

UCLA

UCLA Previously Published Works

Title

Estimation of a Preference-Based Summary Score for the Patient-Reported Outcomes Measurement Information System: The PROMIS®-Preference (PROPr) Scoring System

Permalink

<https://escholarship.org/uc/item/8c3497zd>

Journal

Medical Decision Making, 38(6)

ISSN

0272-989X

Authors

Dewitt, Barry
Feeny, David
Fischhoff, Baruch
et al.

Publication Date

2018-08-01

DOI

10.1177/0272989x18776637

Peer reviewed

Estimation of a Preference-Based Summary Score for the Patient-Reported Outcomes Measurement Information System: The PROMIS[®]-Preference (PROPr) Scoring System

Barry Dewitt¹, David Feeny, Baruch Fischhoff, David Cella, Ron D. Hays, Rachel Hess, Paul A. Pilkonis, Dennis A. Revicki, Mark S. Roberts, Joel Tsevat, Lan Yu, and Janel Hanmer

Abstract

Background. Health-related quality of life (HRQL) preference-based scores are used to assess the health of populations and patients and for cost-effectiveness analyses. The National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS[®]) consists of patient-reported outcome measures developed using item response theory. PROMIS is in need of a direct preference-based scoring system for assigning values to health states. **Objective.** To produce societal preference-based scores for 7 PROMIS domains: Cognitive Function–Abilities, Depression, Fatigue, Pain Interference, Physical Function, Sleep Disturbance, and Ability to Participate in Social Roles and Activities. **Setting.** Online survey of a US nationally representative sample ($n = 983$). **Methods.** Preferences for PROMIS health states were elicited with the standard gamble to obtain both single-attribute scoring functions for each of the 7 PROMIS domains and a multiplicative multiattribute utility (scoring) function. **Results.** The 7 single-attribute scoring functions were fit using isotonic regression with linear interpolation. The multiplicative multiattribute summary function estimates utilities for PROMIS multiattribute health states on a scale where 0 is the utility of being dead and 1 the utility of “full health.” The lowest possible score is -0.022 (for a state viewed as worse than dead), and the highest possible score is 1. **Limitations.** The online survey systematically excludes some subgroups, such as the visually impaired and illiterate. **Conclusions.** A generic societal preference-based scoring system is now available for all studies using these 7 PROMIS health domains.

Keywords

health-related quality of life, health utility, PROMIS, US general population

Date received: November 1, 2017; accepted: April 21, 2018

Health-related quality of life (HRQL) is often assessed with measures for specific domains, such as physical function, depressive symptoms, and social function. Such measures are used in evaluating health interventions, conducting epidemiologic studies, and monitoring population health. Measures of societal preferences for these states allow incorporating them in decision and cost-effectiveness analyses of medical interventions. Currently used preference measures include the

EQ-5D-3L/5L, Health Utilities Index (HUI) Mark 2 and Mark 3, SF-6D, and Quality of Well-Being Scale.^{1–6} Their respective strengths and weaknesses have been widely discussed.^{7–11}

Corresponding Author

Barry Dewitt, Department of Engineering & Public Policy, Carnegie Mellon University College of Engineering, 5000 Forbes Ave, Pittsburgh, PA 15213, USA (barrydewitt@cmu.edu).

Each such measure applies a scoring function that associates a number (*utility*) with each state in a *state space* of *health profiles* (or *health states*). These numbers are treated as cardinal (interval-scale) utilities, representing preferences for health.^{12,13} Various conventions are followed to create societal measures from the preferences of a sample of individuals.^{14,15}

Since 2004, the National Institutes of Health has funded the development and dissemination of the Patient-Reported Outcomes Measurement Information System (PROMIS),^{16–18} a health profile measurement system that uses item response theory (IRT) to produce efficient, well-characterized measures. PROMIS has item banks^{19,20} for many HRQL domains (e.g., pain, physical function, sleep, social activity). These item banks are freely available, customizable for specific uses, and comparable across studies.^{21,22} Here, we apply decision theory methods to estimate the utility of health states for selected PROMIS domains, so that utility scores can be used in research, population health management, and policy analyses that also use the PROMIS measures. We call the resulting scoring system the PROMIS-Preference (PROPr) scoring system.

PROPr is grounded in utility theory and designed to avoid the ceiling and floor effects sometimes observed

with other measures.¹¹ Figure 1 provides an overview of our approach. From the left, PROMIS scores (A) are inputs to PROPr single-attribute scoring functions (B) that yield utilities for each domain (C). PROPr then applies a multiattribute function to combine the single-domain scores (D) and produce a summary score (E). Hanmer and colleagues^{23–25} and the PROPr technical report²⁶ (available in the online appendix) describe the development process for PROPr. Its methods and single- and multiattribute scoring functions are described here.

Methods

The PROPr scoring system is based on the normative theory of preferences embodied in multiattribute utility theory (MAUT).^{12,13,27,28} If its underlying assumptions are met, MAUT procedures produce utility functions that can be treated as *cardinal*, meaning that they are measured on an interval scale, thereby allowing comparisons between differences in utility.²⁹ Cardinality is required to create quality-adjusted life years, which combine the utility of morbidity and mortality. The PROPr scoring system applies the MAUT-based methodology of the Health Utilities Index Mark 2 and Mark 3^{1,2,15,30} to elicit preferences for PROMIS-defined health states, using a nationally representative US sample. The next 2 sections describe the health-state space used in PROPr and the preference survey. They are followed by descriptions of the analytical methods used to produce the 7 PROPr single-attribute scoring functions and the summary multiattribute scoring function.

Health-State Space

A multidimensional health-state space includes all states that can be described by its constituent dimensions. For example, a state space might include physical function and depressive symptoms. One state in that space, (x_p, x_d) , might be $x_p =$ limited physical activity and $x_d =$ no depressive symptoms.

PROPr focuses on a set of 7 PROMIS domains, chosen to span the overall space, so that they would form a common set that would be important to the public, patients, and researchers. We also imposed the constraint, required by MAUT, that the domains be *structurally independent*, in the sense that all states could conceivably occur.¹³ For example, physical function and depression are structurally independent if one can imagine a high score on one and a low score on the other, high scores on both, and low scores on both. Two domains

Carnegie Mellon University, Department of Engineering and Public Policy, Pittsburgh, PA, USA (BD, BF); McMaster University Faculty of Social Sciences, Hamilton, ON, Canada (DF); Northwestern University Feinberg School of Medicine, Chicago, IL, USA (DC); University of California Los Angeles David Geffen School of Medicine, Los Angeles, CA, USA (RDH); University of Utah, Salt Lake City, UT, USA (RH); University of Pittsburgh Medical Center, Pittsburgh, PA, USA (PAP, LY, JH); Evidera Inc, Bethesda, MD, USA (DAR); University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA (MSR); and University of Texas Health Science Center at San Antonio, San Antonio, TX, USA (JT). Parts of these analyses were presented at the 37th Annual Meeting of the Society for Medical Decision Making in St. Louis, MO, October 18–21, 2015, as well as the 39th Annual Meeting of the Society for Medical Decision Making in Pittsburgh, PA, October 22–25, 2017. The authors thank Alexander Davis, Dennis Fryback, Murray Krahn, William Lawrence, Christopher McCabe, Simon Pickard, and Milton Weinstein for helpful discussions. Barry Dewitt was partially supported by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship. Janel Hanmer was supported by the National Institutes of Health (KL2 TR001856). Data collection was supported by the National Institutes of Health (UL1TR000005). David Cella, Ron D. Hays, Paul A. Pilkonis, and Dennis A. Revicki were supported by a grant from the National Cancer Institute (1U2C-CA186878-01) and a supplement to the PROMIS statistical center grant (3U54AR057951-04S4). Baruch Fischhoff was partially supported by the Swedish Foundation for the Humanities and Social Sciences. It should be noted that David Feeny has 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.

