Check for updates

Original Article

MDM
Medical Decision Making

Medical Decision Making
1–11

DOI: 10.1177/0272989X19856617
journals.sagepub.com/home/mdm
SAGE

Exclusion Criteria as Measurements I: Identifying Invalid Responses

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

Background. In a systematic review, Engel et al. found large variation in the exclusion criteria used to remove responses held not to represent genuine preferences in health state valuation studies. We offer an empirical approach to characterizing the similarities and differences among such criteria. **Setting.** Our analyses use data from an online survey that elicited preferences for health states defined by domains from the Patient-Reported Outcomes Measurement Information System (PROMIS®), with a U.S. nationally representative sample ($N = 1164$). **Methods.** We use multidimensional scaling to investigate how 10 commonly used exclusion criteria classify participants and their responses. **Results.** We find that the effects of exclusion criteria do not always match the reasons advanced for applying them. For example, excluding very high and very low values has been justified as removing aberrant responses. However, people who give very high and very low values prove to be systematically different in ways suggesting that such responses may reflect different processes. **Conclusions.** Exclusion criteria intended to remove low-quality responses from health state valuation studies may actually remove deliberate but unusual ones. A companion article examines the effects of the exclusion criteria on societal utility estimates.

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

Date received: August 14, 2018; accepted: May 15, 2019

Utility-based measures of health-related quality of life (HRQL) provide quantitative estimates of preferences for health states. They are used in cost-effectiveness and cost-utility analyses, decision analyses, clinical trials, and population health studies and management.[1] When elicited from representative samples of individuals, these estimates are often treated as representing societal preferences.[2,3] However, although such studies often go to great lengths to secure such samples, they typically discard many responses, based on exclusion criteria intended to exclude poor-quality responses. In this article and a companion one, we examine the properties of the responses excluded by commonly used exclusion criteria and the implications for analyses that depend on them.

Concerns about data quality have a long history in social science, including its medical applications.[4] The growing accessibility of online data collection has raised particular concerns about the implications of interacting indirectly with participants—reducing the risk of inadvertently cuing particular responses, while reducing the opportunity to clarify often unfamiliar tasks.[5] In medical decision-making research, Engel et al.[6] reviewed 76 utility analyses that use a variety of preference elicitation procedures and utility models. They found large variation in the exclusion criteria that investigators used and called for greater understanding of their meaning. We address that call, beginning with a theoretical discussion that builds on previous health preference studies[6–10] and adds perspectives from behavioral decision research.[11–13] Our analysis distinguishes exclusion criteria that reflect properties of the responses (e.g., unusually high values)

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

and properties of the process producing them (e.g., too brief a survey completion time), which various researchers have interpreted as indicating that the responses do not represent participants' preferences. Although we demonstrate our approach with responses to a health utility survey, it could be applied to any research that removes data for quality control purposes (e.g., many discrete-choice experiments).[9,14]

Exclusion criteria in health utility surveys are meant to remove survey responses that do not represent genuine preferences. Each criterion implies a somewhat different mechanism. Researchers sometimes remove responses produced quickly, arguing that participants were not paying attention. They sometimes remove unusually high (or low) responses, concerned that participants might have been confused, misled, or distracted by unintended features of the user interface in online surveys or nonverbal cues during in-person interviews. Researchers sometimes remove participants who are not confident in their responses, taking those self-reports at face value. Drawing on Engel et al.,[6] Table 1 illustrates the space of exclusion criteria and their rationales.[6,15] We call those at the top *process-based* criteria, reflecting how participants behaved (e.g., how confident they were in their responses). We call those at the bottom *preference-based* criteria, reflecting what participants said on preference elicitation tasks (e.g., did their responses violate the utility theory axioms).

As Engel et al. noted,[6] exclusion criteria affect the representativeness of the resulting utility scores by disproportionately removing individuals with particular responses. However, relatively little is known about whom the various criteria exclude or whether they reflect the mechanism imputed to them. For example, participants who complete surveys quickly could be adept rather than thoughtless. Participants who give unusually high responses could have unconventional preferences rather than haphazard ones. Participants who appear confused to an interviewer could just be idiosyncratic. The diversity of criteria, with different rationales, and implemented in different ways, suggests that investigators are also uncertain, or disagree, about which criteria are theoretically permissible and how to operationalize them. Sensitivity analyses sometimes examine the effect of repeating analysis with data sets reflecting different exclusion criteria. Here, we aim to understand the processes that those criteria reflect.

