

Exclusion Criteria as Measurements II: Effects on Utility Functions

Barry Dewitt , Baruch Fischhoff, Alexander L. Davis, Stephen B. Broomell, Mark S. Roberts, and Janel Hanmer 

Background. Researchers often justify excluding some responses in studies eliciting valuations of health states as not representing respondents' true preferences. Here, we examine the effects of applying 8 common exclusion criteria on societal utility estimates. **Setting.** An online survey of a US nationally representative sample ($N = 1164$) used the standard gamble method to elicit preferences for health states defined by 7 health domains from the Patient-Reported Outcomes Measurement Information System (PROMIS[®]). **Methods.** We estimate the impacts of applying 8 commonly used exclusion criteria on mean utility values for each domain, using beta regression, a form of analysis suited to double-bounded scales, such as utility. **Results.** Exclusion criteria have varied effects on the utility functions for the different PROMIS health domains. As a result, applying those criteria would have varied effects on the value of treatments (and side effects) that change health status on those domains. **Limitations.** Although our method could be applied to any health utility judgments, the present estimates reflect the features of the study that produced them. Those features include the selected health domains, standard gamble method, and an online format that excluded some groups (e.g., visually impaired and illiterate individuals). We also examined only a subset of all possible exclusion criteria, selected to represent the space of possibilities, as characterized in a companion article. **Conclusions.** Exclusion criteria can affect estimates of the societal utility of health states. We use those effects, in conjunction with the results of the companion article, to make suggestions for selecting exclusion criteria in future studies.

Keywords

exclusion criteria, study design, health state valuation, preference-based measures

Date received: August 14, 2018; accepted: June 5, 2019

Utility-based measures of health-related quality of life provide quantitative estimates of preferences for health states and are commonly used in cost-effectiveness and cost-utility analyses, decision analyses, clinical trials, and population health studies.¹ Here, we address a problem that the creators of such measures often face: applying exclusion criteria to remove responses that appear not to reflect true preferences, a process that Engel and colleagues² have shown can often remove a substantial proportion of the collected data, sometimes more than half. A companion article examines 10 common exclusion criteria in terms of how and why they agree and disagree about which responses to treat as unacceptable.³ Here, we consider the effects of applying 8 of these criteria on mean societal valuations of health states. We propose a general method, illustrated with utility data for 1 widely used set

of health-state measures, the Patient-Reported Outcomes Measurement Information System (PROMIS[®]).

PROMIS, an initiative of the National Institutes of Health, offers psychometrically constructed scales for eliciting self-reported health states on many domains.⁴ The PROMIS-Preference (PROPr) Scoring System⁵ creates societal utility scores for 7 PROMIS domains: Cognitive Function–Abilities (*cognition*), Emotional Distress–Depression (*depression*), Fatigue (*fatigue*), Pain–Interference (*pain*), Physical Function (*physical function*), Sleep Disturbance (*sleep*), and Ability to

Corresponding Author

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

Participate in Social Roles and Activities (*social roles*). PROPr also offers a multiattribute utility function⁶ for estimating a single health utility score from these 7 domains. Following convention,^{7–9} those utilities reflect the responses of representative samples of the general public to questions using the standard gamble (SG) method.

Two features of single-domain utility functions (utility curves) determine their impact on health policy analyses: their elevation (absolute value), showing how much intermediate health states are valued relative to the worst and full health states, and their sensitivity (curvature), showing how utility changes with changes in health status. Exclusion criteria that increase elevation potentially reduce the value of interventions designed to improve a given health state and the aversiveness of side effects that degrade it.¹⁰ Exclusion criteria that reduce the elevation could do the opposite. Exclusion criteria that increase the curvature of a health utility curve increase the value of treatment that moves people to better health states and the aversiveness of side effects that move people to poorer states. Exclusion criteria that result in flatter curves do the opposite.

We focus on single-domain utility functions because they are the input data used to calculate multiattribute utility scores. PROPr's multiattribute scoring system for its 7 domains applies some of the exclusion criteria studied below: the removal of extreme responses and the responses of those who completed its associated data collection survey in less than 15 minutes. Dewitt et al.⁵ analyzed the effects of several exclusion criteria on the multiattribute score. That sensitivity analysis complements the analysis here, which reveals the effects of exclusion on mean utility estimates without the extra structure required to produce

the multiattribute score, in terms that are meaningful to those who might use them (i.e., cost-effectiveness analysts). That structure can obscure the effects of exclusion criteria on the included preferences, by requiring, for example, single-domain utility functions to go through 0 and 1 at prescribed points. Focusing on the single-domain utility functions allows us to see the variety of effects with different combinations of health domains and exclusion criteria.

The next section introduces the 8 exclusion criteria and the PROPr survey. We then explain the modeling approach, beta regression; apply it to the PROPr survey responses; discuss implications; and, offer recommendations for evaluating exclusion criteria.

Methods

Data

Our analyses use data from the PROPr Scoring System survey, described more fully in references 5 and 11–13, the companion article,³ and section A in the Supplementary Appendix. Briefly, 1164 participants were sampled to be representative demographically of the US general population. They evaluated health states on 1 of 7 PROMIS health domains. The visual analog scale (VAS) was completed first to familiarize them with the domains and was followed by the SG, which was used to estimate the PROPr health state utilities, given its normative properties.¹⁴ We focus on the SG responses here. Participants were randomly selected to evaluate 1 of the 7 health domains. Depression and social roles were evaluated by 167 participants and the other domains by 166. Participants also evaluated other health states, such as dead and the all-worst state. They answered several other tasks as well, described in the other sources.

Exclusion Criteria

Exclusion criteria seek to distinguish true preferences from confused, inattentive, or strategic (deliberately biased) ones. Criteria can be *preference-based*, reflecting a respondent's choices (e.g., unusually high values), or *process-based*, reflecting how respondents produced them (e.g., too quickly to be thoughtful).

