Measurement of Health-Related Quality of Life

Preference Aggregation, Exclusion, and Public Policy

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Engineering & Public Policy

Barry D. Dewitt

B. Arts & Sci, Mathematics, McMaster University
M. Sc., Mathematics and the Foundations of Computer Science, University of Oxford
M. S., Engineering & Public Policy, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA
December, 2017

Acknowledgements

The College of Engineering's requirements happily coincide with my own wishes – to, first, thank my committee: Baruch Fischhoff (chair), Alex Davis, Janel Hanmer, and Mark S. Roberts.

This section is insufficient to thank Baruch for our first meeting, from 4-4:30pm on Monday, February 11th, 2013, and for all subsequent and future meetings. I owe an uncountably(!) infinite amount of gratitude to him – and subsequently, to each of Alex and Janel – for taking a chance on an ex-math-grad-student, who in 2013 knew far more about the *topos* approach to quantum theory than anything in the realm of EPP or (medical) decision making. I now know (i.e., remember) less about *topoi*, and all I know and continue to learn about the latter group of subjects – and much else besides – is due to them. My thanks to Mark for agreeing to be the external member of my dissertation committee and for making "external" mean "supportive at every turn."

Next, I must thank the generous financial support of the Department of Engineering & Public Policy and the College of Engineering of Carnegie Mellon University, the Social Sciences and Humanities Research Council of Canada, the Swedish Foundation for the Humanities and the Social Sciences (Riksbankens Jubileumsfond) Program on Science and Proven Experience (Vetenskap och beprövad erfarenhet), and the Federal Alliance for Safe Homes. Data collection was supported by the National Institutes of Health (Grant UL1TR000005). Thanks also to SAGE Publications for permission to reprint the content of Part I of the dissertation, which is published in *Medical Decision Making*.

Special thanks to Teddy Seidenfeld, whose graduate seminar inspired much of the content of Part I; to Janel, for inviting me to join PROPr, and to Janel along with the whole PROPr advisory team for trusting me with Part II; and, to Stephen Broomell, for many conversations about Part III (and many other conversations these past years).

Not a requirement of the College, though I am sure the College supports it: my thanks to my family and friends for the constant support and encouragement. Special thanks, in randomly generated order, to my siblings Aaron and Beth, partner Sarah, parents David and Susie, and friends Mike Yu and Paul Tisa, for putting up with me more than most.

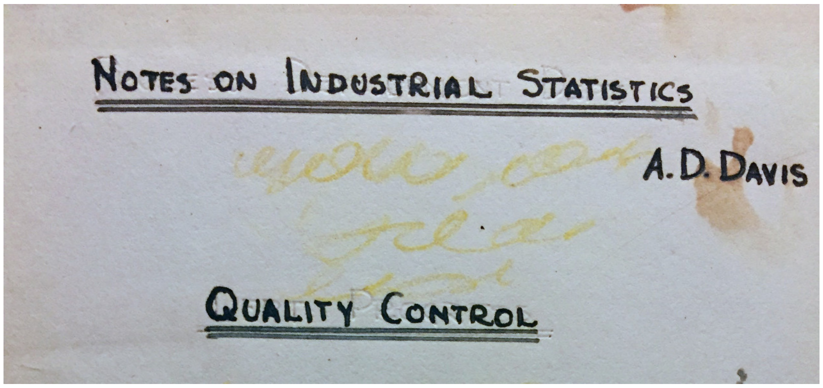
Finally, my thanks to the faculty, staff, and students of EPP, to everyone whom I have learned from while at CMU, Oxford (especially Andreas Döring), UW (especially Bob Dumas), McMaster (especially Megumi Harada, Matt Valeriote, and John Browning), and to the many before my post-secondary education began, who nurtured and encouraged me.