Our approach has 2 components. The first, reported here, uses multidimensional scaling (MDS)[17–19] to compare the responses excluded by 10 commonly (but inconsistently) used exclusion criteria. It then asks whether those patterns are consistent with the rationales typically given for the criteria and when those patterns are similar for criteria with different rationales, so that explicitly applying one means implicitly applying the other. In a companion article (Exclusion Criteria as Measurements II: Effects on Utility Functions; this issue), we ask how applying each criterion affects health utility estimates. Both articles use the data underlying the Patient-Reported Outcomes Measurement Information System (PROMIS) Preference (PROPr) Scoring System, which offers generic societal preference-based measures of HRQL for health states described by PROMIS.[16,20–22]

The Methods section first describes the PROPr data and the exclusion criteria studied. We then introduce MDS, apply it to the PROPr data, report the results, and discuss their implications. Sections C1-C3 in the Supplementary Appendix contain detailed methods and results.

Methods

Online Survey for the PROPr

The PROPr scoring system includes 7 PROMIS domains, chosen to represent health states of greatest concern to the public, patients, and researchers: Cognitive Function – Abilities (*cognition*), Emotional Distress – Depression (*depression*), Fatigue (*fatigue*), Pain – Interference (*pain*), Physical Function (*physical function*), Sleep Disturbance (*sleep*), and Ability to Participate in Social Roles and Activities (*social roles*). Hanmer et al.[23] describe how the domains were chosen.

Each PROMIS domain is represented as a continuous scale. A given amount of functional capacity or symptom

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 List of Common Exclusion Criteria[a]

Exclusion Criterion	Description of Criterion	Rationale for Exclusion	Notes
Low numeracy	Scored too low on a numeracy scale	Low numeracy implies the participant could not understand the elicitation task	
Low understanding	Interviewer rated participant or participant rated themselves too low on a rating of ability to perform the survey	Participant is unable to use the task to communicate their preferences	
Completed survey too quickly	Completed the survey in less than a minimum time	Completing the survey too quickly implies careless responses	
Provided constant utilities	Excluded if the participant assigned the same utility to every health state	Considered an implausible response pattern, such that the responses cannot be communicating true preferences	
Used too little of the utility scale	Excluded if the participant uses too little of the 0-1 utility scale	An extension of providing constant utilities; considered implausible	"Too little" defined by the researcher, e.g., 10% of the scale
Valued too few health states	Participant removed if the participant valued too few health states	Too few responses imply that the responses given are not reliable	"Too few" defined by the researcher, e.g., fewer than 3
Violated dominance	Valued a state describing health that is at least as good on every dimension as some second health state as worse than that second state	Violations of dominance show that the participant did not understand the task and thus their responses are not preference data; some researchers claim that such responses, if true, cannot be used to represent the preferences of the population	The number of violations of dominance leading to the participant being excluded varies widely; one can also decide by how much one rating must be above the other to count as a violation, allowing the participant some error in their utility assignments
Valued a lower anchor (e.g., dead) as equal or better than full health	Valued one of the states assigned as the origin of the utility scale the same as full health	A specific example of violating dominance; makes certain modeling tasks impossible or uninterpretable, depending on the modeling strategy	
Valued dead or the all-worst state as equal or better than all or some health states	Valued dead or the all-worst state as equal or better than full health or some other health state(s) that describe higher functional capacities than dead and the all-worst state	A specific example of the above two criteria; makes certain modeling tasks impossible or uninterpretable, depending on the modeling strategy	
Did not value dead, all-worst state, or full health	Missing data for one of these 3 states	Makes certain modeling tasks impossible or uninterpretable, depending on the modeling strategy	
Valuations too high	Responses are excluded if they fall in the top $x\%$ of the distribution of responses	Responses in the upper tail are seen as "outliers," thus implausibly high	Together with removing responses that are too low, known as "$2x\%$ trimming"
Valuations too low	Responses are excluded if they fall in the bottom $x\%$ of the distribution of responses	Responses in the lower tail are seen as "outliers," thus implausibly high	Together with removing responses that are too high, known as "$2x\%$ trimming"

[a]Common exclusion criteria used in health state valuation studies, including the rationales for their use. It is based on Table 2 in Engel et al.[6] and adds others based on the 2 categories used in that table—"lack of understanding/engagement" and "model requirements"—including criteria from specific studies.[7,16] Unshaded rows indicate preference-based criteria; shaded rows indicate process-based criteria.

burden on a domain is called a *level* of *theta* (a score in item response theory, which underlies PROMIS). A health state is represented by a vector with a theta score for each domain. The PROPr scoring system provides single-attribute scores (utilities) for each domain and a multiattribute score, combining the domains. Scores on each domain were derived by eliciting utilities from members of a representative US sample for levels of theta. (See Table B1 in the Supplementary Appendix for the theta levels and descriptors.) Those levels were described to participants in qualitative terms (see Figure B1 in the Supplementary Appendix). The multiattribute utility score was derived using multiattribute utility theory, to produce a multiplicative summary scoring function.[24] Dewitt et al.[22] and Hanmer and Dewitt[20] provide details.