Table 1 shows 10 criteria, selected to represent the space of commonly invoked rationales, including both preference-based and process-based ones; section B in the Supplementary Appendix shows more examples. The companion article³ applies multidimensional scaling (MDS) to characterize these criteria in terms of how similarly they select participants for exclusion. Two criteria, *low-range* and *no-variance*, are nested, in the sense

Department of Engineering & Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA (BD, BF, ALD); the Institute for Politics and Strategy, Carnegie Mellon University, Pittsburgh, PA, USA (BF); Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, USA (SBB); Division of General Internal Medicine, University of Pittsburgh, Pittsburgh, PA, USA (MSR, JH); and Department of Health Policy and Management, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA (MSR). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This research was completed at the Division of General Internal Medicine, the University of Pittsburgh, and the Department of Engineering & Public Policy, Carnegie Mellon University. Barry Dewitt received partial support from a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship. Janel Hanmer was supported by the National Institutes of Health through grant KL2 TR001856. Data collection was supported by the National Institutes of Health through grant UL1TR000005. Baruch Fischhoff and Barry Dewitt were partially supported by the Swedish Foundation for the Humanities and Social Sciences. The funding agreements ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Table 1 Core Exclusion Criteria

Exclusion Criteria (<i>shorthand</i>)	Requirements for Exclusion
Score on the Subjective Numeracy Scale of less than 2.5 (<i>numeracy</i>)	A participant scored less than 2.5 on the 3-item short form of the Subjective Numeracy Scale. ¹⁵
Self-assessed understanding equal to 1 or 2, on a scale of 1 = <i>not at all</i> to 5 = <i>very much</i> (<i>understanding</i>)	A participant rated themselves a “1” or a “2” on the self-assessed understanding question, which occurred after the preference elicitations.
15-minute time threshold (<i>time</i>)	A participant completed the PROPr survey in under 15 minutes.
Violated dominance on the SG (<i>violates-SG</i>)	A participant, using the standard gamble (SG), violated dominance at least once.
Violated dominance on the VAS (<i>violates-VAS</i>)	A participant, using the visual analog scale (VAS), violated dominance at least once.
Valued the all-worst state or dead as the same or better than full health (<i>dead-all-worst</i>)	A participant valued the all-worst state or dead as the same or better than full health, using the SG.
Used less than 10% of the utility scale (<i>low-range</i>)	A participant’s valuations, using the SG, represent less than 10% of the range of the utility scale.
Provided the same response to every SG (<i>no-variance</i>)	A participant valued every state the same, using the SG.
In the top 5% of responses for an SG (<i>upper-tail</i>)	A response falls in the upper 5% of responses for a health state, using the SG.
In the bottom 5% of responses for an SG (<i>lower-tail</i>)	A response falls in the bottom 5% of responses for a health state, using the SG.

Core exclusion criteria, implemented with the PROMIS-Preference (PROPr) data. Unless otherwise indicated, valuations refer to the valuations of the single-domain states. Unshaded rows indicate preference-based criteria; shaded rows indicate process-based criteria.

that they apply the same rule, one more stringently than the other. Here, we use only *low-range*, which subsumes *no-variance*. We exclude 1 criterion (*violates-VAS*) that does not apply to the SG but that might be examined with the present analytical framework in a comparison of the 2 elicitation procedures.

Previous studies have considered varied health domains and exclusion criteria and found mixed results.² Most have focused on violations of dominance, with some finding that applying criteria had little effect on the multiattribute utility model^{16–18} and some finding large effects.^{19,20} Similarly varied results have been found when applying criteria to calculating the mean value of specific health states.^{2,16,19} The results below complement these studies by modeling the utility for combinations of sets of health states and exclusion criteria, each selected to represent their universe—health states in PROPr and exclusion criteria in the companion article.³

Beta Regression

A single-domain utility function assigns a value of 0 to the worst possible outcome and 1 to the best. (See Table

A1 in the Supplementary Appendix for the scale values corresponding to utilities of 0 and 1 for each domain.)

Double-bounded variables exhibit properties that make them difficult to model using normal-theory regression, such as substantial skew and heteroskedasticity. Several regression methods have been developed to model bounded data. In health utility applications, the Tobit model and censored least absolute deviations (CLAD) model are common.²¹ Here, we use beta regression. Both Tobit and CLAD assume censored data, where values outside the bounds are theoretically possible but not observed because of the measurement procedure (e.g., tests that bound knowledge or ability at 100% scores). In contrast, the utility values of 0 and 1 are theoretical bounds, in the sense that more extreme values do not exist, by definition. CLAD has the additional limitation of estimating medians, rather than the means typically used in health utility analyses.

Beta regression models variance and skew directly,²² assuming that, conditional on each regressor (predictor or covariate), the dependent variable follows a beta distribution $Beta(\omega, \tau)$, defined over (0,1) by two shape parameters, $\omega > 0$ and $\tau > 0$. That distribution can assume many shapes. For example, when $\omega = \tau = 1$, it becomes

the uniform distribution; when $\omega = \tau > 1$, it is bell-shaped (but truncated at 0 and 1). In general, ω pulls the density toward 1, and τ pulls it toward 0, producing skewed distributions when the 2 are unequal.

The probability density function of a beta random variable $y \sim \text{Beta}(\omega, \tau)$ is given by

$$f(y, \omega, \tau) = \frac{\Gamma(\omega + \tau)}{\Gamma(\omega)\Gamma(\tau)} y^{\omega-1} (1-y)^{\tau-1},$$

where $\Gamma(\cdot)$ is the complete gamma function. The mean is

$$E(y) = \frac{\omega}{\omega + \tau}$$

and the variance is

$$\text{Var}(y) = \frac{E(Y)(1 - E(Y))}{\omega + \tau + 1}.$$

We follow Paolino,²³ who provided an alternative parametrization that has now become standard.^{22,24,25} If $\mu = E(y)$ and $\phi = \omega + \tau$, then $\omega = \mu\phi$ and $\tau = \phi - \mu\phi$. Therefore, $\text{Var}(y) = \frac{\mu(1-\mu)}{\phi+1}$, making the variance a function of both μ and ϕ . The parameter ϕ is called the *precision* of the distribution (and ϕ^{-1} the *dispersion*), because variance increases as ϕ decreases. In models predicting health state utilities, the health states and exclusion criteria are the regressors. We focus on modeling the (conditional) mean, which is typically used in health policy analyses.^{5,8,26-28}

In PROMIS, health states are expressed as values of *theta* (a parameter in item response theory), which are constructed from responses of the PROMIS reference population, such that $\theta = 0$ for the mean response, and a 1-unit change in θ equals the standard deviation. The PROMIS reference population is close enough to the general US population²⁹ to interpret these values as probability-sample estimates for that population. Larger θ values describe better functioning for 3 domains (cognition, physical function, social roles) and more symptoms for 4 (depression, fatigue, pain, sleep disturbance).