Abstract

Societal preference-based health-related quality of life (HRQL) measures provide the numbers that researchers and policy-makers need for quantifying the value of health. This dissertation provides a normative, descriptive, and prescriptive analysis of the design of such measures. Part I analyzes the normative foundations of the preference aggregation procedures that define societal preferences, concluding that conventional procedures represent a small subset of those that would be normatively permissible given conventional assumptions. Each aggregation procedure represents different ethical principles, reflecting differences in the treatment of preference heterogeneity, and Part I presents an analytical-deliberative framework for choosing among them. Part II describes the creation of a new HRQL measure, the Patient-Reported Outcomes Measurement Information System (PROMIS®) Preference (PROPr) Scoring System. The PROPr Scoring System is a free, open-source tool that combines best practices in health profile measurement with those for creating preference-based scores. It is designed for integration with the PROMIS initiative of the National Institutes of Health, which is poised to become the U.S. standard for patient-reported outcomes. Part III characterizes the empirical relationships between exclusion criteria – meant to capture the data from preference surveys that are not true preferences – and the implications for those who choose to implement them. We show that conventions for data cleaning via exclusion criteria represent disparate mechanisms for exclusion, that exclusions impact the preference data available for analyses, and that exclusion likely affects the societal representativeness of the included preferences. We argue that there is sufficient empirical evidence to recommend only a subset of criteria, and describe procedures for implementing them in order to minimize wrongful exclusion. We outline future work to produce criteria with better classification properties and surveys that minimize the need for exclusion.

Dedication

For my grandfathers, the late Abraham Derek Davis (z"l) and the late Vernon Dewitt (z"l). Vernon's education was cut short by the Second World War, during which he served in an engineering corps of Her Majesty's Armed Forces. Abraham graduated with a Bachelor's of Science in Civil Engineering from the University of New Brunswick, in 1934, but ultimately joined his wife's family business. I grew up safe in Toronto, and followed a circuitous path to an engineering department. I hope my PhD from the Carnegie Institute of Technology would have made them proud. (Despite the fact that, by the end of it, I cannot build roads like they could, nor even print as nicely as Abraham – unless assisted by a computer. See Figure 1.)



NOTES ON INDUSTRIAL STATISTICS

A.D. DAVIS

QUALITY CONTROL

Figure 1: Notebook of Abraham Davis, circa 1930.

Contents

List of Tables

List of Figures

1

Preface

This dissertation focuses on *health-related quality of life*. Measures of health-related quality of life – abbreviated “HRQL” or “HRQoL” in the literature – define HRQL as a multi-dimensional construct, where each dimension is a *health domain*, such as cognition, physical functioning, sleep quality, etc. HRQL measures rely on the work of psychometricians to produce numeric scales for each of those domains, allowing health to be represented as a vector of numbers where each number is a measurement on one health domain. That vector defines a *health state*, *health profile*, or *health status*.

Other measures – variously called *preference-based*, *societal preference-based*, or *utility-based* measures of HRQL – go one step further: they assign a *utility* to each vector, meant to represent society’s preferences for the health described by the vector. That process allows comparing every vector with every other vector, and it is the utilities (also called *preference scores*) that are used in policy analyses, such as cost-effectiveness analyses.

This dissertation analyzes the normative foundations of societal preference-based HRQL (Part I), describes the construction of a new HRQL measure (Part II), and argues that conventions for excluding data from preference studies can be improved (Part III). More specifically:

- In Part I, we use social choice theory to show that conventional *preference aggregation procedures*, used to turn lay preference data into societal utility values, are normatively justified; however, the analyst or researcher has many procedures to choose from, each with ethical and practical implications. We suggest an analytical-deliberative framework for choosing among those procedures, informed by behavioral decision research.

- In Part II, we describe the creation of a new utility-based HRQL measurement system called the Patient-Reported Outcomes Measurement Information System (PROMIS®) Scoring System (the PROPr Scoring System). It combines best practices for the measurement of health domains with those for creating utility scores out of lay preference data, in a free, open-source tool.
- In Part III, we analyze the conventional *exclusion criteria* used in health state valuations studies. These studies collect preference data from the lay public for the creation of societal utilities; the exclusion criteria define the data considered "high-quality" and the data that will be excluded from subsequent work. We show that current conventions for data cleaning via exclusion criteria represent disparate potential mechanisms for exclusion, that those mechanisms have varying and often uncertain ability to capture non-preference data, and we characterize the policy implications for those who wish to implement them.

Although the focus is on preferences for health, the dissertation is relevant to any work synthesizing the preferences of the lay public, and to surveys where data quality is a concern – such as many online surveys. For example, much of the dissertation's content could be applied to the design and interpretation of discrete choice experiments eliciting lay preferences for policy options.