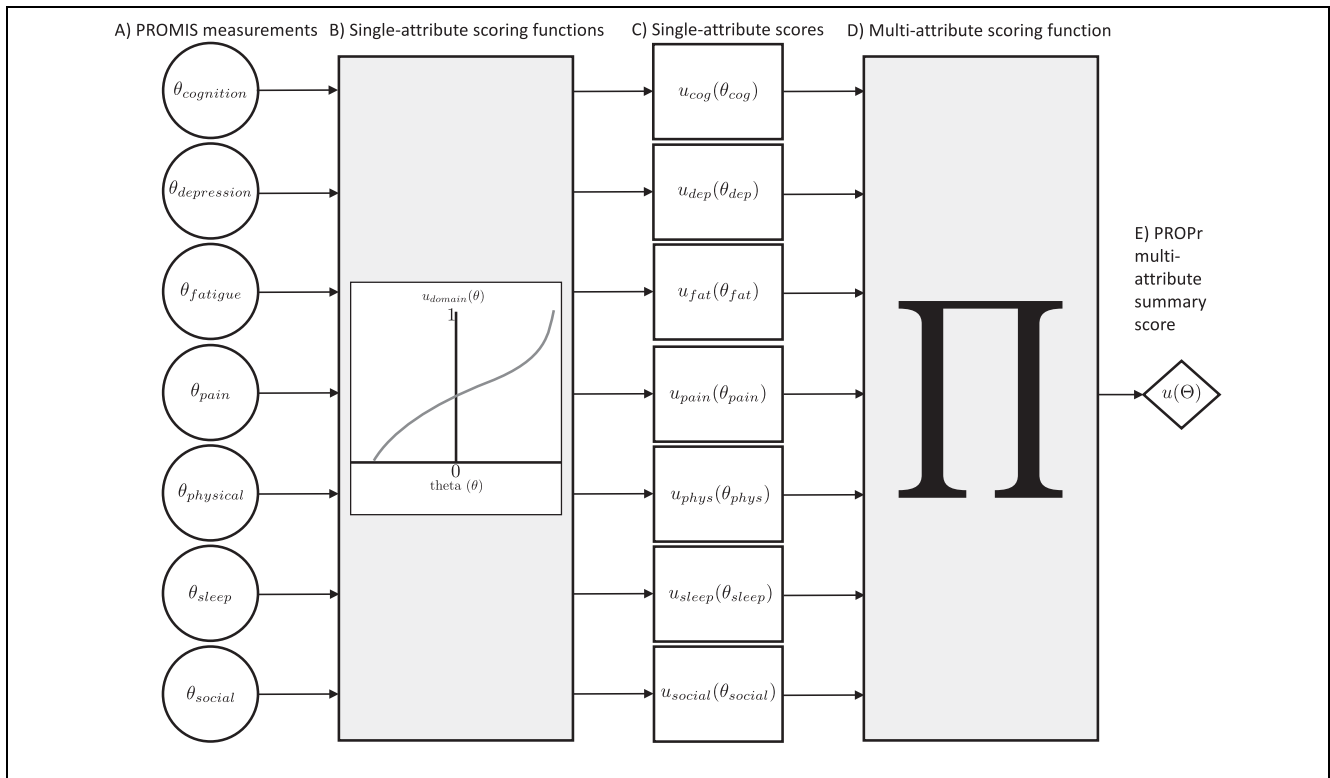


Figure 1 The PROMIS-Preference (PROPr) scoring system conceptual model. In (A), a measurement on one of the 7 PROMIS domains used in PROPr, denoted θ , is the input to its single-attribute scoring function u_{domain} . In (B), the output of $u_{domain}(\theta)$ is a score on the scale where 0 is the utility of that domain's disutility corner state and 1 is the utility of full health, the state with the highest functional capacity on all domains. If we have all 7 PROMIS measurements, then we can take the outputs from the 7 single-attribute scoring functions (C) and use them as inputs to the multiplicative multiattribute scoring function (D). The multiattribute function produces a summary score, $u(\Theta)$, for the entire vector Θ of 7 PROMIS measurements, on the scale where 0 is the utility of dead and 1 is the utility of full health (E).

can be structurally independent even if they are empirically correlated. Hanmer and colleagues²⁴ describe the procedure that selected the 7 PROMIS health domains in the PROPr state space: Cognitive Function–Abilities v2.0 (*cognition*), Depression v1.0 (*depression*), Fatigue v1.0 (*fatigue*), Pain–Interference v1.1 (*pain*), Physical Function v1.2 (*physical function*), Sleep Disturbance v1.0 (*sleep*), and Ability to Participate in Social Roles and Activities v2.0 (*social roles*). All currently available physical function item bank versions (v1.0, v1.1, v1.2, and v2.0) and pain item bank versions (v1.0 and v1.1) can be used with PROPr. PROPr requires at least v2.0 of the cognition and social roles item banks—their 1.0 versions cannot be used. When new item banks become available, the PROMIS documentation will describe whether they are compatible with those used to develop PROPr; if so, they can be used with the PROPr scoring system.

PROMIS treats each health domain as a continuous latent construct, called *theta* (in IRT). That construct is theoretically unbounded in both directions. It is expressed in *T scores*, constructed to have population mean 50 and standard deviation 10. PROPr uses a standardized transformation of T scores into *z scores*, such that the mean is 0 and the standard deviation is 1. Actual scores rarely fall outside the range -4 to 4 on *theta* (T score range = 10–90).

A functional capacity on a domain is called a *level* of *theta*. A health state (or profile) in PROPr is a vector with 7 elements, each representing a level on 1 domain. Each domain was represented by 2 items, which appear in Figure 2 (e.g., cognition was expressed as ability to concentrate and ability to remember). Levels of those items (e.g., not at all, a little bit) were chosen to represent 8 or 9 health states that spanned the space of *theta* values (Table 1; see Hanmer and Dewitt²⁶ for fuller details).

Cognition	I have been able to concentrate. . .	Not at all	A little bit	Somewhat	Quite a bit	Very much
	I have been able to remember to do things, like take medicine or buy something I needed . . .	Not at all	A little bit	Somewhat	Quite a bit	Very much
Depression	I felt unhappy . . .	Always	Often	Sometimes	Rarely	Never
	I felt that nothing was interesting . . .	Always	Often	Sometimes	Rarely	Never
Fatigue	How often were you too tired to take a bath or shower? . . .	Always	Often	Sometimes	Rarely	Never
	How often did you feel tired?	Always	Often	Sometimes	Rarely	Never
Pain	How often was your pain so severe you could think of nothing else? . . .	Always	Often	Sometimes	Rarely	Never
	How often was pain distressing to you? . . .	Always	Often	Sometimes	Rarely	Never
Physical Function	Are you able to dress yourself, including tying shoelaces and buttoning up your clothes? . . .	Unable to do	With much difficulty	With some difficulty	With a little difficulty	Without any difficulty
	Are you able to run 100 yards (100 m)? . . .	Unable to do	With much difficulty	With some difficulty	With a little difficulty	Without any difficulty
Sleep	I got enough sleep . . .	Never	Rarely	Sometimes	Often	Always
	I woke up too early and could not fall back to sleep . . .	Always	Often	Sometimes	Rarely	Never
Social Roles	I have trouble taking care of my regular personal responsibilities . . .	Always	Usually	Sometimes	Rarely	Never
	I have trouble participating in recreational activities with others. . .	Always	Usually	Sometimes	Rarely	Never

Figure 2 Health-state descriptions in the PROPr survey. Health-state descriptions were given as a table like the one above, with one answer selected for each item (row). For example, the health state describing the highest functional capacity on each domain (called *full health*) would have the rightmost column selected for all items. The health state describing the lowest functional capacity on each domain (called the *all-worst state*) would have the leftmost column selected for all items.

