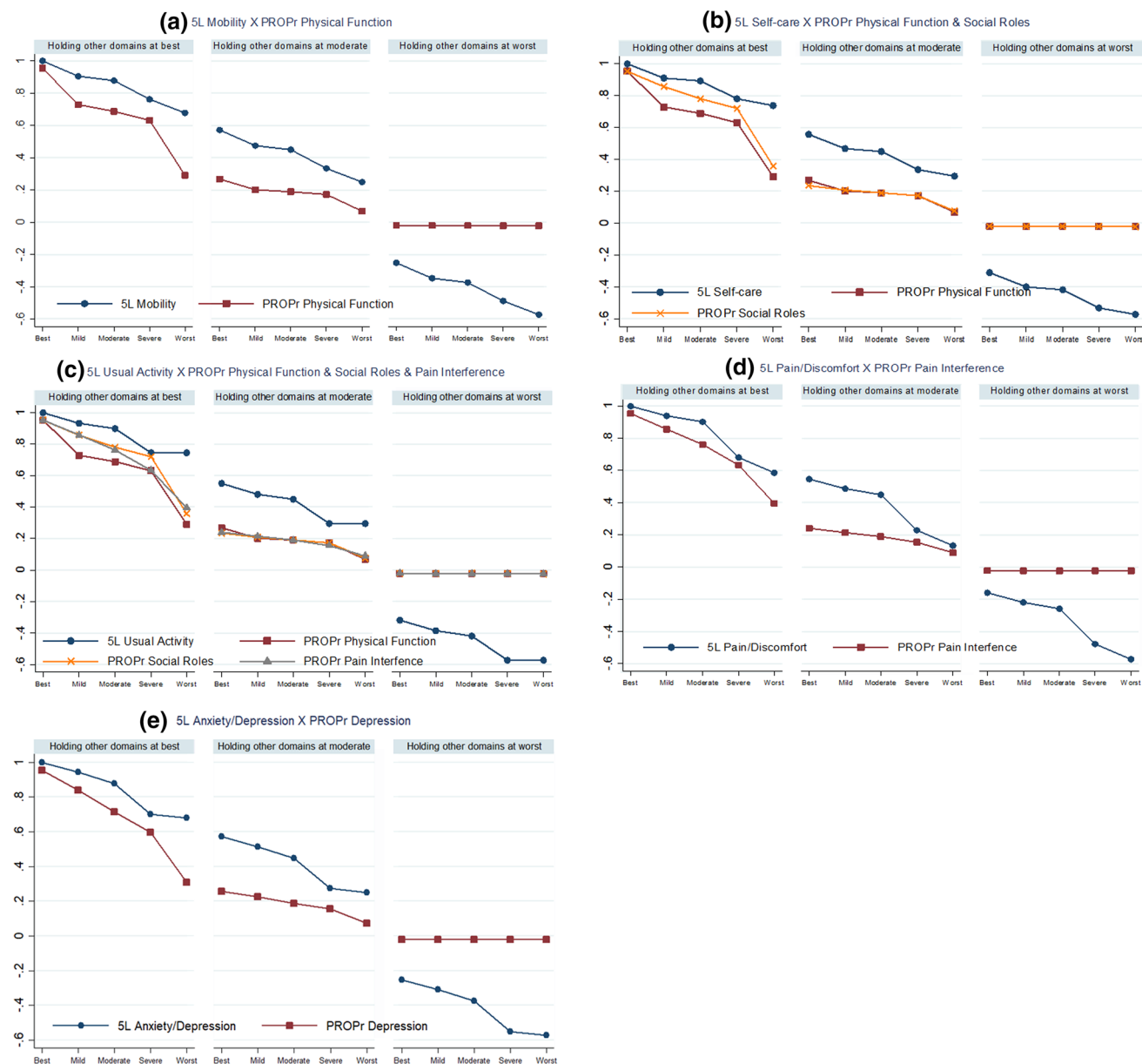


Fig. 2 Changes in utilities between adjacent states between ‘comparable’ dimensions in EQ-5D-5L US value set and PROPr, when holding other domains at best, moderate and worst level respectively

and much less so by what aspect of health it is experienced in. One possible explanation for this is that the participants may have had difficulty using the SG valuation process to distinguish between these states.