At the time of writing, Part I is published in *Medical Decision Making* (Dewitt, Davis, Fischhoff, & Hanmer, 2017). Part II will soon be submitted, and the reader can refer to the PROPr website (`http://janelhanmer.pitt.edu/PROPr.html`) for more detail on PROPr and references to papers completed under the PROPr project. Part III is in preparation for submission, as multiple manuscripts.

Barry Dewitt
Pittsburgh, September 2017

Part I

Preference Aggregation in Health-Related Quality of Life

2

An Approach to Reconciling Competing Ethical Principles in Aggregating Heterogeneous Health Preferences

2.1 Introduction

Generic health-related quality of life (HRQL) measures place individuals' health status on a common scale, allowing researchers to compare the effects of clinical trials across individuals and summarize the results of population health studies (Wilson & Cleary, 1995). Utility-based measures are a subset of HRQL measures, and attach scores to states of health. These scores can be used for outcomes, e.g., quality-adjusted life years, providing the estimates needed by regulatory analysts (Gold, Siegel, Russell, & Weinstein, 1996; Neumann, Saunders, Russell, Siegel, & Ganiats, 2016).[1]

To produce societal preference-based HRQL scores, the common practice is to aggregate the preferences of a sample of individuals designated as representing the target population. That might mean eliciting the preferences of patients with a disease, experts in a disease, or individuals held to represent society as a whole. For example, a sample of individuals with naturally varying vision might be asked to assess, in numeric terms, the relative quality of life with blindness and

[1] This chapter is published in *Medical Decision Making* and can be accessed at `http://journals.sagepub.com/doi/abs/10.1177/0272989X17696999`. The *MDM* version should be considered the version of record (Dewitt et al., 2017).

20/20 vision.

Although analysts need such aggregation procedures so that they can incorporate HRQL estimates into their models, aggregating individuals' estimates to societal ones poses a fundamental problem in social choice theory (Keeney & Raiffa, 2003; Arrow, 1951; Hirose, 2015; Rawls, 1971; Sen, 1970). It has long been known that, under certain general assumptions, there is no unique solution to this aggregation problem if there is any heterogeneity in individuals' preferences. Rather, many solutions are possible. Selecting a method to collapse a distribution of preferences into a single number implies an ethical judgment about what distributional information matters.

Foundational work in utility-based HRQL measurement (e.g., (Torrance, Boyle, & Horwood, 1982)) recognized the preference aggregation problem. Concurrent research in social choice theory revealed the implications of alternative aggregation procedures (e.g., (Roberts, 1980)). Here, we integrate the two fields, taking advantage of advancements in both during the ensuing years. We use the Health Utilities Index (HUI) as an exemplar (Torrance et al., 1996; Furlong et al., 1998; Feeny et al., 2002), examining it in terms of key concepts in social choice theory (Arrow, 1951; Sen, 1970; Roberts, 1980), thereby providing a concrete example that could be followed with other societal preference-based scores.[2]

We first identify the conditions under which HUI is normatively justified, and then show the range of acceptable aggregation procedures, each of which expresses an ethical stance. Finally, we offer an approach for choosing among these options. Related concerns can be found in assessments of other forms of analysis (Asaria, Griffin, & Cookson, 2015; Dolan, Shaw, Tsuchiya, & Williams, 2005; Fischhoff, 2015; Nord, Pinto, Richardson, Menzel, & Ubel, 1999). Our approach is generalizable to any societal preference-based HRQL measurement system.

2.2 Methods

Below, we describe relevant results from social choice theory, applicable to specifying any societal preference-based HRQL measurement system. Additional background material on societal preference-based HRQL scores and the HUI system is available in the Appendix.

[2] We use "HUI" to refer to both the HUI Mark 2 and Mark 3 systems. When it is necessary to distinguish between them, we do so with the acronyms "HUI:2" and "HUI:3."