As with generalized linear models (e.g., logistic regression), beta regression uses a link function^{30,31} to connect the statistic being modeled with the regressors, so that both are unbounded. For the mean (μ), the most frequently used link function is the logit ($\log(\frac{\mu}{1-\mu})$), producing model coefficients that reflect log-odd changes for μ . For ϕ , the link function is frequently the natural logarithm (i.e., $\log(\phi)$).

One limit to beta regression is that the dependent variable cannot equal 0 or 1, because the link function maps

the random variable to the entire real line and the logit is undefined at those values. For data sets with 0 and 1 values, the convention is to squeeze the data,²² by applying the transformation $\frac{y(n-1) + 0.5}{n}$, where y is a dependent value (possibly 0 or 1) and n is the sample size. Doing so transforms all data, unlike a transformation that affects only the endpoints (e.g., adding $\epsilon > 0$ to any 0 and subtracting it from any 1). By applying this transformation to all data,²² the squeeze transformation preserves the ratios of distances between each pair of data points, treating the data as interval scaled, as is assumed for utility.^{28,32,33} Sections C3-C5 in the Supplementary Appendix report the sensitivity of the present results to the choice of transformation.

Beta Regression Models for Health State Utilities

A beta model is fully specified by 2 parameters: its mean and its precision. If responses are conditionally beta distributed, then the mean and precision characterize the entire response distribution. Under that assumption, our beta regression models for an exclusion criterion are:

$$\log(\mu_{\text{criterion, domain}}) = \beta_0 + \beta_1 \text{theta}_{\text{domain}} + \beta_2 \text{criterion} + \beta_3 \text{theta}_{\text{domain}} : \text{criterion}$$

Equation 1 Beta regression model for the logit of the conditional mean (μ), as a function of the health domain (θ) and an exclusion criterion.

$$\log(\phi_{\text{criterion, domain}}) = \zeta_0 + \zeta_1 \text{theta}_{\text{domain}} + \zeta_2 \text{criterion} + \zeta_3 \text{theta}_{\text{domain}} : \text{criterion}$$

Equation 2 Beta regression model for the log of the conditional precision (ϕ), as a function of the health domain (θ) and an exclusion criterion.

Here, μ and ϕ are the mean and precision parameters for the beta distribution, respectively; *theta* is a continuous variable representing health states; and *criterion* is a dummy variable equal to 1 if a response is excluded and 0 otherwise. As mentioned, we focus on the effects of applying exclusion criteria to mean utilities (i.e., Equation 1). We also focus on intermediate health states and do not use the endpoints of the health domains to estimate the single-domain functions. The utility values of those endpoints were fixed in the survey and not elicited from participants.

In the model for the mean (Equation 1), β_0 (the intercept or constant) gives the mean log-odds utility for included responses, when θ is 0 (the mean population

health status on that domain); β_1 gives the change in log-odds utility for a 1-unit (1 standard deviation) change in theta for included responses; β_2 gives the difference in the intercept for excluded responses; and $\beta_3 + \beta_1$ gives the change in log-odds utility for a 1-unit change in theta for excluded responses, so that β_3 is the difference in slope (on the log-odds scale) between the included and excluded groups.

Any coefficient involving a *theta* term estimates the slope of a best-fit line on the log-odds utility scale (and the curvature of the corresponding line on the utility scale). The greater the slope (curvature on the utility scale), the more sensitive estimated utilities are to changes in theta. The lower the intercept, the lower the utility of the health state describing the population average (theta = 0) and the lower the utility of all health states, given a fixed curvature.

As these estimates are for log-odds utility, the estimate for mean utility is

$$\mu_{\text{criterion, domain}} = \frac{e^\eta}{1 + e^\eta}$$

Equation 3 Equation for the mean μ on the utility scale.

where $\eta = \beta_0 + \beta_1\text{theta}_{\text{domain}} + \beta_2\text{criterion} + \beta_3\text{theta}_{\text{domain} : \text{criterion}}$. See Section C in the Supplementary Appendix for more details on the beta regression models used here.

To estimate these coefficients, we use the **betareg** package in R.²⁴ Equation 1 models the parameters as a linear function of theta, from utilities elicited for 6 or 7 values of theta for each domain. As one test of goodness-of-fit, our sensitivity analyses include models that treat theta as a factor (i.e., a categorical) variable. See Section C3–C5 in the Supplementary Appendix for these and additional sensitivity analyses, including ones that use a more flexible mixture-model procedure, called *zero-one inflated beta regression*, which treats responses of 0 and 1 separately, removing the need to squeeze the data.

By analyzing Equation 1 for all domain-criterion pairs, we make judgments for how mean preferences differ between groups excluded by each criterion, analyzing the magnitude and direction of the effects across domains. As each domain was evaluated by a different sample, the 7 domains can be seen as 7 implementations of the criteria with different samples undertaking the same survey, with only the domain differing between them. We then combine those results with those in the companion piece, in which we analyze exclusion criteria as binary classifiers, to provide recommendations for readers planning on applying exclusion criteria or

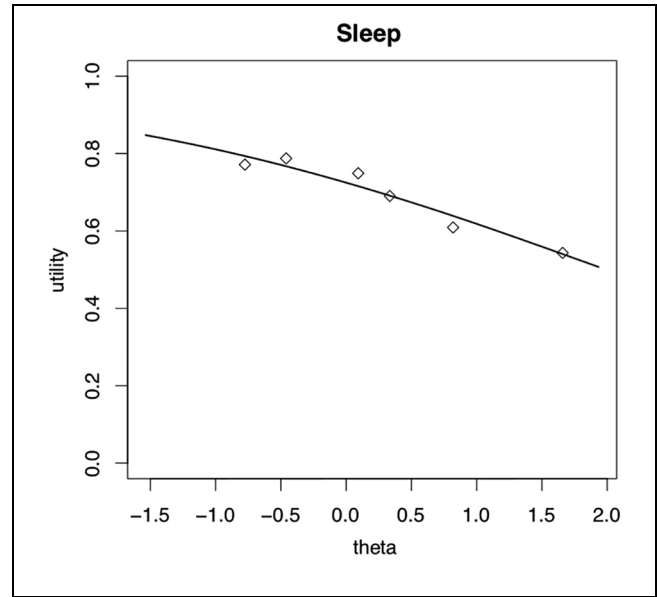


Figure 1 Modeling mean sleep disturbance utilities as a function of the health states. The solid curve is the line of best fit for the model treating health states as a continuous variable (i.e., theta in item response theory). The diamonds are the result of treating the health states as factors (i.e., a categorical variable)

interested in using our approach to evaluate their own criteria or improve survey design.