Table 1 PROMIS Theta Scores Used in PROPr Elicitation Tasks^a

	Highest Functional Capacity								Lowest Functional Capacity
Cognition	1.12	0.52	0	-0.37	-0.65	-0.9	-1.24	-1.57	-2.05
Depression	-1.08	-0.26	0.15	0.6	0.91	1.39	1.74	2.25	2.7
Fatigue	-1.65	-0.82	-0.09	0.3	0.87	1.12	1.69	2.05	2.42
Pain	-0.77	0.1	0.46	0.83	1.07	1.41	1.72	2.17	2.73
Physical Function	0.97	0.16	-0.21	-0.44	-0.79	-1.38	-1.78	-2.17	-2.58
Sleep Disturbance	-1.54	-0.78	-0.46	0.09	0.34	0.82	1.66	█	1.93
Social Roles	1.22	0.49	0.08	-0.28	-0.62	-0.96	-1.29	-1.63	-2.09

^aThe table shows the theta values corresponding to the health state descriptions valued in the PROPr survey. The levels between the unhealthiest and the healthiest correspond to the intermediate states valued in valuation set (i) of the elicitation task. The unhealthiest levels, together, define the *all-worst state*, while the healthiest levels, together, define *full health*. 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. Elicitations for the sleep disturbance domain had 6 health states; all others had 7.

Survey Overview

We collected preference data with an online instrument administered by ICF (<https://www.icf.com/services/research-and-evaluation>) and SurveyNow (<http://www.surveynowapp.com/>). Full descriptions appear in the online appendix and technical report.²⁶ The present analyses focus on the preference elicitation task, which was preceded by demographic questions and ones about participants' health, using the PROMIS-29 inventory and Cognition 4-item short form.^{22,31} The preference elicitation task asked each participant to evaluate states spanning the range for 1 health domain randomly chosen from the 7 PROPr domains and to evaluate several multidomain health states (as described below).

As compensation, participants who completed the survey could choose among products that included gift cards and reward program points. The ICF International Institutional Review Board approved the survey (ICF IRB FWA00002349). Responses were anonymized before the authors received them.

In pretesting, we found that participants could not thoughtfully read the essential introductory instructions and then complete the survey in under 15 minutes. Therefore, we only used data from surveys completed in at least 15 minutes.

Multiattribute Scoring Function

As mentioned, PROPr associates a cardinal utility with each health state in PROPr's 7-domain state space. MAUT specifies the models for scoring functions that map states onto interval scales in normatively justified ways.¹³ The 3 most commonly used models are the *linear additive*, *multiplicative*, and *multilinear*. They differ in their assumptions about interactions among preferences, that is, how evaluations of levels on one attribute (here, PROMIS domains) depend on the levels on other attributes. The linear additive model is the most restrictive; it assumes that preferences do not interact. The multilinear model allows pairs of attributes to be *preference complements* or *preference substitutes*.² For example, the domains of physical function and social roles would be preference complements if the magnitude of the change in utility caused by being immobile and socially isolated were greater than the magnitude of the change caused by each condition individually but less than the magnitude of the sum of the 2 individual changes. Those 2 domains would be preference substitutes if the magnitude of the change in utility caused by being immobile and socially isolated were greater than the magnitude of the sum of the 2 individual changes. The multiplicative model allows

all pairs of domains to be preference complements or substitutes but not both.³⁰ The linear additive model is a special case of the multiplicative model.

Following the methods described by Furlong and colleagues³⁰ and Feeny and colleagues,² our preference elicitation survey collected responses needed to fit a multiplicative model. The PROPr procedures evaluate the appropriateness of the linear additive model (step 3, below). Although more flexible, the multilinear model has unrealistic data requirements, in terms of sample size and participant burden.

A general multiplicative utility function u for m attributes assigns a number $u(\Theta)$ to every state $\Theta = (\theta_1, \theta_2, \dots, \theta_m)$ in its state space and has the following form:

$$u(\Theta) = \frac{1}{k} \left(\prod_{i=1}^m (1 + k \cdot k_i \cdot u_i(\theta_i)) - 1 \right), \quad (1)$$

where

$$\left(\prod_{i=1}^m (1 + k \cdot k_i) \right) - k - 1 = 0. \quad (2)$$

The k_i terms are utilities of the *corner states*, defined as ones with the best level on the i th attribute and the worst on all other attributes. The k term is the *global interaction constant*, which measures preference interactions among all the attributes: a negative value indicates that the domains are preference substitutes; a positive value indicates that they are complements.^{2,13}

Following the method described by Feeny and colleagues,² the procedure asks participants to envision *disutility corner states*, with the unhealthiest level on the i th domain and the healthiest level on all other domains (1). As a result, the PROPr function is calculated in *disutility* terms and then transformed to utility, with utility = 1 - disutility.

Preference Elicitation

Participants valued 2 sets of states, first using a visual analog scale (VAS) and then a standard gamble (SG).^{32,33} The VAS task was intended to introduce the health states to be valued in the SG task.³³ The SG task was used for PROPr because of its grounding in expected utility theory.^{12,13}

The VAS had a 0 to 100 scale (sometimes called a Feeling Thermometer), where 0 is the value of a lowest health state and 100 the value of *full health*, the state with the highest functional capacity on all domains. Figure 3