2.2.1 Social choice theory

Social choice theory characterizes preference aggregation procedures that define societal preferences. It begins with a set of axioms that a preference aggregation procedure must satisfy to be deemed rational. One commonly used set has these three axioms:[3]

i) *Unrestricted Domain*: Any set of individual preferences is allowed.

ii) *Independence of Irrelevant Alternatives:* If two groups of utility functions agree on a subset of health states, then the societal preferences of the two groups agree on that same subset.

iii) *Weak Pareto Criterion*: If all individuals prefer health state x to health state y, then the societal preference should as well.

A fundamental result from social choice theory is Arrow's Impossibility Theorem (Arrow, 1951), which states that the only guaranteed way to aggregate individual preferences that satisfies these axioms is dictatorship: impose the preferences of one individual on the entire group. It is called an *im*possibility theorem because "non-dictatorship" is an axiom in Arrow's framework, making it impossible to satisfy the full set of axioms.

Arrow's result has been interpreted as precluding *any* non-dictatorial aggregation procedure from being normatively justified. However, as Sen (1970) initially showed, and Roberts (1980) elaborated, Arrow's theorem is a special case of a more general result about preference aggregation. Sen and Roberts identified two aspects of individuals' preferences that determine the type of aggregation that is possible: *informational content* and *interpersonal comparability*. Together, they constitute the *informational basis* of the preferences.

The informational content of an individual utility function reflects its measurement scale; for example, *ordinal* or *interval*. Ordinal preferences provide only enough information to rank options. For example, assigning 15 to option A, 5 to option B, and 0 to option C means that A is preferred to B, B is preferred to C, and A is preferred to C, but nothing more. With cardinal preferences, utility is on an interval scale, so that units of utility have consistent meaning across the scale. Thus, in the example, A is preferred to B by twice as much as B is preferred to C, meaning that there is *intra*personal comparability of utility differences. Ordinal and cardinal preferences, on ordinal and

[3] These axioms are only applied to preference aggregation procedures defining societal preferences that produce a complete, reflexive, and transitive ordering over the state space.

interval scales, respectively, define what can be said about an individual's preferences, but say nothing about interpersonal comparisons among individuals' preferences.

Following Sen and Roberts, we distinguish two types of interpersonal comparability: *level comparability* and *unit comparability*. Level comparability lets us say whether one person's HRQL is better or worse than another's, whereas unit comparability lets us say whether a change in one person's HRQL is greater or less than a change in another's (e.g., one person improves more with a treatment) (Roberts, 2005).

If preferences have cardinal informational content, level comparability, and unit comparability, then they satisfy *cardinal full comparability*. Sen and Roberts showed that Arrow's result applies to the special case of preferences that are ordinal and completely non-comparable, and that non-dictatorial aggregation procedures are possible when preferences are cardinal or allow interpersonal comparability.

We apply the Sen-Roberts framework to determine when aggregation procedures in HRQL measurement can be normatively justified. We use HUI as an example, and then consider general conditions. We begin by examining the assumptions that HUI makes about the informational content and comparability of individual preferences. We then describe the preference aggregation procedures that these assumptions allow. Finally, we describe an empirical framework for choosing among these procedures, which can be applied to current aggregation procedures and might suggest new ones.

2.3 Results

2.3.1 Utility elicitation

The normative justification of any aggregation procedure, including those used in HUI, depends on its ability to meet social choice theory's demands regarding the informational basis of the preferences and the appropriateness of its axioms. (HUI uses two aggregation procedures: 1) produce a mean multi-attribute utility function by averaging individuals' multi-attribute utility functions, or 2) the *person-mean* approach, where single-attribute utility functions are averaged over individuals and then combined into a multi-attribute utility function for the "person-mean," a hypothetical individual whose preferences equal the mean of individual preferences within each attribute. HUI:2 produces functions with both 1) and 2), and HUI:3 uses 2). See the Appendix for more detail.)

In the case of HUI, one must establish the informational content (ordinal or cardinal) and interpersonal comparability (level and unit) of preferences elicited on a scale from 0 to 1, where 0 represents the utility of some lower anchor state (e.g., dead, the most-disabled state) and 1 represents the utility of the full health state (i.e., the most-able state). Participants then assign numbers between 0 and 1 to a set of intermediate health states using methods that rely on the standard gamble technique, an approach used to produce cardinal preferences (Torrance, 1986). Furthermore, HUI assumes full comparability: both utility values and changes in utility values are treated as comparable across people. In the development of HUI, individuals' responses are transformed during the creation of the systems' respective societal utility functions. For example, in the HUI:2 system, if participants assigned 0 to dead, their responses are transformed using a strictly positive affine transformation, so that 0 represents the utility of the most-disabled state, and 1 the utility of full health (Torrance et al., 1996). All of the transformations used throughout HUI produce utilities that are assumed to have origin and scale comparability; i.e., cardinal full comparability.