Preferences were collected online, with an instrument administered by ICF (https://www.icf.com/services/research-and-evaluation) and SurveyNow (http://www.surveynowapp.com/). As compensation, participants could choose among several products, including gift cards and reward program points. The ICF International Institutional Review Board approved the survey (ICF IRB FWA00002349). Responses were anonymized before researchers received them. Because completing valuation tasks for the entire health state space of PROPr would be unduly burdensome, each participant was randomly assigned to a single health domain. They valued 7 or 8 levels of that health state, depending on the domain (see Supplementary Appendix B1 for details). They also valued the health states of "dead" or "all-worst" relative to the other and to "all-best" (known as "full health"). Participants performed warm-up exercises before undertaking the valuations. Once working on the survey proper, they could not alter previously recorded responses. An introductory question elicited their self-reported life expectancy. They were asked to use that value as the time they would spend in any health state that they evaluated.

Participants valued the health states in 2 ways. The first was a warm-up exercise using a visual analogue scale (VAS) ranging from 0 to 100, sometimes called a feeling thermometer, in which 0 is the value of a lower anchor state and 100 is the value of full health. An example appears in Figure B2 in the Supplementary Appendix. This task was used to introduce the health states. The responses were used only in applying 1 exclusion criterion (*violates-VAS*), as explained below.

The second task used a standard gamble (SG)[25] to elicit utilities for the selected health states, a procedure that was pretested in the development of PROPr. The PROPr scoring system uses the SG method because of its grounding in expected utility theory.[24,26] The specific SG task presented a health state and then offered a choice between 1) that state with certainty and 2) a lottery with probability p of full health and probability (1-p) of a lower anchor state. For single-domain valuation tasks, that lower anchor state had the worst level of functioning on that domain and the highest on all others; for multidomain valuation tasks, the lower anchor was the all-worst state or dead (depending on which was worse for each participant, as indicated on earlier questions). Participants received a series of such choices, which varied p until they were indifferent between the lottery 2) and the certain option 1). Following utility theory assumptions, this probability, $\hat{p}$, is the utility of the intermediate state. The maximum for $\hat{p}$ is thus 1, the utility of full health, and the minimum is 0, the utility of the lower anchor state. Figure B3 in the Supplementary Appendix presents an example.

The PROPr online survey was completed by 1164 participants selected to match the US 2010 Census as closely as possible for several demographic characteristics (see Table B2 in the Supplementary Appendix for demographic information on the sample). The final sample was slightly older, more educated, with higher income and a greater proportion of white individuals than the overall US population. Excellent health was reported by 12.5%, very good health by 39.4%, good health by 33.8%, fair health by 12.4%, and poor health by 1.9%.

Exclusion Criteria

We selected 10 exclusion criteria representing the space defined by Table 1. Table 2 presents those 10 criteria. Some remove all responses (e.g., when participants do not pass a numeracy test threshold), whereas others remove only individual responses (e.g., those "trimmed" as too high or too low).

We created this subset by 1) eliminating criteria not applicable in the PROPr data and 2) choosing the most stringent of "nested" criteria. Thus, we did not use "valued too few health states" (Table 1, row 3) because the PROPr survey required participants to value all health states in the survey. "Nested" refers to criteria in which exclusion by one necessarily implies exclusion by another; Table A1 in the Supplementary Appendix shows examples of ways one might implement the criteria in Table 1 with the PROPr data, which include many nested examples. For example, some nested criteria differ in how many violations of dominance they tolerate before removing a participant. We used the most stringent of these criteria, which excludes even one such violation,

Table 2 Core Exclusion Criteria[a]

Exclusion Criteria (*shorthand*)	Requirements for Exclusion
Score on the Subjective Numeracy Scale of less than 2.5 (*numeracy*)	A participant scored less than 2.5 on the 3-item short form of the Subjective Numeracy Scale.[27]
Self-assessed understanding equal to 1 or 2, on a scale of 1 = *not at all* to 5 = *very much* (*understanding*)	A participant rated themselves a "1" or a "2" on the self-assessed understanding question, which occurred after the preference elicitations.
15-min time threshold (*time*)	A participant completed the PROPr survey in under 15 minutes.
Violated dominance on the SG (*violates-SG*)	A participant, using the standard gamble (SG), violated dominance at least once.
Violated dominance on the VAS (*violates-VAS*)	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 (*dead-all-worst*)	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 (*low-range*)	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 (*no-variance*)	A participant valued every state the same, using the SG.
In the top 5% of responses for an SG (*upper-tail*)	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 (*lower-tail*)	A response falls in the bottom 5% of responses for a health state, using the SG.

[a] Core exclusion criteria, implemented with the Patient-Reported Outcomes Measurement Information System (PROMIS) Preference Scoring System (PROPr) data. Not every criterion from Table 1 is represented, because some would exclude no one by virtue of the design of the PROPr survey (e.g., there are no missing 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.

designated *violates-SG*. We did, however, keep one pair of nested criteria: *low-range* and *no-variance*. Obviously, responses with no variance have a low range, meaning that *no-variance* is nested in *low-range*. However, *no-variance* has a unique relationship with *violates-SG*, in that someone excluded by one cannot also be excluded by the other, because giving the same utility for every health state (no variance) does not violate dominance. However, violating dominance means valuing 2 states differently (and in the "wrong" way), hence having variance. We also deviated from some researchers' practice by having separate criteria for excluding the top 5% of responses (*upper-tail*) and the bottom 5% of responses (*lower-tail*), rather than combining them (*10% trimming*).