Results

For expository purposes, we first model utility as a function of theta for all responses (i.e., with no exclusions) for 1 domain, sleep disturbance (Table 2; Figure 1). We then repeat the analyses applying 2 exclusion criteria, one process-related, *numeracy* (Table 2; Figure 2), and one preference-related, *violates-SG* (Table 2; Figure 3).

The first column of Table 2 shows regression coefficients for the mean model for all responses to sleep disturbance states (i.e., Equation 1 without the criterion variable). The entries are on the logit (log-odds) scale, so an entry of value x equals $\text{logit}^{-1}(x) = \frac{e^x}{1 + e^x}$ on the utility scale. We explain each value in turn.

The value of the constant in the regression table is the log-odds utility (0.969) of the health state described by theta = 0 (the population average), which is sleep of moderate quality. That equals a utility of 0.725 (on the 0-1 utility scale). The coefficient on theta shows how log-odds utility decreases as sleep disturbance worsens (and theta increases). For example, moving from theta = 0 to

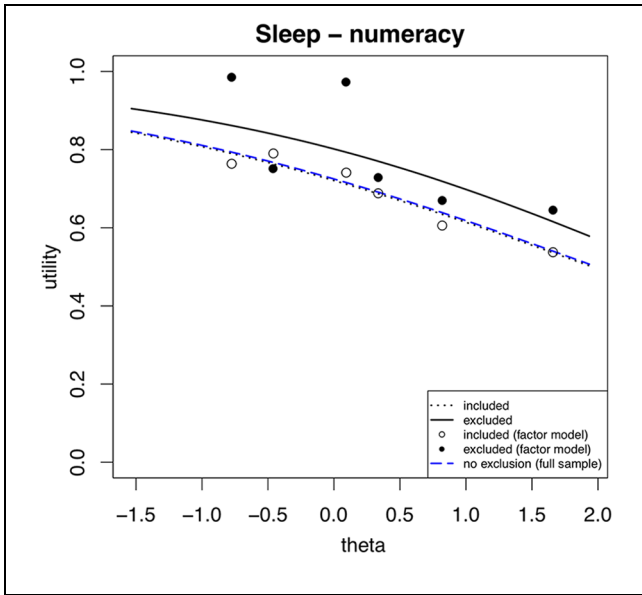


Figure 2 Modeling sleep utilities as a function of health states and the *numeracy* criterion, treating health states as continuous (lines) and as factors (dots).

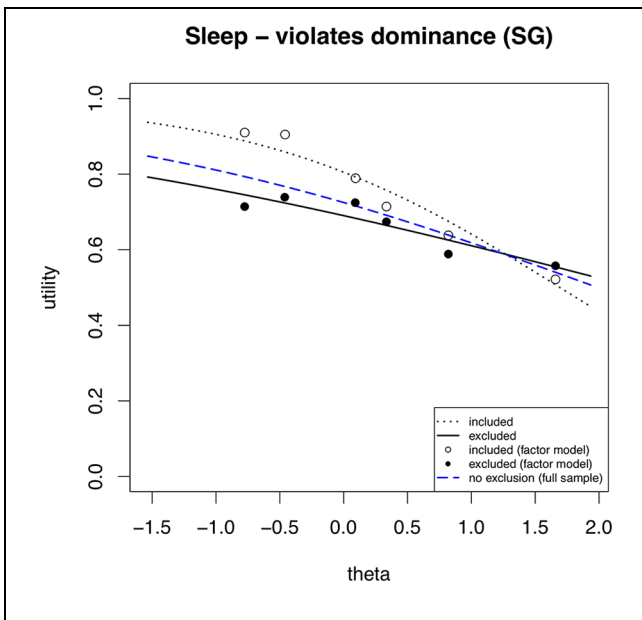


Figure 3 Modeling sleep utilities as a function of health states and the *violates-SG* criterion, treating health states as continuous (lines) and as factors (dots).

theta = 1 reduces utility from 0.725 to 0.618 [= $\text{logit}^{-1}(0.969 - 0.487) = \text{logit}^{-1}(0.482)$.] As the units are in log-odds, the change in utility caused by a 1-unit

change in theta depends on where it occurs on the theta scale. Figure 1 shows the conditional mean curve estimated from the model. It also shows the associated factor model (the diamonds), treating the health states as categorical rather than continuous variables.

The *numeracy* exclusion criterion discards all responses of any participant who scores below 2.5, after averaging the 3 questions (scored 1-6) on the short form of the Subjective Numeracy Scale (second column of Table 2).^{15,34} Figure 2 shows the effects of applying this criterion in 2 ways. The first applies beta regression separately to the included and excluded responses, seen in the dotted and solid black lines, respectively. The second is the factor model, which presents conditional means of included and excluded responses for each theta value separately, seen in the open and solid black dots, respectively.

The regression models find that participants excluded by numeracy have higher utility for sleep, for all values of theta, compared with participants who have a high score on the numeracy test. That is, those who are excluded reported utility values for intermediate sleep states closer to the utility of the best sleep state. The same result holds for the factor model, except for one value of theta. Given the greater stability of the regression models, which incorporate all data, we focus on them but discuss the factor model in sensitivity analyses (see sections C3 and C4 in the Supplementary Appendix). The dashed blue curve is the regression for the full sample, as in Figure 2. The error in estimating the curves depends on the number of responses in each group and their variability (see Table 3 as well as Section C1 and C2 in the Supplementary Appendix).

The constant corresponds to the utility of a theta score of 0 for participants not excluded by the numeracy criterion (dummy = 0). The log-odds value of 0.948 (the constant in Table 2) equals 0.721 on the 0-1 utility scale. The log-odds value of -0.484 for the theta coefficient says that 1 standard deviation of worse sleep disturbance—for example, from theta = 0 to theta = 1—reduces the estimated mean utility from 0.721 [= $\text{logit}^{-1}(0.948) = 0.721$] to 0.614 [= $\text{logit}^{-1}(0.948 - 0.484 \times 1) = \text{logit}^{-1}(0.948 - 0.484) = \text{logit}^{-1}(0.464)$].

The numeracy coefficient indicates the extent to which the excluded group (solid line in Figure 2) assigned higher values to sleep quality and the extent to which excluding them reduces the societal utility of sleep quality.