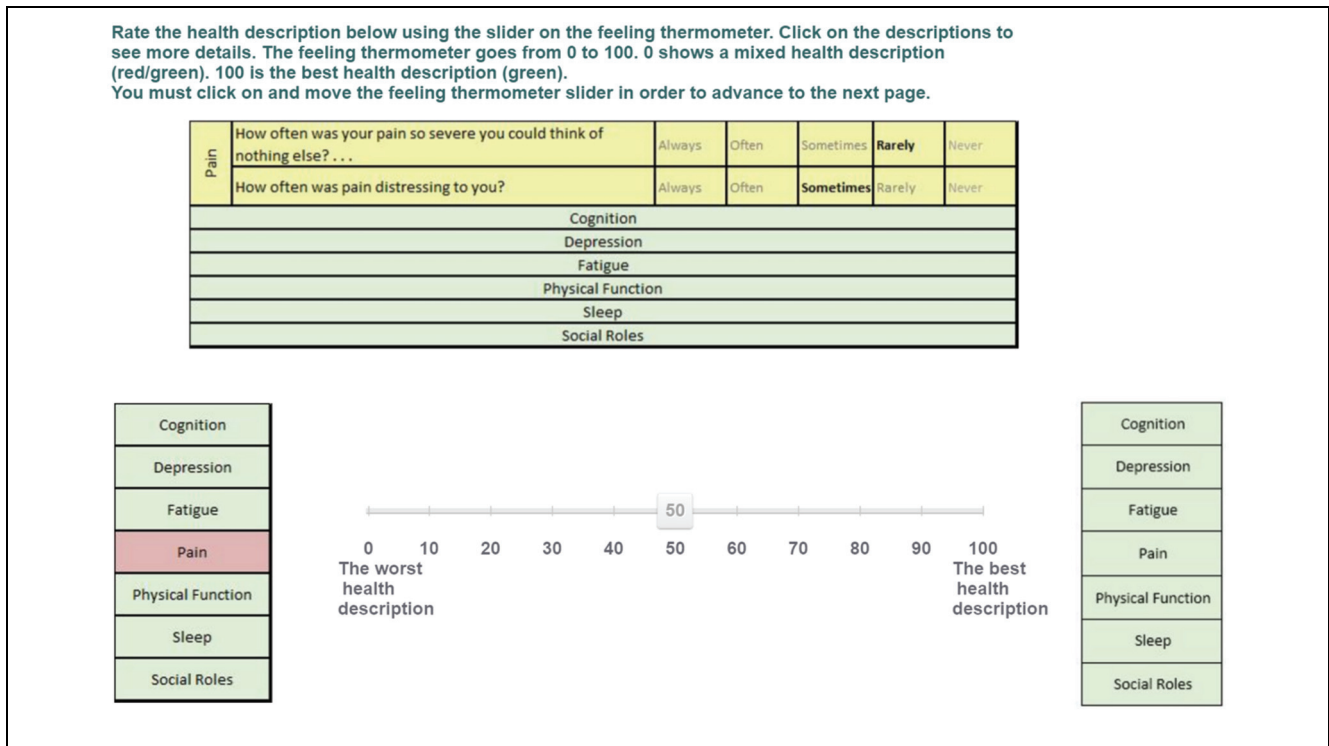


Figure 3 An example valuation, using the visual analog scale (VAS).

shows an example of the VAS, which elicits the value of an intermediate state for pain, by asking participants to rate the health state that is perfect in all respects except for rarely having pain so severe that they could think of nothing else and sometimes having pain that is distressing.

The SG task for the same intermediate health state poses a choice between a) having this state with certainty and b) a lottery with probability p of full health and $(1 - p)$ for the bottom state (see below). The SG procedure offers a series of choices, varying the probability p until the participant is indifferent between the options. Following utility theory assumptions, this probability, \hat{p} , is the utility of the intermediate state. Figure 4 shows one gamble in such a sequence, with the intermediate state described at the top right (choice B), and the gamble at the left (choice A) showing a 0.8 probability of full health and a 0.2 probability of the most severe level of pain.

(a) Set (i)

We randomly assigned participants to assess 1 of the 7 health domains (e.g., cognition). Each participant evaluated 6 or 7 states for that domain, selected to represent the intermediate theta values in Table 1 and described in verbal terms in Figure 2. The bottom state on these

valuations was always the disutility corner state for the given domain (corresponding to the unhealthiest possible level in that domain and the healthiest level on all others, as in Figure 3 and Figure 4). Figure 5A illustrates this process for the cognition domain.

(b) Set (ii)

Recognizing that participants may consider some states to be worse than dead², we asked them whether they preferred the dead state or the state with the unhealthiest level on all 7 domains (the *all-worst state*). We treated the option *not* chosen as the bottom state for the participant's valuations in this set. Participants then valued the disutility corner state for their assigned domain in set (i). They also valued 2 other states, randomly selected from the disutility corner states for the other domains, and 3 *marker states*, chosen to span the health state space.² Finally, participants valued either dead or the all-worst state, depending on which they had selected as better (Figure 5B,C).

Calculating the PROPr Scoring System

To create the PROPr scoring system, we first calculate a single-attribute scoring function for each PROMIS domain, with 0 equal to the utility of that domain's

Would you prefer the gamble on the left (choice A) or the sure thing on the right (choice B)?

Choice A		Choice B – 100% chance						
80% % chance	20% % chance	Pain	How often was your pain so severe you could think of nothing else? ...	Always	Often	Sometimes	Rarely	Never
			How often was pain distressing to you?	Always	Often	Sometimes	Rarely	Never
Cognition	Cognition	Cognition						
Depression	Depression	Depression						
Fatigue	Fatigue	Fatigue						
Pain	Pain	Physical Function						
Physical Function	Physical Function	Sleep						
Sleep	Sleep	Social Roles						
Social Roles	Social Roles							

Choice A
Choice B
 About Equal

NEXT

Figure 4 An example step in a standard gamble (SG) valuation. Choice A shows some gamble between the best and worst health states in the given domain—in this case, pain. Choice B shows the sure thing of some intermediate health state.

disutility corner state and 1 equal to the utility of full health. The 7 single-attribute functions are combined to produce a multiattribute summary scoring function, where 0 is the utility of dead and 1 is the utility of full health, with scores less than 0 corresponding to states judged worse than dead. Specifically, the creation of the PROPr scoring system follows these steps:

- (1) Estimate single-attribute disutility functions for each health domain.
- (2) Calculate the mean values of the disutility corner states.
- (3) Check the fit of the linear additive and multiplicative functional forms; calculate the global interaction constant.
- (4) Combine results from steps 1 to 3 to produce the multiattribute *disutility* function.
- (5) Transform the disutility function to a utility function, and then rescale so that the utility of dead = 0.
- (6) Perform sensitivity analyses.

Following Feeny et al.² and Furlong et al.,³⁰ we excluded the highest and lowest 5% of elicited utilities (10% *trimming*) for each health state.

1. Estimate single-attribute disutility functions for each health domain (set (i)).

Creating the PROPr scoring system required addressing 3 technical issues. One is how to estimate utilities for states between the levels of theta corresponding to the health state descriptions that participants valued (Table 1). (In previous work, single-attribute functions [e.g., HUI:2 or HUI:3] have been estimated over a discrete state space.) The second is how to translate the unbounded PROMIS scores into the bounded scales required by MAUT, which assign a utility (disutility) of 1 (0) to full health and a utility (disutility) of 0 (1) to the disutility corner state of the domain. The third is ensuring that the function be monotonically increasing with increased functional capacity (lest it lead to paying for treatments that *worsen* health).

To address these concerns, we combined isotonic regression with linear interpolation. Isotonic regression imposes monotonicity on the mean values of the dependent variable (here, utility) associated with successive values of the independent variable (here, health states) by replacing any nonmonotonic set of 2 means with their average, weighted by the number of observations involved in each. Intermediate values are estimated by connecting the means with lines.