In our analyses of exclusion criteria, we treat each criterion as labeling each participant either for exclusion or inclusion, regardless of whether the criterion would exclude all or only some responses in an actual utility analysis.

Multidimensional Scaling

MDS provides a holistic picture of the similarities (and differences) between a set of objects,[18] by translating pairwise comparisons among the objects into a graphical representation in which the distance between objects is a proxy for their similarity.[17] Here, the objects are the exclusion criteria, and the similarity metric assesses how well they agree about which participants to exclude. Each exclusion criterion is a binary classifier, either excluding or including each participant. The agreement of 2 binary classifiers can be represented in a confusion matrix like that in Table 3.

There are many possible summary indices for a confusion matrix.[28–30] We chose one that incorporates each cell:

$$\frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}.$$

Known as the phi (ϕ) coefficient or the Matthews' correlation coefficient,[28,29,31,32] it is also the Pearson correlation for 2 binary variables. (The interpretation of phi differs somewhat from its continuous analog; for example, in how the distribution of each variable affects the possible values it can take.)

In the present application, given n exclusion criteria, the basic input to MDS is the n-by-n proximity matrix, denoted $\boldsymbol{p}$, whose (i,j) entry p_{ij} is the phi value for the proximity of exclusion criteria i and j. The diagonal of

Table 3 Confusion Matrix[a]

	Excluded by Criterion 2	**Included by Criterion 2**
Excluded by criterion 1	*a*	*b*
Included by criterion 1	*c*	*d*

[a] A *confusion matrix* (or *2-by-2 contingency table*), which shows the possible outcomes of 2 joint binary variables (e.g., binary classifiers). We can consider each exclusion criteria as a classifier that categorizes a participant with a 1 if at least 1 of the participant's responses would be excluded by the criterion and a 0 otherwise. The table entries give the counts of the combinations of relationships: *a* is the number of participants excluded by both; *b* is the number excluded by criterion 1 that are included (not excluded) by criterion 2; *c* is the number excluded by criterion 2 that are included by criterion 1; and, *d* is the number included by both.

the proximity matrix is 1 (i.e., $p_{ii} = 1$ for all i); the matrix is symmetric (i.e., $p_{ij} = p_{ji}$ for all i and j), by construction.

When plotting the criteria, MDS uses the phi value of each pair of criteria and also considers the extent to which a pair identifies participants for exclusion in the same way as other criteria, by comparing their 2 rows of phi values in the proximity matrix, iterating over every pair. In doing so, MDS offers a visual display that demonstrates graphically the similarities and differences between criteria with respect to whom to exclude.

MDS algorithms produce m-dimensional configurations or plots, in which the distance between objects (here, the exclusion criteria) best approximate the values in the matrix (here, phi values), as measured by a goodness-of-fit (or *stress*) value. We used the ordinal algorithm because it makes the weakest assumptions about the data.[30] For interpretability sake, we focus on $m = 2$. Higher-dimensional configurations necessarily provide a better fit but are more difficult to interpret.[17] One of our sensitivity analyses adds a third dimension.

Structure in the MDS plot is seen in how the plotted objects cluster and align themselves dimensionally. In the present case, when criteria are close, it means that they exclude (and include) similar participants. The dimensions are not fixed, as the plot is equivalent under certain transformations, such as rotations. However, as with other procedures for reducing dimensionality (e.g., principal components analysis), there is no guarantee that any set of axes is interpretable.

We used the **smacof** package for the statistical software R[33] to identify the best-fit (lowest stress) solution for placing the criteria graphically, such that the distances between them represent the phi values and thus the similarity of their exclusion patterns.

MDS sensitivity analysis. We performed the following sensitivity analyses:

1. *Dimensionality*: We compared the 2-dimensional MDS solution with a 3-dimensional one. Adding dimensions necessarily improves the fit but need not produce new interpretable structures. We compared the 2- and 3-dimensional solutions to see if the latter revealed new relationships.
2. *MDS jackknife*: We repeated the analysis after removing each criterion,[34] to see if any had a disproportionate effect on the overall MDS solution.
3. *MDS algorithm*: We compared the 2-dimensional MDS solution using the ordinal algorithm, with solutions produced using ratio, interval, and spline algorithms, in order to assess the sensitivity to assumptions about the scale type of the input data.
4. *Clustering*: We applied k-means clustering,[35] to compare it with the graphical approach of MDS.

Section C2 of the Supplementary Appendix reports the results of these sensitivity analyses, none of which produced materially different patterns.