The coefficient for the theta: numeracy interaction term equals the difference in the change in predicted mean utility as theta changes for the groups included and excluded by *numeracy*. As seen in Figure 2, the sensitivity of the excluded group (solid line) is only slightly more

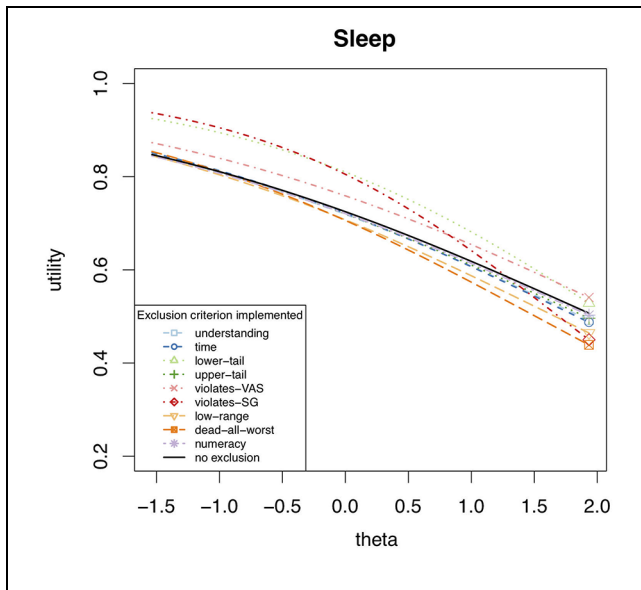


Figure 4 The estimated conditional mean utility curve for sleep, after applying the indicated exclusion criterion (or none). Note the y-axis begins at 0.2, to magnify the utility curves.

pronounced than that of the nonexcluded group (dotted line). The closeness of the models for the full sample (dashed blue) and its nonexcluded subgroup (dotted line) reflects the relatively small number excluded by *numeracy* (Table 3) and the small interaction term.

The *violates-SG* exclusion criterion discards all responses of participants who assign a higher utility to any health state than to one describing a higher level of functioning or lower level of symptom burden. Figure 3 and the third column of Table 2 show the results of applying *violates-SG* to judgments for the sleep disturbance domain. Unlike *numeracy*, for which the small theta-criterion interaction term and the small number of excluded participants mean relatively parallel utility curves, here, both the interaction term and the number of excluded participants are much larger, producing different utility functions. Figure 3 shows that excluded participants assigned lower values to high sleep quality, similar values to moderate sleep quality, and higher values to poor sleep quality. Figure 4 shows the full sample curve and reduced sample (included) curves, applying each criterion to the sleep domain responses.

Figure 5 compares the full sample and reduced sample curves for all domains, applying each exclusion criterion (details in the Supplementary Appendix, Figures C1-C8). Tables C1-C7 in the Supplementary Appendix summarize the regression coefficients for all domain and criteria combinations.

The patterns revealed in the sensitivity analysis were generally similar to those in the main analysis (see sections C3-C5 in the Supplementary Appendix).

Discussion

We begin by discussing the implications of these results for the worked example of sleep disturbance and then summarize other results, in the form of patterns found across health domains and potential recommendations for selecting exclusion criteria.

Sleep Disturbance Example

As seen in Table 2 and Figure 2, applying *numeracy* lowers the utility curve for sleep disturbance. That could give sleep-related treatments higher priority because each level of sleep is less satisfactory and leaves more room for improvement. If a policy decision is sensitive to the difference, then investigators would need to decide why the excluded responses were different. If less numerate participants were simply less able to perform the SG task, then excluding them might be justified, by arguing that their preferences are better represented by the responses of society's more numerate members. However, if the excluded participants genuinely assign higher utility to all health states, then excluding them misrepresents societal utilities and inappropriately increases the value of treatments that improve sleep quality. The analysis of exclusion patterns in the companion paper suggests that the former is the case. The similarity of the slopes of the curves for the full and reduced samples means, however, that exercising the criterion would not affect decisions that depend on treatment effectiveness (or side effects), captured in the *change* in utility across health states. The small number of excluded participants (5.4%) mitigates the effect of wrongfully excluding (or including) those participants.

Similarly, applying the *violates-SG* exclusion criterion increases the value of treatments for very poor sleep, because that part of the curve is lower for nonexcluded responses. On the other hand, it decreases the value of treatments that make good sleep even better, because that part of the curve is higher. The steeper slope of the utility curve after applying *violates-SG* makes a unit of improved sleep more valuable, no matter where it occurs. The companion article suggests that participants excluded by *violates-SG* struggled with the task but were earnestly engaged. Because 64.5% of responses violated this criterion (Table 3), applying it implies a tradeoff between potentially not representing the sample and potentially not representing the preferences of those in it.