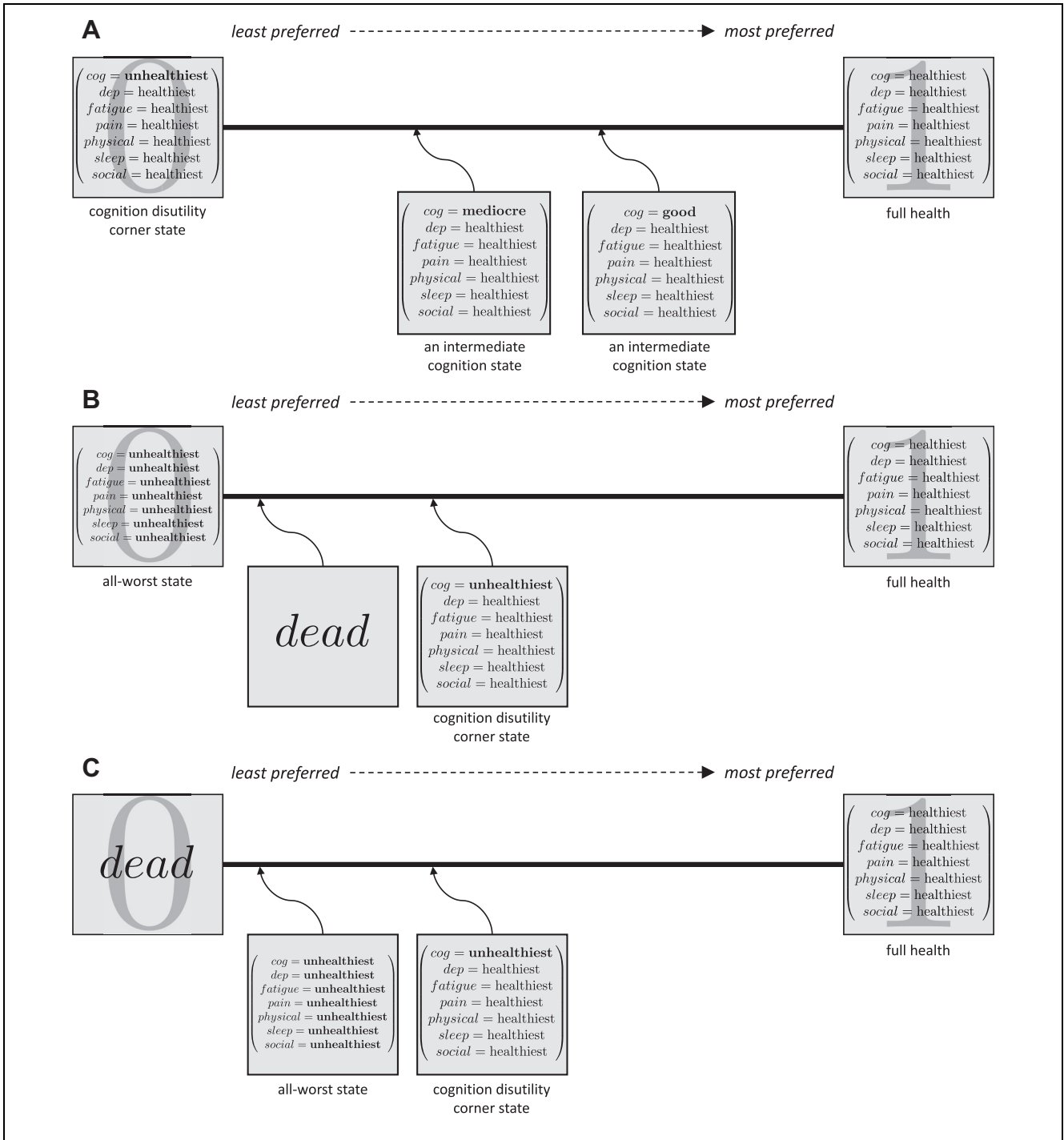


Figure 5 An example, using the cognition domain, of the data produced by the preference elicitations, in utility terms. (The associated disutility scale is produced by taking $1 - \text{utility}$.) In (A), the participant values intermediate states of cognition on a scale from the unhealthiest level of cognition (the cognition disutility corner state) to full health. In (B), a participant who prefers the state of dead to the all-worst state values dead and the cognition disutility corner state on a scale from the all-worst to full health; panel (C) shows the output of someone who prefers the all-worst state to dead. Panel (A) corresponds to set (i) in the main text and panels (B) and (C) to set (ii).

2. Calculate the mean values of the disutility corner states (set (ii)).

The mean values of the disutility corner states are calculated separately for participants who preferred the all-worst state to dead and for participants who preferred dead to the all-worst state. An affine transformation translated the values produced by the former group to the scale of the latter and then combined estimates from the 2 groups, weighting each by its size. Thus, the resulting disutilities are on a scale where the all-worst state has a disutility of 1 and full health has a disutility of 0.

3. Check the fit of the linear additive and multiplicative functional forms; calculate the global interaction constant.

MAUT determines the fit of the multiplicative and linear additive models by the sum of the k_i s, with the linear additive being superior only if that sum equals 1,¹³ in which case, the global interaction constant is 0. If the multiplicative model is superior, the global interaction constant is determined by solving equation (2), using the disutility corner state values calculated in step 2. Because equation (2) is a polynomial, it can have several real roots. MAUT offers theorems for determining which is the constant.¹³

4. Create the multiattribute disutility function.

If the multiplicative model is the better functional form, then the multiattribute disutility function uses equation (1). Written in disutility terms, it becomes equation (3):

$$\bar{u}_{AW}(\Theta) = \frac{1}{c} \left(\prod_{i=1}^7 (1 + c \cdot c_i \cdot \bar{u}_i(\theta_i)) - 1 \right). \quad (3)$$

Here, \bar{u}_{AW} is the disutility function on the all-worst to full-health scale, and Θ is the vector of PROMIS scores for a health state. The constant c is the global interaction constant, the constant c_i is the mean disutility corner state value for domain i , and \bar{u}_i is the single-attribute disutility function for that domain.

If the linear additive form is superior, the disutility function equals the sum of the \bar{u}_i , each multiplied by its respective c_i .

5. Transform the disutility function to a utility function, and rescale so that the utility of dead = 0.

The utility function $u_{AW}(\Theta)$ equals $1 - \bar{u}_{AW}(\Theta)$. Following the transformation procedure of step 2, the

disutility function is rescaled to the utility function, $u(\Theta)$, where 0 equals the utility of dead and 1 equals the utility of full health:

$$u(\Theta) = 1 - \frac{\bar{u}_{AW}(\Theta)}{\bar{u}_{AW}(dead)}, \quad (4)$$

Here, $\bar{u}_{AW}(\Theta)$ is equation (3), and $\bar{u}_{AW}(dead)$ is the mean disutility value of dead on the all-worst to full-health scale.

6. Perform sensitivity analyses.

Societal health utility measures aggregate the preferences of individuals. Those in the present sample were selected to represent the US adult population with enough vision and literacy to complete the survey. As mentioned, we excluded those who spent less than 15 minutes on the SG task.

As an additional quality control measure, for each set of utility estimates, we followed precedent and applied a 10% trimming rule,² excluding the highest and lowest 5% of values, treating them as noisy responses, reflecting inattention on that survey item. In some cases, that practice might have eliminated thoughtfully produced but unusual responses, where the health domain was particularly important or unimportant for the participant (e.g., physical function for someone who is athletic or sedentary). The 10% trimming procedure removed individual responses, not entire participants.

We did not exclude cases where SG produced “out-of-bounds” responses, below 0 or above 1, but rounded them to 0 and 1, respectively, treating them as reflecting imprecision rather than confusion.

To assess the effects of these data-handling decisions, we conducted 4 sensitivity analyses by repeating the analysis with the following:

- i. No minimum completion time threshold and 10% trimming
- ii. Fifteen-minute completion threshold and no 10% trimming
- iii. Fifteen-minute completion threshold and 10% trimming, excluding “out-of-bounds” responses (rather than adjusting them to 0 or 1)
- iv. A “stringent criteria” subsample

Case (iv) excluded participants who met any of the following exclusion criteria: spent less than 15 minutes on the survey; violated dominance more than twice; used less than 10% of the scale for all valuations; rated their understanding as less than 2 (where 1 = *Not at all* to

6 = *Very much*); had a numeracy score of less than 2.5, on a scale from 1 to 6³⁴; or rated dead or the all-worst state as equal to or better than full health. Similar exclusions have been used in other studies.³⁵

We estimated the multiattribute scoring function for the core, *analysis sample*, as defined by 10% trimming and 15-minute completion threshold, and for the 4 sensitivity analysis samples. We then applied each of these 5 functions to estimate the health utility of each participant in the analysis sample, using that individual's health profile defined using the 7 PROMIS domains, as reported on the survey's PROMIS-29 inventory and Cognition 4-item short form. As a measure of the sensitivity of the scoring function to the choice of sample, we calculated linear correlations between these 5 utility scores. As the disutility corner state values determine how the single-attribute functions are weighted in the final summary scoring function (equation (1)), we also calculated linear correlations between the disutility corner state values estimated for the analysis sample and the 4 sensitivity analysis samples.