Results

Table 4 shows how many PROPr participants are excluded by each criterion, which ranges from 7.8% (*numeracy*) to 84.7% (*violates-VAS*). Except for *violates-VAS*, the preference-based criteria are applied to participants' SG valuations.

Table 5 shows the matrix of phi values correlating the 10 exclusion criteria (Table 2). As phi is symmetric in its arguments, the correlation matrix in Table 4 is symmetric. The greatest agreement (0.98) is for (*no-variance*, *low-range*), the nested pair; the greatest disagreement (–0.58) is for (*no-variance*, *violates-SG*), the mutually exclusive pair. This matrix is the input to the MDS algorithm. (See the Supplementary Appendix, section C1, for details.) As mentioned, similarity is forced with the 2 nested criteria, *low-range* and *no-variance*, which artifactually inflates goodness-of-fit for this solution.

Figure 1 presents the 2-dimensional MDS plot of the exclusion criteria, with the distance between them showing how similarly they exclude participants. If exclusion criteria reflect similar mechanisms (e.g., inattention), they should be clustered closely.

Table 4 Number of Participants Flagged by Each Criterion[a]

Exclusion Criterion	**Number Excluded (Total Sample, N = 1164)**
Score on the Subjective Numeracy Scale of less than 2.5 (*numeracy*)	91 (7.8%)
Self-assessed understanding equal to 1 or 2, on a scale of 1 = *not at all* to 5 = *very much* (*understanding*)	165 (14.3%)
15-min time threshold (*time*)	181 (15.6%)
Violated dominance on the SG (*violates-SG*)	833 (71.6%)
Violated dominance on the VAS (*violates-VAS*)	986 (84.7%)
Valued the all-worst state or dead as the same or better than full health (*dead-all-worst*)	326 (28.0%)
Provided the same response to every SG (*no-variance*)	137 (11.8%)
Used less than 10% of the utility scale (*low-range*)	142 (12.2%)
In the top 5% of responses for an SG (*upper-tail*)	914 (78.5%)
In the bottom 5% of responses for an SG (*lower-tail*)	513 (44.1%)

VAS, visual analogue scale; SG, standard gamble.

[a]The number of participants in the PROPr data flagged by the each criterion from Table 2. The total number of participants in the sample is 1164. Unshaded rows indicate preference-based criteria, shaded rows indicate process-based criteria.

Table 5 Proximity Matrix[a]

	Understanding	*Time*	*Numeracy*	*No-variance*	*Low-range*	*Lower-tail*	*Upper-tail*	*Violates-SG*	*Dead-all-worst*	*Violates-VAS*
Understanding	1	0.04	0.06	0.10	0.1	0.03	0.11	0.02	0.06	0.02
Time	0.04	1	0.02	0.09	0.08	–0.04	0.06	0.03	0.13	0.02
Numeracy	0.06	0.02	1	0.05	0.06	–0.01	0.10	–0.04	0.06	0.03
No-variance	0.10	0.09	0.05	1	0.98	–0.27	0.07	–0.58	0.25	–0.05
Low-range	0.10	0.08	0.06	0.98	1	–0.27	0.05	–0.56	0.24	–0.05
Lower-tail	0.03	–0.04	–0.01	–0.27	–0.27	1	0.08	0.15	–0.10	0.06
Upper-tail	0.11	0.06	0.10	0.07	0.05	0.08	1	–0.04	0.30	0.07
Violates-SG	0.02	0.03	–0.04	–0.58	–0.56	0.15	–0.04	1	–0.08	0.09
Dead-all-worst	0.06	0.13	0.06	0.25	0.24	–0.10	0.30	–0.08	1	0.03
Violates-VAS	0.02	0.02	0.03	–0.05	–0.05	0.06	0.07	0.09	0.03	1

VAS, visual analogue scale; SG, standard gamble.

[a]The proximity matrix for the core criteria: each entry is the ϕ (phi) value of the exclusion criteria in the row and column. The shaded row and column titles indicate the process-based criteria, while the unshaded rows indicate the preference-based criteria.

Discussion

Process-Based Criteria

The 3 process-based criteria have related rationales: 1) *numeracy*: low-numeracy participants may struggle to understand the demanding SG task; 2) *time*: participants who complete the survey quickly may not have made the effort needed to understand it; and 3) *understanding*: participants who report not understanding the task may have been confused—they may also be unusually self-critical. The relative proximity of these criteria in Figure 1 suggests that they reflect a common underlying process: inability or unwillingness to perform the SG task.

Preference-Based Criteria

The *no-variance* and *low-range* criteria are necessarily atop one another (at the center bottom of the figure), as they are nested, with identical responses (*no-variance*) being a special case of highly similar responses (*low-range*). Conversely, the most distant criteria in the figure are *violates-SG* and *no-variance*, which necessarily exclude different participants (as it is impossible to violate dominance when assigning the same value to all health states). The other relationships in the figure are empirical, rather than necessary ones.