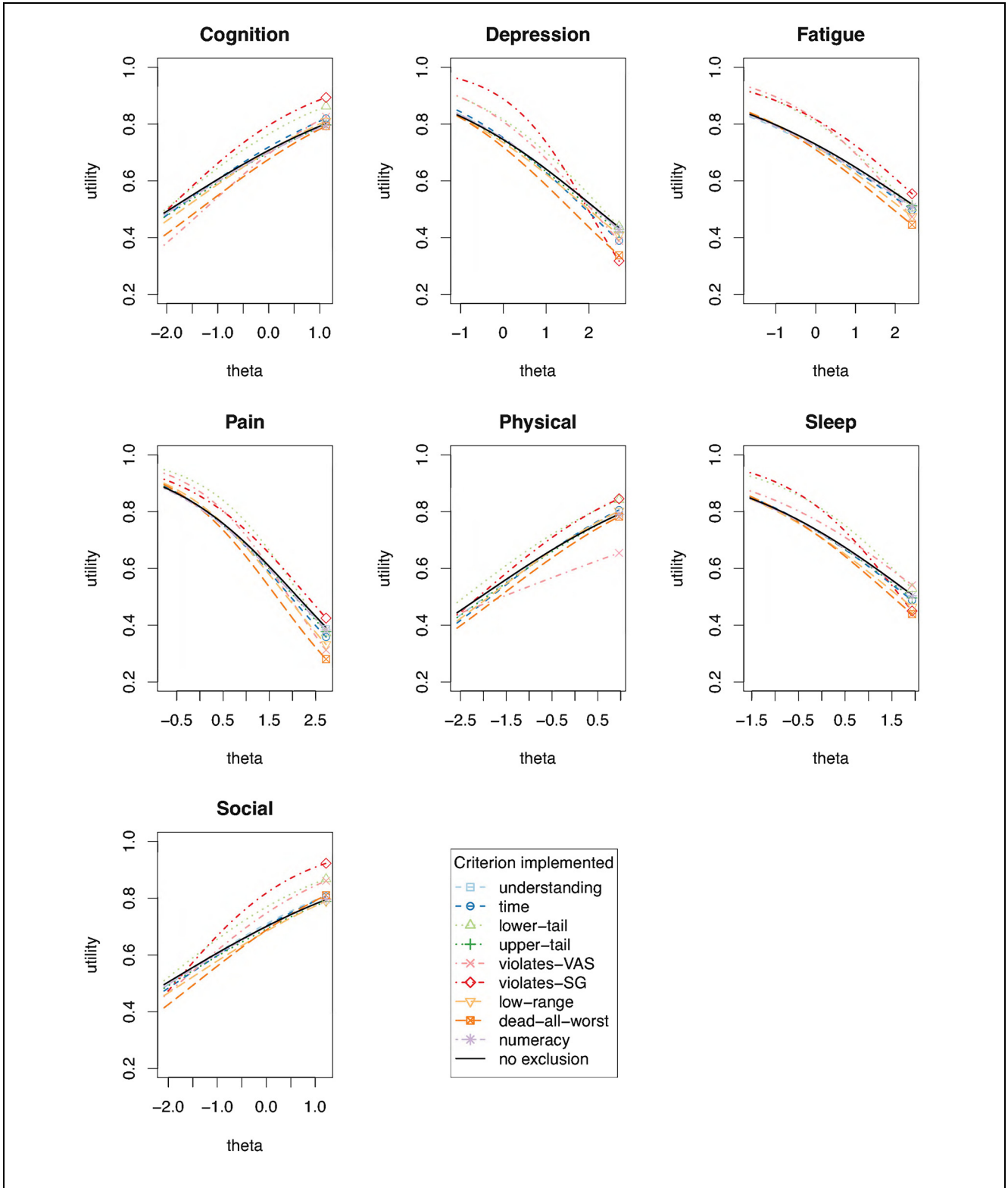


Figure 5 The estimated conditional mean utility for each domain, after applying each exclusion criterion (or none). Note the y-axis starts at 0.2, to magnify the utility curves.

Table 2 Modeling Mean Utilities for the PROMIS Sleep Disturbance Domain

	Dependent Variable		
	Log-Odds Utility		
	(1)	(2)	(3)
Constant (intercept)	0.969*** (0.050)	0.948*** (0.051)	1.419*** (0.093)
theta	-0.487*** (0.056)	-0.484*** (0.057)	-0.837*** (0.096)
numeracy		0.448* (0.241)	
theta:numeracy		-0.073 (0.260)	
violates-SG			-0.618*** (0.111)
theta:violates-SG			0.486*** (0.118)
Observations	996	996	996
R ²	0.076	0.081	0.111
Log likelihood	1561.564	1563.459	1581.657

The first column shows the model with no exclusion criterion (utility as a function of theta only). The second column shows the model with the *numeracy* criterion. The third column shows the model with the *violates-SG* criterion.

* $P < 0.1$; *** $P < 0.01$.

Table 3 Proportion of Participants Flagged for Exclusion by Each Criterion, by Domain

Exclusion Criterion (% Excluded in Total)	Cognition ($n = 166$)	Depression ($n = 167$)	Fatigue ($n = 166$)	Pain ($n = 166$)	Physical Function ($n = 166$)	Sleep ($n = 166$)	Social ($n = 167$)
<i>numeracy</i> (7.8%)	8.4%	9.0%	9.0%	12.7%	4.2%	5.4%	6.0%
<i>understanding</i> (14.3%)	17.5%	10.8%	14.5%	14.5%	15.1%	12.0%	15.0%
<i>time</i> (15.6%)	12.0%	17.4%	16.9%	16.3%	17.5%	13.9%	15.0%
<i>violates-SG</i> (71.6%)	72.3%	74.9%	72.3%	71.1%	77.7%	64.5%	68.3%
<i>violates-VAS</i> (84.7%)	85.5%	80.8%	88.6%	80.1%	89.8%	85.5%	82.6%
<i>dead-all-worst</i> (28.0%)	28.9%	25.7%	26.5%	30.7%	24.7%	28.9%	30.5%
<i>low-range</i> (12.2%)	12.7%	7.2%	15.1%	15.7%	9.6%	13.3%	12.0%
<i>no-variance</i> (11.8%)	12.0%	6.6%	14.5%	15.1%	9.0%	13.3%	12.0%
<i>upper-tail</i> (78.5%)	78.9%	77.8%	80.1%	74.1%	77.1%	83.7%	77.8%
<i>lower-tail</i> (44.1%)	42.2%	44.9%	45.8%	42.8%	52.4%	38.0%	42.5%

The proportion of participants in the PROMIS-Preference (PROPr) data flagged by each criterion from Table 1, per domain. Each column label is 1 of the 7 PROPr domains, with the number of participants assigned to value that domain in parentheses, with the sum = 1164. Each row is one of the core criteria (Table 1), with the percentage of all participants excluded by each criterion in parentheses. Unshaded rows indicate preference-based criteria; shaded rows indicate process-based criteria.

Recommendations

Exclusion criteria assume that excluded individuals are better represented by the preferences of the participants who remain than by the ones that they themselves reported. In this and the companion paper, we have analyzed commonly used criteria, in terms of whom they exclude and how they affect utility estimates. Here, we summarize our recommendations for using each exclusion criterion, by examining all criterion-by-domain pairs in Figures C1–C8 of the Supplementary Appendix.

The need to consider exclusion criteria at all means that more inclusive elicitation procedures are needed. Our recommendations appear in Table 4.

We begin with the 3 process-related criteria.

Process-related exclusion criteria

Numeracy. Applying *numeracy* produces a lower utility curve for all domains except depression. As mentioned, the companion article concluded that less numerate participants' difficulty with the task produces artifactually high estimates. We recommend excluding responses of less numerate participants. Given their small number in the demographically diverse PROPr sample (7.8%), however, the choice might make little practical difference, as in the sleep example (Figure 2).