Results

Of the 2026 individuals invited to the survey, 1779 completed the consent form (87.8%) and 1164 (57.5%) completed the entire survey. Of the 615 people who completed the consent form but not the full survey, 331 dropped out before the health state valuation section. Median survey completion time was 25 minutes, with 983 participants spending at least 15 minutes—defining the analysis sample. Overall, 630 (64.1%) participants chose dead to be better than the all-worst state and the remainder (353) the opposite. As mentioned, 10% trimming removed responses, not participants.

Sample Demographics

The sample's demographic characteristics largely match the US 2010 Census except that the analysis sample reported being slightly older, more educated, with higher income, and a larger proportion of white individuals than the US population (Table 2). In the analysis sample, reported overall health status was excellent for 12.5%, very good for 39.4%, good for 33.8%, fair for 12.4%, and poor for 1.9%.

1. Estimate single-attribute disutility functions for each health domain.

Figure 6 shows the 7 single-attribute disutility functions, where the x-axis is the construct measured on the PROMIS z score scale (theta) and the y-axis is disutility. For example, the upper left graph shows disutilities of the PROMIS cognition domain. The curves for cognition, physical function, and social roles slope downward because higher theta scores indicate higher functioning, whereas higher theta scores indicate higher symptom burden for the other domains.

2. Calculate the mean value of the disutility corner states.

The corner states had a range of disutility values (Table 3). The larger the number, the more weight that domain had in the final multiattribute utility model.

3. Check the fit of the linear v. the multiplicative functional form; calculate the global interaction constant.

The sum of the disutility corner states was 4.45, indicating a multiplicative MAUT model. Using the disutility corner state values and equation (2), and following the procedure specified in Appendix 6B of Keeney and Raiffa,¹³ the global interaction constant for that model is -0.999 . That value indicates that the domains are preference complements, as has been the case in all versions of the HUI.²

4. Create the multiattribute disutility function.

Using these estimates for the disutility corner states, the global interaction constant, and the single-attribute disutility functions, we calculated the multiattribute disutility function \bar{u}_{AW} on the all-worst to full-health scale with equation (3).

5. Transform the disutility function to a utility function, and rescale so that the utility of dead = 0.

The mean utility value of dead on the all-worst to full-health scale is 0.021. Using that value and the function $\bar{u}_{AW}(\Theta)$ (equation (3)) from step 4, the PROMIS-Preference (PROPr) multiattribute scoring function is given by $u(\Theta)$ in equation (4). After rescaling so that dead has a utility of 0, the all-worst state has a utility of -0.022 . By construction, 1 is the highest possible score.

6. Perform sensitivity analyses.

Table 2 Sample Demographics

	US 2010 Census, %	Total Sample, %	Core Sample, %
Sex		(<i>n</i> = 1164)	(<i>n</i> = 983)
Female	51.0	52.7	54.1
Male	49.0	47.0	45.8
Other	NA	0.3	0.1
Age, y			
18–24	13.0	12.0	10.0
25–34	17.0	18.0	16.0
35–44	17.0	15.0	14.0
45–54	19.0	17.0	18.0
55–64	16.0	17.0	17.0
65–74	9.0	11.0	13.0
75–84	6.0	6.0	7.0
85 +	3.0	5.0	5.0
Hispanic			
Yes	16.0	17.0	16.0
No	84.0	83.0	84.0
Race			
White	72.0	75.4	77.0
African American	12.0	12.5	11.7
American Indian	1.0	1.0	1.0
Asian	5.0	5.5	4.5
Native Hawaiian	1.0	0.2	0.2
Other	6.0	3.2	3.6
Multiple races	3.0	2.2	2.0
Education for those age 25 and older		(<i>n</i> = 1029)	(<i>n</i> = 888)
Less than high school graduate	13.9	11.9	12.2
High school graduate or equivalent	28.0	26.3	26.8
Some college, no degree	21.0	21.7	21.5
Associate's degree	7.9	6.9	7.0
Bachelor's degree	18.0	19.4	19.4
Graduate or professional degree	11.0	13.8	13.2
Income			
Less than \$10,000	2.0	3.7	3.4
\$10,000 to less than \$15,000	4.0	3.5	3.8
\$15,000 to less than \$25,000	14.0	10.3	10.5
\$25,000 to less than \$35,000	17.0	15.8	15.9
\$35,000 to less than \$50,000	20.0	18.5	17.8
\$50,000 to less than \$65,000	15.0	16.4	16.9
\$65,000 to less than \$75,000	6.0	6.0	6.2
\$75,000 to less than \$100,000	10.0	11.1	11.0
\$100,000 or more	12.0	14.7	14.6
	Quota, %	Total, %	Core, %
Self-rated health			
Excellent	NA	14.9	12.5
Very good	NA	38.7	39.4
Good	NA	33.1	33.8
Fair	NA	11.5	12.4
Poor	NA	1.8	1.9

NA, not applicable

We repeated steps 1 to 5 for each of the 4 sensitivity analysis samples. The 5 resulting multiattribute scoring functions were then applied to the health states reported by the 983 participants. Linear correlations between the

individual utilities estimated with the 5 scoring functions were all ≥ 0.98 ($P < 0.001$). The disutility corner state values estimated using the 4 alternative samples were correlated above 0.90 with those estimated using the analysis

Table 3 Mean Disutility Values of the Disutility Corner States, the c_i in Equation (3)

Domain	Disutility (All-Worst State = 1, Full Health = 0)
Physical function	0.688
Depression	0.666
Pain	0.653
Fatigue	0.639
Cognition	0.635
Social roles	0.611
Sleep	0.563

sample (all $P < 0.01$), except for case (iii) (removal of out-of-bounds responses), where the correlation was 0.76 ($P = 0.046$).

Discussion

This article describes the development of 7 single-attribute scoring functions and a multiplicative multiattribute summary scoring function for 7 PROMIS domains: Cognitive Function–Abilities, Depression, Fatigue, Pain–Interference, Physical Function, Sleep Disturbance, and Ability to Participate in Social Roles and Activities. We call this scoring system the PROMIS-Preference (PROPr) scoring system. The single-attribute functions and the multiattribute summary score can be used to compare groups or to track groups over time. For the multiattribute scoring function, 0 is the utility of dead and 1 is the utility of full health. For the single-attribute scoring functions, 0 corresponds to the utility of the state with the unhealthiest level on a domain and the healthiest levels on all other domains (i.e., the disutility corner state of that domain), and 1 corresponds to the utility of full health.

The 7 single-attribute functions suggest that (dis)utility is a nonlinear function of the PROMIS scores (Figure 6). That result is consistent with research showing that nonlinear models typically provide better fits for full (multiattribute) scoring functions.^{10,36} PROPr is, we believe, the first method capable of observing nonlinearity for individual domains. A linear utility function for a PROMIS domain would imply that utility for that domain is the same as the domain construct itself, which is generally not the case for other constructs (e.g., utility of money). The form of nonlinearity, reflected in the different slopes of the line segments of the single-attribute functions, varies by domain, even though the states on each domain cover a similar range of functional

capacity. For example, the single-attribute function for social roles changes abruptly in the mid-range of theta, while the single-attribute function for fatigue has constant slope for a large portion of its range.

The disutility corner state values are all similar (Table 3). A methodological interpretation of this result is that enough participants had enough difficulty with the SG task to blur distinctions among these states. A substantive interpretation is that participants believe that the disutility corner states would similarly affect their overall HRQL. That similarity could reflect the success of our attempt to choose the most important domains²⁴ and to represent each with values that span its range (Table 1). Intuitively, the disutility corner states describe such low levels of functioning that the utilities assigned to them plausibly could be very close.