Both *violates-SG* and *violates-VAS* are applied when participants rate a state with lower functional capacity or more

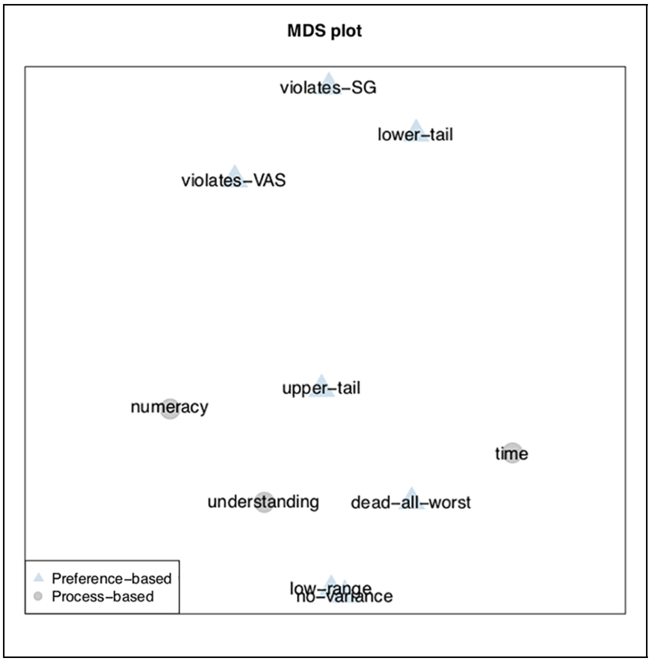

Figure 1 The core 2-dimensional multidimensional scaling configuration.

symptom burden as better than a state with higher functional capacity or less symptom burden, thereby violating dominance. Both exclude many participants (Table 4). The fact that VAS excluded more may reflect its greater precision: the VAS allows increments of 0.01, whereas the SG (as implemented here) only offers probability (hence utility) increments of 0.05. The 2 criteria lie close to one another in the MDS space, indicating that they exclude similar individuals.

Both, however, are far enough away from the process-based criteria (*numeracy*, *understanding*, *time*) that they appear to capture different mechanisms than those 3 measures of low-quality responses. Indeed, the fact that *violates-SG* is so distant suggests that it might remove some participants who are trying to express well-considered utilities but cannot do so without violating dominance. For example, some participants may prefer some depression to none, believing that it confers empathy that is valuable to their health. Similarly, some participants might prefer the worst possible cognitive ability to poor cognitive ability, when the former includes a lack of self-awareness, but the latter does not. Such preferences violate dominance deliberately. Other violations may include participants who have something to say but struggle with either the wording or the mechanics of the task. Researchers who use *violates-SG* as an exclusion criterion typically allow some violations, up to 11 in one study,[6] suggesting that such struggles might be common. Box 1 describes one possible confusion pattern with the interface used here.

The *dead-all-worst* exclusion criterion represents a violation of dominance that arises when participants rate dead or the all-worst state as being at least as good as full health. Its proximity to *time* and *understanding* might indicate the kind of confusion that Box 1 suggests for *violates-SG*. However, the distance between *dead-all-worst* and *violates-SG* indicates different processes. Possibly, participants excluded by *violate-SG* might be trying to express themselves but are frustrated by the interface design, whereas participants excluded by *dead-all-worst* either cannot communicate their preferences or are not trying, leading them to say that they prefer dead or the all-worst state to full health. Doing so is an extreme violation of dominance, whereas *violates-SG* flags any violation, even among relatively similar health states.

The final 2 criteria, *upper-tail* and *lower-tail*, remove the highest and lowest 5% of responses. (In our implementation, they flag anyone with a response eligible for such trimming; often, there are many responses tied at the cut-off point.) Ten percent trimming is a common practice, followed in constructing both the PROPr scoring system and the widely used Health Utilities Index Mark 3.[16] The 2 criteria are far apart in Figure 1, indicating that they exclude different participants. Of the 513 flagged by *lower-tail*, 421 (82.1%) are flagged by *upper-tail*; conversely, *lower-tail* flags 57.0% of the 914 flagged by *upper-tail*. Phi reflects this asymmetry, with a value of 0.08 (Table 5) for the 2 criteria.

Both criteria apply to SG valuations for individual health states and make exclusions based on how a participant's response compares with those of other participants. However, as seen in Figure 1, the exclusion pattern for *lower-tail* is most similar to that for *violates-SG*, which compares a participant's responses across health states, without consideration for how other participants respond. In contrast, the exclusion pattern for *upper-tail* is closest to those for *numeracy* and *understanding*, criteria that seem to exclude people who cannot or will not perform the task, which is more in line with the stated rationale of both trimming criteria. Thus, 10% trimming may remove 2 very different groups of participants: those who understand the task and explicitly express unusually low utilities and those who are confused by the task and inadvertently produce unusually high utilities.