Time. This criterion excludes participants who spent less than 15 minutes taking the survey, deemed the minimum for thoughtful responses, based on pretests. Across

Table 4 Summary of Recommendations for Exclusion Criteria

Exclusion Criteria (<i>Shorthand</i>)	Recommendations
Score on the Subjective Numeracy Scale of less than 2.5 (<i>numeracy</i>)	We endorse this criterion. However, a researcher must consider any problems with representing the preferences of less numerate individuals with their more numerate counterparts.
Self-assessed understanding equal to 1 or 2, on a scale of 1 = <i>not at all</i> to 5 = <i>very much</i> (<i>understanding</i>)	We do not endorse this criterion, as it appears likely that it captures conscientious participants.
15-minute time threshold (<i>time</i>)	We endorse this criterion, as its rationale (inattention) is supported by its empirical effects.
Violated dominance on the SG (<i>violates-SG</i>)	We do not endorse this criterion. Our results suggest it captures many who are engaged with the task.
Valued the all-worst state or dead as the same or better than full health (<i>dead-all-worst</i>)	We endorse this criterion. It represents the most egregious violation of dominance, and our analysis suggests a response process for it that is different from <i>violates-SG</i> and more likely to produce responses that are not preferences.
Used less than 10% of the utility scale (<i>low-range</i>)	We recommend this criterion and more stringent versions of it (e.g., <i>no-variance</i>). Our results support the claim that it captures inattentive responses.
In the top 5% of responses for an SG (<i>upper-tail</i>)	We do not endorse the criterion—usually combined with <i>lower-tail</i> —because of the mismatch between the basis for it and our empirical results.
In the bottom 5% of responses for an SG (<i>lower-tail</i>)	We do not endorse the criterion—usually combined with <i>upper-tail</i> —because of the mismatch between the basis for it and our empirical results.

SG = standard gamble. We summarize our recommendations for each criterion, based on our results from this article and its companion.³ Note that any criterion includes the risk of wrongful exclusion. The magnitude of that risk is partly a function of the number of participants affected by the criterion. The extent to which that varies across studies is an empirical question.

the 7 health domains, the utility curves for those excluded by this criterion were not consistently higher or lower than those for the remaining participants. However, they were consistently flatter. Although that response pattern could reflect insensitivity to health states, the analyses in the companion article suggest that these participants were inattentive. We recommend excluding them. Including them would produce inappropriately flat utility curves, diminishing the value of treatments that improve health states—and underestimating the importance of side effects that degrade health states. They represent 15.6% of the PROPr sample.

Understanding. This criterion excludes participants who reported not understanding the task. Applying it would have little effect on the elevation of the utility curves, other than lowering the utility curve for fatigue while slightly increasing the slope for all domains. The similarity in the curves suggests that those who reported not understanding the task may have set a higher standard for themselves, rather than actually experiencing more difficulty. As a result, we recommend including them, even if the individuals involved are uncertain that they should be included. They represent 14.3% of the PROPr sample.

Preference-related criteria

Dead-all worst. This criterion excludes participants whose utility for the dead or all-worst state was not lower than the utility for the best health state. These participants had systematically higher and flatter curves than the others. The companion article (Box 1) suggests how the mechanics of the SG interface might have inadvertently led to unduly high responses. We recommend excluding these participants. They represent 28.0% of the PROPr sample.

Violates-SG. This criterion excludes participants who rated at least 1 health state more highly than another strictly better health state. For all 7 domains, the utilities of these participants were lower than those of other participants, for most theta values. Their responses showed less curvature on all domains except fatigue and pain. However, the mean utility curves of these participants do not violate dominance, in the sense of decreasing, rather than increasing, as health states improve. The lack of an overall effect suggests that the individual violations reflect the noisiness of a challenging task, consistent with a previous finding that violations are more likely with more similar health states.^{2,35} That interpretation provides one explanation for our *violates-SG* and *dead-all-worst* results

in both articles: the former could be capturing many engaged participants who are trying to distinguish similar health states, whereas the latter violation involves such distant health states that it is unlikely an engaged participant would produce it. Given the large number of participants with at least 1 such violation (71.6% of the PROPr sample), we recommend not applying this exclusion criterion.

Upper-tail and lower-tail. These criteria exclude the highest and lowest 5% of responses, for each health state. As they are the only criteria we considered that exclude individual responses and not entire individuals, we examined differences between those who are eligible for exclusion by these criteria and those who are not. Most studies that apply these criteria combine them, in a procedure known as 10% trimming. However, the companion article found that they identify different response processes. By definition, *upper-tail* excludes participants with the highest utility values at a given state, whereas *lower-tail* excludes participants with the lowest—but they say nothing about their utilities for the states at which they are not among the extremes. For cognition, depression, pain, physical function, and sleep disturbance, those excluded by *upper-tail* have less sensitive utility curves (and equally sensitive ones for the other 3 domains). For cognition, depression, fatigue, pain, sleep, and social roles, *lower-tail* excludes participants whose curves are less sensitive to changes in health states (and equally sensitive for physical function). One difference not captured by the regressions is that a much higher percentage of responses fall in the upper tail than in the lower tail, 78.5% versus 44.1%. Both percentages are much higher than 5% because of ties. Because they disqualify so many responses, standard practice is to sample at random enough eligible responses to reach 5% of the total sample. The large number captured by each criterion supports the need for improved elicitation methods; see also Box 1 in the companion article³. We recommend not applying them, because of the mismatch between their combined rationale and our empirical results (i.e., that they do not act symmetrically).

Low-range. This criterion excludes participants who used less than 10% of the utility scale. Their utility curve is necessarily less sensitive to health states. As a result, removing them increases the sensitivity of the utility curve. Highly similar responses could mean either insensitivity to the health states or inattention. As noted in the companion article, most of these responses were 1s, suggesting that participants rushed through the survey

and hence were inattentive. As a result, we recommend using this exclusion criterion. It applies to 12.2% of the PROPr sample.

Conclusion

Exclusion criteria for health state preference surveys seek to identify responses that are not valid representations of participants' preferences. In this article and the companion one, we offer an approach to assessing the properties of exclusion criteria and their impacts on utility estimates. We demonstrate the approach with responses from a nationally representative US sample, evaluating health states on 7 domains from the PROMIS inventory, producing the PROPr scoring system.

The approach has 2 components. The first, in the companion article,³ uses MDS to characterize the agreement among criteria regarding whom to include and exclude. Applied to the PROPr data, it found differences between the usual rationales of criteria and their empirical effects, such as when 2 criteria that are typically combined have quite different exclusion patterns ("trimming" the highest and lowest 5% of responses).

The second component of our approach, described here, estimates the impact of applying exclusion criteria on health state utilities. It uses beta regression, a procedure suited to modeling double-bounded variables, such as health utility. Applied to the PROPr data, the beta regression analyses found that some criteria had little impact, because relatively few responses were involved or preferences were similar for the included and excluded groups. It also found that some criteria affected the elevation of health utility functions (hence, the acceptability of those health states) or their sensitivity to changes in health state (hence, the importance of changes).