We have found that PROPr has lower values than EQ-5D-5L for ‘comparable’ mild states. In interpreting this finding, it is important to note that PROPr inherits the characteristics of PROMIS measures, which do not exhibit the ceiling effects observed in EQ-5D instruments. Thus, PROPr makes it less likely for a respondent to self-report

themselves as being at the best health state [9].⁸ This means that mild states in PROPr are *very* mild compared with EQ-5D-5L mild states, so our finding that they have lower utilities is notable.

At the other end of the scale, the worst health problems described by PROMIS-29 + 2 have values close to 0 (− 0.022), whereas those described by EQ-5D-5L

⁸ For example, the best physical functioning in PROPr is “able to dress yourself, including tying shoelaces and buttoning up your clothes without any difficulty and able to run 100 yards (100 m) without any difficulty.” In contrast, the best level of physical functioning in EQ-5D-5L is “I have no problems walking”.

have larger negative values (-0.573 in the US value set, -0.285 and -0.676 for English and Australian value sets, respectively). In general, PROPr utilities are also higher than EQ-5D-5L values for more severe health states. However, there is an exception to that; the utilities of corner states in PROPr are substantially lower than those in EQ-5D-5L among corresponding domains. This pattern is also reflected in our analysis of differences in utility between adjacent states (Fig. 2)—EQ-5D-5L utilities are *higher* than PROPr utilities when holding other domains at best or moderate level, but become lower than PROPr when holding other domains at worst level.

The characteristics of PROPr and EQ-5D-5L utilities observed in this paper are likely to be a product both of what is being measured, and fundamental differences in the stated preference methods used and how utility algorithms are estimated from them. This includes mode of administration, elicitation technique, selection of the sub-set of health states to value, display of health states, number of health states valued by each respondent, the approach to anchoring at 0 and to handling values worse than dead, quality control and data exclusion criteria, modelling approach and choice of final model. For example, for EQ-5D-5L, the TTO task bounds individual respondents' utilities between 1 and -1 by design [25]. In contrast, PROPr asks respondents whether they prefer the all-worst state or dead. Instead of $\text{dead} = 0$ being established for each respondent (as is the case in EQ-5D-5L value sets), the PROPr approach entails calculating the mean utility of *dead* for all respondents who considered dead the worst state; the mean utility of *dead* for all those who considered the all-worst state to be worst, and rescaling and combining the two, weighted by group size [29]. There are a number of unresolved methodological issues regarding the stated preference methods used to value HRQoL. Recent papers raise fundamental questions about the way such utilities are anchored at 0 [30, 31]. A different approach to anchoring at 0 would have an important effect on both PROPr and EQ-5D utilities [32], and indeed the values for any PROMs.

It is important to note that both PROPr utilities and EQ-5D-5L utilities are based on stated preference data—there is no 'gold standard' against which to judge the veracity of value sets and revealed preferences data cannot be used to validate them. Nevertheless, the extent of the differences between utilities for PROPr and the EQ-5D-5L is striking, given both are based on the preferences of the US general public.

The contrasting characteristics of the utilities we have reported in this paper will combine with differences in the descriptive systems to affect data and results arising from their use and application. For example, cross-sectional validation studies of PROPr based on the general population