Several limitations should be considered when interpreting the findings of this study. First, only 57% of invited participants completed the entire survey. Although relatively few of those who dropped out did so during the SG task (37 of 615 = 6%), there is always the concern that losing participants who find the task particularly challenging removes individuals with systematically different preferences.

Second, participants were recruited from an online panel.³⁷ As part of its efforts to recruit a representative sample, the survey company released invitations in waves to ensure the final sample's demographic characteristics matched the 2010 US Census. As noted, the final sample's demographic characteristics generally matched those of the US adult population on several variables potentially related to health utilities.

Third, because we used a community sample, the utilities of individuals who have experienced ill health on each domain are reflected only to the extent of their prevalence in the population. The choice of sample is an ethical question,^{38,39} with uncertain empirical implications.^{40–48} Our choice of a community sample reflects the recommendations of the Second Panel on Cost-Effectiveness in Health and Medicine.⁴⁹

Fourth, we excluded some participants based on their responses. Our analysis sample excluded individual participants who took fewer than 15 minutes, the minimum time needed for thoughtful responses. We also excluded (“trimmed”) individual responses in the top or bottom 5% of the utility distribution for each health state. These exclusion criteria sought to balance external validity (having a more representative sample) and internal validity (having better quality responses). Sensitivity analyses found that the multiattribute scoring function for the analysis sample produced similar utility estimates for

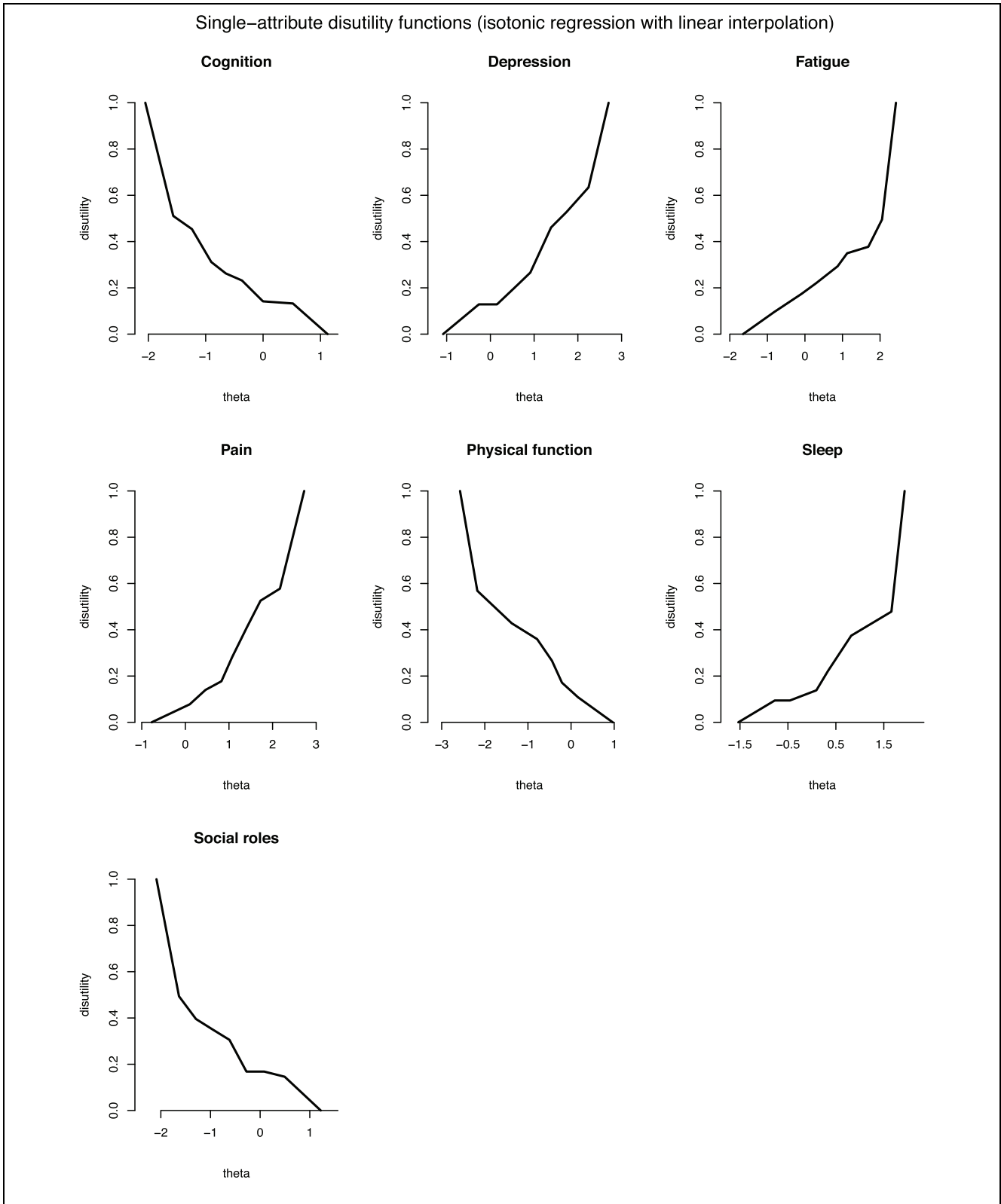


Figure 6 Single-attribute disutility functions. Isotonic regression with linear interpolation modeling the conditional mean disutility for each level of theta corresponding to Table 1.

participants' health states as the scoring functions created with the 4 samples using other exclusion criteria. Estimates for corner state disutilities were similar as well. Nonetheless, it is possible that some trimmed responses reflect thoughtful but uncommon valuations,^{35,50} a topic for future research.

Finally, the current procedure asks participants whether they prefer the all-worst state or dead and then uses the worse of the 2 as the origin for some valuations. To place all responses on a common scale, calculations for the 2 groups were done separately and then combined, weighting by group size. An alternative approach is to transform each participant's valuations individually. We did not use that approach because it would exclude more responses.

The PROMIS measures provide greater granularity than other approaches to describing health states.⁵¹ PROPr inherits this granularity when producing utilities. The PROMIS measures also avoid 2 problems commonly observed with other methods: substantial ceiling effects in the general population and floor effects in unhealthy populations.^{20,50} The range of the PROMIS domains included in PROPr was chosen to avoid these effects as well; future work should verify this assumption. By using health states that represent the range of PROMIS scores, PROPr should be applicable to studies using these domains, whatever their specific design. An important future evaluation of the PROPr scoring system will be to compare the preference scores derived from surveys composed of different sets of PROMIS items for the PROPr health domains.


PROPr can also address an issue that has proven difficult with earlier systems, quantifying the statistical uncertainty in its utility estimates.⁵² IRT allows estimating the precision of the PROMIS questions used to elicit individuals' health states, which can then be propagated into their PROPr scores.

We offer a general societal preference-based scoring system for 7 selected PROMIS health domains. Clinical, population, and health services research studies that use these PROMIS domains can use PROPr to estimate preference-based scores. Thus, PROPr links IRT-based health state measures (PROMIS) with utility theory, allowing a more unified assessment of health outcomes for clinical and health policy studies. In the spirit of PROMIS, PROPr seeks to make health valuation as easy as possible for researchers, clinicians, and policy makers. Standardized code is available, at no cost, for users of R and SAS for calculating PROPr scores.²⁶

Supplementary Material

Supplementary material for this article is available on the *Medical Decision Making* Web site at <http://journals.sagepub.com/home/mdm>.