Conclusion

Removing responses from data sets is a common practice in health utility studies[6] and other empirical research.[36]

Box 1

One possible account for high valuations is that the first question in any SG valuation presents a degenerate gamble where 1 option has a 100% probability of full health and 0% probability of the low anchor state (e.g., dead or the all-worst state), and the other option is the sure-thing of the intermediate state whose utility is being estimated. The participant can choose the gamble, the sure thing, or indifference. Choosing the gamble leads to a choice between a different gamble and the sure thing. Choosing the sure thing or expressing indifference completes the task and implies a utility estimate of 1 (or greater than 1 for those who select the sure thing). Therefore, making either of these last two choices leads to a response in the upper tail of the utility distribution. Thus, the mechanics of the procedure might lead to confused or inattentive responses being recorded as utilities in the upper tail of the response distribution, given that two of the three initial choices lead to an extreme utility value. Less numerate participants might be more likely to experience such confusion.

Exclusion criteria formalize the removal process. Here, we analyze 10 exclusion criteria, chosen to represent those commonly used in health utility studies. We use MDS to compare their removal patterns, using the PROPr data set as a testbed.[22] Those criteria include preference-based ones, reflecting what participants said (e.g., with unusually low utilities), and process-based ones, reflecting how they responded (e.g., unusually quickly).

The clustering of the 3 process-based criteria in Figure 1 suggests that they reflect related aspects of poor performance. In contrast, the spatial distribution of the preference-based criteria suggests that they reflect different mechanisms. As interpreted above, those mechanisms are sometimes at odds with the rationales given for using the criteria (Table 1). For example, *upper-tail* and *lower-tail* are far apart, despite commonly being combined using the same rationale (i.e., "aberrant" responses). Our analysis suggests that upper-tail responses reflect confusion or inattention, making them, in effect, preference-based reflections of response processes. In contrast, lower-tail responses appear to be deliberate expressions of low utilities. Thus, we propose that the 2 trimming criteria not be combined when the standard gamble is implemented in the same way as in PROPr (see Box 1). As a result, we analyze them separately in the companion article (Exclusion Criteria as Measurements II: Effects on Utility Functions; this issue), which assesses the effects of applying these criteria on health state utility estimates.

These results and interpretations suggest ways in which future elicitation procedures might be improved, rendering fewer responses and participants as candidates for exclusion. Some investigators prefer in-person SG elicitations, such as the paper standard gamble,[37] in order to reduce cognitive demands and any confusion caused by an unfamiliar computer interface. If online survey methods[38,39] are to achieve the potential benefits of low-cost data collection from demographically diverse samples, they need to reduce the cognitive demands of the SG. One possible strategy is the use of interface designs that provide real-time feedback to help users increase their understanding of the task without biasing their content. Those designs might include references to exclusion criteria, attention checks,[40–43] or manipulation checks,[44] asking whether participants understand the task, thereby communicating to the participant what the researcher expects from them.[45] PROPr used in-person pilot testing to refine its survey design, as well as having participants complete the VAS to familiarize them with the health domain and consider their preferences for health, before completing the SG task. The approach demonstrated here and in the companion article (Exclusion Criteria as Measurements II: Effects on Utility Functions; this issue), asking how preferences differ between the excluded and included responses, could be used to assess the impacts of competing designs. Those tests could be applied to other preference elicitation tasks as well, such as the time-tradeoff and discrete-choice experiments.[46,47]

None of the exclusion criteria considered here explicitly examine whether responses are informed,[48] in the sense that preferences are based on considered reflection, possibly including personal experience of health states (e.g., through illness or caregiving). Choosing only informed preferences could be defined as an exclusion criterion, if there are measurements to operationalize it. Doing so requires analysts to consider the debate over whether HRQL measurement should reflect the preferences of the general population or the people most directly involved.[49]

Exclusion criteria pose a tradeoff between potentially improving data quality and potentially reducing sample representativeness. The present analyses provide insight into which responses are removed by different criteria and why. The companion paper analyzes their effects on estimates of health state utilities, complementing sensitivity analyses that reanalyze data with and without data exclusions. Its concluding section offers overall recommendations, drawing on the present results and those reported there. Both articles assume that exclusion criteria should be selected in advance, based on their rationale, and applied only if the data are consistent with that rationale. They offer complementary approaches to determining whether that is the case.