have shown that the absolute differences in utilities between different socio-demographic groups were smaller for PROPr compared with EQ-5D-5L [33]. The smaller range of PROPr utilities means the change in QALYs from HRQoL-increasing interventions might be reduced. However, the implications of these differences in utility for QALY estimates are quite complex. For example, interventions that result in improvements from the mildest states may yield larger utility gains when estimated by PROPr rather than EQ-5D-5L. This is because PROPr utilities for mild states are *much* lower than that of EQ-5D-5L, so any improvement from a mild state to full health will yield a relatively large incremental gain in PROPr utility. PROPr utilities for disutility corner states (i.e. health states with the worst level in one domain and the best on all others) are substantially lower than the utilities for comparable EQ-5D-5L states, so the changes in utilities between these corner states and their 'next best' adjacent states are larger in PROPr. However, PROPr utilities for the all-worst health states are much higher than those for EQ-5D-5L. For the all-worst state in PROPr, an intervention that leads to an improvement in just one domain (e.g. a reduction in pain) results in no change in utility (e.g. even if pain is eliminated altogether), whereas in EQ-5D-5L such improvements generate large gains in utility. It is worth noting that, as utilities will be applied to both treatment and control groups, some of these utility differences between instruments may cancel out. Furthermore, the impact of differences between utilities on QALY gains also depends on whether the change is from improving quality of life or extending length of life, or both. This requires further investigation using clinical cohort data.

There are a number of limitations of this study. First, we have not assessed in this study whether our matched states are equivalent. In our main analysis, we considered comparable PROMIS-29 + 2 states to EQ-5D-5L health states based on domain raw scores, but this 'direct' match did not take advantage of the IRT-based descriptive system for PROMIS-29 + 2. In the sensitivity analysis (see Appendix D in ESM 1), we classified PROMIS-29 + 2 states based on T-scores, although it is unknown where EQ-5D-5L fits on that construct. Notwithstanding these limitations, the results of both sets of analyses were qualitatively similar in the differences in utility between adjacent health states and of matched states. Second, there are different options for estimating PROPr scores. Our main analysis focused on the PROPr scores calculated by PROMIS-29 + 2, given it is the most widely used PROMIS profile measure. It is possible that some characteristics of PROPr may differ depending on the PROMIS instrument chosen (in addition to the way T-scores are obtained and the set of intervals chosen, as mentioned previously), limiting the generalisability of the results to other PROMIS instruments. To check for this, we also calculated PROPr using PROMIS-29 (results presented

in Appendix C in ESM 1). Although PROPr scores differ somewhat in the range of values (– 0.021 to 0.905 when calculated by PROMIS-29) and there are minor differences in PROPr scores of matched states, the results of both sets of analyses were qualitatively similar in terms of the differences in utility between adjacent health states and of matched states when compared with EQ-5D-5L value sets and the overall key conclusions drawn about the nature of PROPr and EQ-5D-5L values. Third, we have not tested the impact of EQ-5D-5L value sets and PROPr using patient-reported data, which would enable us to directly compare QALY estimates produced from each. More research is needed using both instruments in a variety of clinical contexts and longitudinally [11, 34, 35]. Our ongoing research aims to address these gaps in evidence using patient data [26].

It is important for decision makers and clinical trialists to be aware of the substantial differences between the characteristics of utilities for these two widely used PROMs, and their potential implications in QALYs estimation used in decision making. The different dimensions included in the classification systems, combined with the non-trivial differences in the characteristics of the utilities in PROPr and EQ-5D-5L value sets reported here, means that which instrument is chosen will potentially yield different results and evidence for decision making. To date, PROPr is available only for the US. Given the growing importance and use of PROMIS measures, we suggest that international PROPr utilities are required, in different country settings, and further comparisons of these utilities with the EQ-5D-5L and other measures are needed to understand the impact of different measures and methods used on the values elicited.

5 Conclusions

There are important differences between the characteristics of EQ-5D-5L and PROPr utilities. It is important for decision makers and clinical trialists to understand these differences, as they may exert an effect on utility-weighted patient data produced by each instrument. The smaller range of PROPr means the change in QALYs from HRQoL-increasing interventions might be reduced. However, interventions that result in improvements from disutility corner states and from the mildest states may yield larger QALY gains when estimated by PROPr. Of course, these differences in utilities will apply to both treatment and control groups and the effects on estimates of QALY gains also depend on whether these come from improvements in quality of life or by extending length of life. The characteristics of PROPr may be attributable to the specific methods used to obtain stated preferences, suggesting further research is required to understand that, and potentially to investigate the merits of

alternative methods. The methods we develop in this paper potentially provide a means of comparing the properties of utilities between any two HRQoL instruments.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40273-021-01109-3>.