ORCID iD

Barry Dewitt  <https://orcid.org/0000-0003-1622-6736>

References

1. Torrance GW, Feeny D, 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(7):702–22.
2. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Med Care*. 2002;40(2):113–28.
3. EuroQol Group. EuroQol—a new facility for the measurement of health related quality of life. *Health Policy (New York)*. 1990;(16):199–208.
4. Herdman M, Gudex C, Lloyd A, 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.
5. Kaplan RM, Anderson JP, Ganiats TG. The quality of well-being scale: Rationale for a single quality of life index. In: Walker SR, Rosser RM, editors. *Quality of life assessment: Key issues in the 1990s*. Dordrecht: Springer; 1993.
6. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-12. *J Health Econ*. 2002;21(9):271–92.
7. Tosh J, Brazier J, Evans P, Longworth L. A review of generic preference-based measures of health-related quality of life in visual disorders. *Value Health*. 2012;15(1):118–27.
8. McNamee P, Seymour J. Comparing generic preference-based health-related quality-of-life measures: advancing the research agenda. *Expert Rev Pharmacoecon Outcomes Res*. 2005;5(5):567–82.
9. Kaplan RM, Tally S, Hays RD, et al. Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. *J Clin Epidemiol*. 2011;64(5):497–506.
10. Feeny D, Krahn M, Prosser LA, Salomon JA. Valuing health outcomes—online appendices. 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; 2016. p 167–99.
11. Fryback DG. Measuring health-related quality of life. Paper prepared for the Workshop on Advancing Social Science Theory: The Importance of Common Metrics. National Academies, Washington, DC, February 25–26, 2010.
12. von Neumann J, Morgenstern O. *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press; 1944.

13. Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley; 2003.
14. Dewitt B, Davis A, Fischhoff B, Hanmer J. An approach to reconciling competing ethical principles in aggregating heterogeneous health preferences. *Med Decis Making*. 2017;37:647–56.
15. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Oper Res*. 1982;30(6):1043–69.
16. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap Cooperative Group during its first two years. *Med Care*. 2007;45(5):3–11.
17. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*. 2007;16(Suppl. 1):133–41.
18. Cella D, Riley W, Stone A, 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.
19. Embretson S, Reise SP. *Item Response Theory for Psychologists*. Hillsdale, NJ: Lawrence Erlbaum; 2000.
20. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(1):S22–S31.
21. Collins FS, Riley WT. NIH's transformative opportunities for the behavioral and social sciences. *Sci Transl Med*. 2016;8:366ed14.
22. Gershon RC, Rothrock N, Hanrahan R, Bass M, Cella D. The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. *J Appl Meas*. 2010;11(3):304–14.
23. Hanmer J, Feeny D, Fischhoff B, et al. The PROMIS of QALYs. *Health Qual Life Outcomes*. 2015;13:122.
24. Hanmer J, Cella D, Feeny D, et al. Selection of key health domains from PROMIS® for a generic preference-based scoring system. *Qual Life Res*. 2017;26(12):3377–85.
25. Hanmer J, Cella D, Feeny D, et al. Evaluation of options for presenting health-states from PROMIS® item banks for valuation exercises. *Qual Life Res*. 2018;doi:10.1007/s11136-018-1852-1
26. Hanmer J, Dewitt B. PROMIS-Preference (PROPr) score construction—a technical report. 2017. Available from: janelhanmer.pitt.edu/PROPr.html
27. Luce RD, Suppes P. Preferences, utility, and subjective probability. In: Luce RD, Bush RR, Galanter E, eds. *Handbook of Mathematical Psychology*. New York: John Wiley; 1965. p 249–410.
28. Savage LJ. *The Foundations of Statistics*. 2nd rev. New York: Dover; 1972.
29. Coombs CH, Dawes RM, Tversky A. *Mathematical Psychology: An Elementary Introduction*. New York: Prentice-Hall; 1970.
30. Furlong W, Feeny D, Torrance GW, et al. *Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report*. Hamilton, ON: McMaster University Centre for Health Economics and Policy Analysis; 1998. Report No.: 98–11.
31. PROMIS. Applied Cognition—Abilities. 2015. Available from: http://www.healthmeasures.net/administrator/components/com_instruments/uploads/PROMIS%20Short%20Form%20v2.0-Cognitive%20Abilities%20Subset%204a%2010-20-2016.pdf
32. Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health Serv Res*. 1994;29(2):207–24.
33. Torrance GW, Feeny D, Furlong W. Visual analog scales. *Med Decis Making*. 2001;21(4):329–34.
34. McNaughton CD, Cavanaugh KL, Kripalani S, Rothman RL, Wallston KA. Validation of a short, 3-item version of the Subjective Numeracy Scale. *Med Decis Making*. 2015;35(8):932–6.
35. Engel L, Bansback N, Bryan S, Doyle-Waters MM, Whitehurst DGT. Exclusion criteria in national health state valuation studies: a systematic review. *Med Decis Making*. 2016;36(7):798–810.
36. van der Pol M, Currie G, Kromm S, Ryan M. Specification of the utility function in discrete choice experiments. *Value Health*. 2014;17(2):297–301.
37. Tourangeau R, Plewes TJ, ed. *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: National Academies Press; 2013.
38. Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: a call to reconsider current guidelines. *Soc Sci Med*. 2016;165:66–74.
39. Dolan P. Whose preferences count? *Med Decis Making*. 1999;19(4):482–6.
40. Dolders MGT, Zeegers MPA, Groot W, Ament A. A meta-analysis demonstrates no significant differences between patient and population preferences. *J Clin Epidemiol*. 2006;59(7):653–64.
41. Mulhern B, Rowen D, Snape D, et al. Valuations of epilepsy-specific health states: a comparison of patients with epilepsy and the general population. *Epilepsy Behav*. 2014;36C:12–7.
42. Neumann PJ, Goldie SJ, Weinstein MC. Preference-based measures in economic evaluation in health care. *Annu Rev Public Health*. 2000;21:587–611.
43. Feeny D, Blanchard C, Mahon JL, et al. Comparing community-preference-based and direct standard gamble utility scores: evidence from elective total hip arthroplasty. *Int J Technol Assess Health Care*. 2003;19(2):362–72.
44. Feeny D, Furlong W, Saigal S, Sun J. Comparing directly measured standard gamble scores to HUI2 and HUI3 utility scores: group- and individual-level comparisons. *Soc Sci Med*. 2004;58(4):799–809.

45. Feeny D, Wu L, Eng K. 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.* 2004;13(10):1659–70.
46. Krahn M, Ritvo P, Irvine J, et al. Preferences for outcomes patient and community in prostate cancer: implications for clinical policy. *Med Care.* 2003;41(1):153–64.
47. Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. *Value Health.* 2010;13(2):306–9.
48. Zethraeus N, Johannesson M. A comparison of patient and social tariff values derived from the time trade-off method. *Health Econ.* 1999;8:541–5.
49. Sanders GD, Neumann PJ, Basu A, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second Panel on Cost-Effectiveness in Health and Medicine. *JAMA.* 2016; 316(10):1093–103.
50. Wittenberg E, Prosser LA. Ordering errors, objections and invariance in utility survey responses: a framework for understanding who, why and what to do. *Appl Health Econ Health Policy.* 2011;9(4):225–41.
51. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim J-S. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Med Decis Making.* 2010;30(1):5–15.
52. Chan KKW, Xie F, Willan AR, Pullenayegum E. Underestimation of variance of predicted health utilities derived from multiattribute utility instruments: the use of multiple imputation as a potential solution. *Med Decis Making.* 2017;37(3):262–72.