Applying these 2 methods clarifies who is excluded by an exclusion criterion and how it affects the resulting societal health utility estimates. That clarification should help researchers make informed tradeoffs between data quality and sample representativeness. It should also help them to inform policy analysts and policy makers how data analytic choices affect health utility estimates and decisions using them.


In addition to contributing new methodologies, the MDS and beta regression results extend previous ones. In their systematic review, Engel et al.² found only 1 study that analyzed the effects on utility models of exclusion criteria other than violations of dominance.¹⁹


Nevertheless, our specific results are limited to the exclusion measures studied, the sample (nationally representative of the US), the 7 health state domains, the

measure (PROMIS), the elicitation procedure (SG, preceded by VAS), administration method (online), and implementation (see sample screenshots, Figures A2 and A3 in the Supplementary Appendix). The effect of changing any of these features is an empirical question.

The SG is attractive because it is rigorously grounded in utility theory.³⁶ Given some participants' apparent difficulty with the SG, we encourage additional research designed to improve the method, especially for online implementation, with its potential for efficient elicitation from large, representative samples. The need for exclusion criteria is primarily attributable to 2 related sources: inattentive participants and difficult survey items. We can reduce exclusions by improving the accessibility of our stimuli, which could include more warm-up exercises that train participants to use the stimuli to communicate their preferences. Our methodology offers a systematic way to evaluate alternative designs, whether those be new implementations of widely used methods or wholly new preference elicitation mechanisms. The better people can understand their tasks and translate their preferences into those terms, the less need there will be to worry about exclusion criteria.

ORCID iDs

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

Janel Hanmer  <https://orcid.org/0000-0001-6159-2482>

Supplemental Material

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

References

1. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. *JAMA*. 1995;273(1):59–65.
2. 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.
3. Dewitt B, Fischhoff B, Davis A, Broomell SB, Roberts MS, Hanmer J. Exclusion criteria as measurements I: Identifying invalid responses. 2019.
4. 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.
5. Dewitt B, Feeny D, Fischhoff B, et al. Estimation of a preference-based summary score for the Patient-Reported Outcomes Measurement Information System: the PROMIS®-Preference (PROPr) scoring system. *Med Decis Making*. 2018;38(6):683–98.
6. Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley & Sons; 1976.
7. Neumann PJ, Saunders GD, Russell LB, Siegel JE, Ganiats TG, eds. *Cost-Effectiveness in Health and Medicine*. 2nd ed. New York: Cambridge University Press; 2016.
8. 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.
9. Rabin R, Charro F De. EQ-SD: a measure of health status from the EuroQol Group. *Ann Med*. 2001;33(5):337–43.
10. Nord E. Cost-value analysis of health interventions: introduction and update on methods and preference data. *Pharmacoeconomics*. 2014;33(2):89–95.
11. Hanmer J, Dewitt B. PROMIS-Preference (PROPr) score construction—a technical report. 2017. Available from: janelhanmer.pitt.edu/PROPr.html
12. 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.
13. Hanmer J, Cella D, Feeny D, Fischhoff B, Hays RD, Hess R, et al. Evaluation of options for presenting health-states from PROMIS® item banks for valuation exercises. *Qual Life Res*. 2018;27(7):1835–43.
14. Feeny D. A utility approach to the assessment of health-related quality of life. *Med Care*. 2000;38(9 Suppl):II151–4.
15. 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.
16. Lamers LM, Stalmeier PFM, Krabbe PFM, Busschbach JJ V. Inconsistencies in TTO and VAS values for EQ-5D health states. *Med Decis Making*. 2006;26(2):173–81.
17. Johnson JA, Ergo A, Coons SJ, Szava-Kovats G. Valuation of the EuroQOL (EQ-5D) health states in an adult US sample. *Pharmacoeconomics*. 1998;13(4):421–33.
18. 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.
19. Bansback N, Tsuchiya A, Brazier J, Anis A. Canadian valuation of EQ-5D health states: preliminary value set and considerations for future valuation studies. *PLoS One*. 2012;7(2):e31115.
20. Devlin NJ, Hansen P, Kind P, Williams A. Logical inconsistencies in survey respondents' health state valuations—a methodological challenge for estimating social tariffs. *Health Econ*. 2003;12(7):529–44.
21. Pullenayegum EM, Tarride JE, Xie F, Goeree R, Gerstein HC, O'Reilly D. Analysis of health utility data when some subjects attain the upper bound of 1: are tobit and CLAD models appropriate? *Value Health*. 2010;13(4):487–94.

22. Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods*. 2006;11(1):54–71.
23. Paolino P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Polit Anal*. 2001;9(4):325–46.
24. Cribari-Neto F, Zeileis A. Beta regression in R. *J Stat Softw*. 2010;34(2):1–24.
25. Verkuilen J, Smithson M. Mixed and mixture regression models for continuous bounded responses using the beta distribution. *J Educ Behav Stat*. 2012;37(1):82–113.
26. 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. McMaster University Centre for Health Economics and Policy Analysis Working Paper. Report No.: 98–11. 1998.
27. 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(6):647–56.
28. 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.
29. Liu H, Cella D, Gershon R, et al. Representativeness of the Patient-Reported Outcomes Measurement Information System internet panel. *J Clin Epidemiol*. 2010;63(11):1169–78.
30. Shalizi CR. *Advanced Data Analysis from an Elementary Point of View*. Cambridge, UK: Cambridge University Press; 2019.
31. Myers RH, Montgomery DC, Vining GG, Robinson TJ. *Generalized Linear Models: With Applications in Engineering and the Sciences*. New York: John Wiley & Sons; 2012.
32. Ellsberg D. Classic and current notions of “measurable utility.” *Econ J*. 1954;64(255):528–56.
33. Koebberling V. Strength of preference and cardinal utility. *Econ Theory*. 2006;27(2):375–91.
34. Fagerlin A, Zikmund-Fisher BJ, Ubel PA, Jankovic A, Derry HA, Smith DM. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Med Decis Making*. 2007;27(5):672–80.
35. Tsuchiya A, Ikeda S, Ikegami N, et al. Estimating an EQ-5D population value set: the case of Japan. *Health Econ*. 2002;11(4):341–53.
36. von Neumann J, Morgenstern O. *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press; 1944.