Acknowledgements This work was funded by the EuroQol Research Foundation. Opinions expressed in the paper are those of the authors and are not necessarily those of the EuroQol Group. The authors would like to thank Qilin Yu, PhD candidate at The University of Melbourne, for his help and advice on generating all the ‘theoretical’ possible health states in Python v3.8. We would like to thank Ernest H. Law, PhD, Pfizer Inc., for his discussion and comments on this paper and participants at the 37th EuroQol Plenary Meeting, 16–17 September 2020, for their questions and comments.

Declarations

Funding This study was funded by the EuroQol Research Foundation. The funding agreement ensured the authors’ independence in designing the study, interpreting the data, writing and publishing the report. The open access fee for publishing this article is funded by the EuroQol Research Foundation.

Conflict of interest ND, RV, RN and BM are members of the EuroQol Group. JH contributed to development of PROPr.

Availability of data and material We generated all theoretically possible health states for EQ-5D-5L and PROMIS-29 using Python v3.8. We calculated the health utility score for each theoretically possible health state using published value sets for each instrument. The data used in this study is available upon request.

Code availability We generated all theoretically possible health states for EQ-5D-5L and PROMIS-29 (PROPr) using Python v3.8. We obtained the PROPr utilities based on code written by Janel Hanmer in September 2017 using SAS 9.4 and transformed this to Stata code. We shared the Stata code as electronic supplementary material (ESM) 2.

Author contributions ND, BM and TP conceived and designed the study. TP led the data analysis and interpretation. ND, BM, RV, RN and JH supported the data analysis and interpretation. TP and ND wrote the first draft; BM, RV, RN and JH commented on and edited drafts of the manuscript.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission

directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

References

- Brazier J, Ara R, Rowen D, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. *Pharmacoeconomics*. 2017;35(1):21–31.
- Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. *Measuring and valuing health benefits for economic evaluation*, 2nd editions. Oxford: Oxford University Press; 2017.
- Kennedy-Martin M, Slaap B, Herdman M, van Reenen M, Kennedy-Martin T, Greiner W, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *Eur J Health Econ*. 2020;21(8):1245–57.
- Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–36.
- Oppe M, Devlin NJ, Hout B, Krabbe PFM, Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014;17(4):445–53.
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol*. 2010;63(11):1179–94.
- Health Measures. List of adult measures. 2020. Cited 27 Apr 2020. <http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/list-of-adult-measures>
- Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, et al. PROMIS((R)) adult health profiles: efficient short-form measures of seven health domains. *Value Health*. 2019;22(5):537–44.
- Dewitt B, Feeny D, Fischhoff B, Cella D, Hays RD, Hess R, et al. Estimation of a preference-based summary score for the patient-reported outcomes measurement information system: the PROMIS((R))-preference (PROPr) scoring system. *Med Decis Mak*. 2018;38(6):683–98.
- Health Measures. Preference-based Scores. 2019. Cited 10 Mar 2020. www.healthmeasures.net/score-and-interpret/calculate-scores/preference-based-scores
- Pan T, Mulhern B, Viney R, Duy Tran A, Norman R, Devlin N. Evidence on the relationship between PROMIS-29 and EQ-5D: a literature review. *Qual Life Res*. 2021. <https://doi.org/10.1007/s11136-021-02911-4>.
- Parkin D, Rice N, Devlin N. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? *Med Decis Mak*. 2010;30(5):556–65.
- Devlin NJ, Brooks R. EQ-5D and the EuroQol group: past, present and future. *Appl Health Econ Health Policy*. 2017;15(2):127–37.
- Devlin N, Parkin D, Janssen B. *Methods for analysing and reporting EQ-5D data*. Berlin: Springer; 2020.
- Devlin N, Roudijk B, Ludwig K. *Value sets for EQ-5D-5L: a compendium, comparative review and user guide*. Berlin: Springer; 2022.
- Pickard AS, Law EH, Jiang R, Pullenayegum E, Shaw JW, Xie F, et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*. 2019;22(8):931–41.
- Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Econ*. 2018;27(1):7–22.
- Norman R, Cronin P, Viney R. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Appl Health Econ Health Policy*. 2013;11(3):287–98.
- Health Measures. PROMIS adult profile instrument. 2020. http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Adult_Profile_Scoring_Manual.pdf
- Hanmer J, Cella D, Feeny D, Fischhoff B, Hays RD, Hess R, et al. Evaluation of options for presenting health-states from PROMIS(R) item banks for valuation exercises. *Qual Life Res*. 2018;27(7):1835–43.
- Hanmer J, Cella D, Feeny D, Fischhoff B, Hays RD, Hess R, et al. Selection of key health domains from PROMIS((R)) for a generic preference-based scoring system. *Qual Life Res*. 2017;26(12):3377–85.
- Measures H. PROMIS® Score cut points. 2020. Cited 23 Jan 2021. <https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/promis-score-cut-points>
- Mulhern B, Feng Y, Shah K, Janssen MF, Herdman M, van Hout B, et al. Comparing the UK EQ-5D-3L and english EQ-5D-5L value sets. *Pharmacoeconomics*. 2018;36(6):699–713.
- Law EH, Pickard AS, Xie F, Walton SM, Lee TA, Schwartz A. Parallel Valuation: A Direct Comparison of EQ-5D-3L and EQ-5D-5L Societal Value Sets. *Med Decis Making*. 2018;38(8):968–82.
- McClure NS, Sayah FA, Xie F, Luo N, Johnson JA. Instrument-defined estimates of the minimally important difference for EQ-5D-5L index scores. *Value Health*. 2017;20(4):644–50.
- Mulhern B, Pan T, Norman R, Duy Tran A, Hanmer J, Viney R, et al. Are generic measures of health related quality of life comparable? Understanding the measurement relationship between EQ-5D-5L, PROMIS-29 and PROPr. EuroQol Academy Meeting, 2021.
- Dewitt B, Yu L, Hanmer J. Using item response theory to estimate EQ-5D-5L utilities from PROMIS. Society for medical decision making 42nd annual meeting. <https://smdm.confex.com/smdm/2020/meetingapp.cgi/Paper/13540>. 2020.
- Dewitt B, Jalal H, Hanmer J. Computing PROPr utility scores for PROMIS® profile instruments. *Value Health*. 2020;23(3):370–8.
- Hanmer J, Dewitt B. The development of a preference-based scoring system for PROMIS® (PROPr): a technical report. 2017. <https://janelhanmer.pitt.edu/documents/technicalreportv1.4.pdf>
- Dewitt B, Torrance GW. Incorporating mortality in health utility measures. *Med Decis Mak*. 2020;40(7):862–72.
- Sampson C, Parkin D, Devlin N. Drop dead: is anchoring at ‘dead’ a theoretical requirement in health state valuation? (Research Papers 002308). London: Office of Health Economics; 2020.
- Norman R, Mulhern B, Viney R. The impact of different DCE-based approaches when anchoring utility scores. *Pharmacoeconomics*. 2016;34(8):805–14.
- Hanmer J, Dewitt B, Yu L, Tsevat J, Roberts M, Revicki D, et al. Cross-sectional validation of the PROMIS-Preference scoring system. *PLoS ONE*. 2018;13(7):e0201093.
- Feeny D, Furlong W, Torrance G. Commentary in praise of studies that use more than one generic preference-based measure. *Int J Technol Assess Health Care*. 2019;35:257–62.
- Hanmer J. Cross-sectional validation of the PROMIS-preference scoring system by its association with social determinants of health. *Qual Life Res*. 2021;30(3):881–9.