ORCID iDs

Barry Dewitt https://orcid.org/0000-0003-1622-6736
Janel Hanmer https://orcid.org/0000-0001-6159-2482

Supplementary 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. Neumann PJ, Goldie SJ, Weinstein MC. Preference-based measures in economic evaluation in health care. *Annu Rev Public Heal*. 2000;21:587–611.
3. 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.
4. Broeck J Van Den, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*. 2005;2(10):e267.
5. Boyd D, Crawford K. Critical question for big data: provocations for a cultural, technological, and scholarly phenomenon. *Informat Commun Soc*. 2012;15(5):662–79.
6. 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.
7. 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.
8. Devlin NJ, Hansen P, Selai C. Understanding health state valuations: a qualitative analysis of respondents' comments. *Qual Life Res*. 2004;13(7):1265–77.
9. Lancsar E, Louviere J. Deleting "irrational" responses from discrete choice experiments: a case of investigating or imposing preferences? *Health Econ*. 2006;15(8):797–811.
10. 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.
11. Fischhoff B, Kadvany J. *Risk: A Very Short Introduction*. New York: Oxford University Press; 2011.
12. Fischhoff B. Judgment and decision making. *Wiley Interdiscip Rev Cogn Sci*. 2010;1(5):724–35.
13. Edwards W. The theory of decision making. *Psychol Bull*. 1954;51(4):380–417.
14. Law EH, Pickard AL, Kaczynski A, Pickard AS. Choice blindness and health-state choices among adolescents and adults. *Med Decis Making*. 2017;37(6):680–7.
15. 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.
16. 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.
17. Borg I, Groenen PJF. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer; 2005.
18. Baird JC, Noma EJ. *Fundamentals of Scaling and Psychophysics*. New York: John Wiley & Sons; 1978.
19. Shepard RN. Multidimensional scaling, tree-fitting, and clustering. *Science*. 1980;210:390–8.
20. Hanmer J, Dewitt B. PROMIS-Preference (PROPr) score construction: a technical report. 2017. Available from: janelhanmer.pitt.edu/PROPr.html
21. Hanmer J, Feeny D, Fischhoff B, et al. The PROMIS of QALYs. *Health Qual Life Outcomes*. 2015;13:122.
22. 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;38:683–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29944456%0Ahttp://journals.sagepub.com/doi/10.1177/0272989X18776637
23. 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;26:3377–85.
24. Keeney RL, Raiffa H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley & Sons; 2003.
25. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Med Decis Making*. 2001;21:329–34.
26. von Neumann J, Morgenstern O. Theory of Games and Economic Behaviour. Princeton, NJ: Princeton University Press; 1944.
27. 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;35:932–6.
28. Gower JC, Legendre P. Metric and Euclidean properties of dissimilarity coefficients. *J Classif*. 1986;3:5–48.
29. Warrens MJ. On association coefficients for 2 x 2 tables and properties that do not depend on the marginal distributions. *Psychometrika*. 2008;73(4):777–89.
30. Borg I, Groenen PJF, Mair P. *Applied Multidimensional Scaling*. New York: Springer; 2012.
31. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem Biophys Acta*. 1975;405(2):442–51.

32. Yule GU. On the methods of measuring the association between two attributes. *J R Stat Soc*. 1912;75(6):579–652.
33. de Leeuw J, Mair P. Multidimensional scaling using majorization: SMACOF in R. *J Stat Softw*. 2009;31(3):1–30.
34. de Leeuw J, Meulman J. A special jackknife for multidimensional scaling. *J Classif*. 1986;3(1):97–112.
35. Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer; 2006.
36. Martinson BC, Anderson MS, de Vries R. Scientists behaving badly. *Nature*. 2005;435(7043):737–8.
37. Ross PL, Littenberg B, Fearn P, Scardino PT, Karakiewicz PI, Kattan MW. Paper standard gamble: a paper-based measure of standard gamble utility for current health. *Int J Technol Assess Health Care*. 2003;19(1):135–47.
38. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect Psychol Sci*. 2011;6(1):3–5.
39. Leeuw ED De. Counting and measuring online: the quality of internet surveys. *Bull Méthodologie Sociol*. 2012;68–78.
40. Hauser DJ, Schwarz N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res Methods*. 2016;48(1): 400–7.
41. Hauser DJ, Schwarz N. It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*. 2015;5(2). Available from: http://sgo.sagepub.com/lookup/doi/10.1177/2158244015584617
42. Abbey JD, Meloy MG. Attention by design: using attention checks to detect inattentive respondents and improve data quality. *J Oper Manag*. 2017;53–56:63–70.
43. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav Res Methods*. 2014;46(4):1023–31.
44. Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: detecting satisficing to increase statistical power. *J Exp Soc Psychol*. 2009;45(4):867–72.
45. Armantier O, Bruine de Bruin W, Potter S, Topa G, van der Klaauw W, Zafar B. Measuring inflation expectations. *Annu Rev Econ*. 2013;5:273–301.
46. Weinstein MC, Torrance G, Mcguire A. QALYs: the basics. *Value Health*. 2009;12:S5–9.
47. Ratcliffe J, Brazier J, Tsuchiya AKI, Symonds T, Brown M. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. *Health Econ*. 2009;18:1261–76.
48. Karimi M, Brazier J, Paisley S. Are preferences over health states informed? *Health Qual Life Outcomes*. 2017;15(1):1–11.
49. 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.