Determining the applicability of the social choice axioms requires evaluating utility elicitation and aggregation procedures in their light. The first axiom, Unrestricted Domain, deals with both utility elicitation and aggregation, while the other axioms deal primarily with aggregation. We consider each axiom in turn.

The axiom of Unrestricted Domain requires that the preference aggregation procedure can be followed with any set of individuals. A classic violation of this axiom is majority rule in Arrow's context, where preferences are ordinal and non-comparable. Majority rule is a permissible aggregation procedure only if one is allowed to remove some number of individuals. Thus, majority rule does not conform to this axiom, because it necessarily excludes individuals in order to produce societal preferences. In contrast, HUI adheres to the axiom of Unrestricted Domain because it can accept any individual's preferences as inputs; although, in practice, HUI disallows incoherent responses, such as valuing a health state higher than one that dominates it (i.e., a health state that is as good or better on all attributes (Torrance et al., 1996)).

Independence of Irrelevant Alternatives requires the societal ranking of any subset of states to be solely a function of individuals' utilities for each of those states. Therefore, preferences for other health states should be irrelevant, as should an individual's current health state. HUI combines preferences according to this principle, defining the societal preference of a given state as a function of individuals' preferences for that state alone.

The Weak Pareto Criterion ensures that unanimous preferences – should they exist – prevail, precluding any other concerns. HUI methods adhere to this axiom, meaning that unanimity about a ranking would result in the societal preference preserving that ranking.

Sen (1970) discusses some contexts where aggregation procedures that incorporate other types of concerns might be desirable, and Roberts (1980) shows how it is possible, under cardinal full comparability, to derive their mathematical representation. More generally, social choice theory examines aggregation procedures that satisfy alternative axiomatic conditions, often involving the weakening of one of the above set (Roberts, 1980; D'Aspremont & Gevers, 2002). Some of these alternative axiomatizations might also be satisfied by societal preference-based HRQL measurement systems.

By examining how HUI elicits individuals' preferences and combines them to produce its aggregate score, we find that HUI assumes cardinal full comparability, and satisfies the three axioms of Unrestricted Domain, Independence of Irrelevant Alternatives, and the Weak Pareto Criterion. The informational basis and the axioms then determine the aggregation procedures that are normatively permissible. An analogous examination of the elicitation and aggregation procedures of any other societal preference-based score should be sufficient to determine its informational basis, and thus its normatively justifiable aggregation procedures. In the next section, we define those procedures for the HUI system.

2.3.2 Normatively justifiable aggregation procedures

The aggregation procedure used in HUI is based on the mean. It can be written as:

$$U = U_{avg} = \frac{1}{n} \sum_{i} u_i. \tag{2.1}$$

Thus, the societal utility function (U) is the average of the individual utility functions (u_i). Applying the Sen-Roberts test means asking whether this function U_{avg} respects the cardinality and interpersonal comparability conditions of the individual utility functions (i.e., cardinal full comparability).

Roberts (1980) describes the set of such normatively justifiable societal utility functions, under many different informational bases. With cardinal full comparability, averaging individuals' utilities is one such function. Thus, the HUI averaging strategy (equation (2.1)) is consistent with social choice theory. However, there are, as Roberts notes, an infinite number of other normatively justified aggregation procedures under cardinal full comparability, each of which adjusts the average (equation (2.1)) by another function. More precisely, any function of the form

$$U(x) = U_{avg}(x) + g\left(u(x, \cdot) - U_{avg}(x)\right), \tag{2.2}$$

where x is a health state and g is a homogeneous function of degree 1 – meaning that $g(\lambda v) = \lambda g(v)$ for any v in the domain of g and all $\lambda > 0$ – is allowed. Following Roberts' notation, $u(x, \cdot)$ denotes the function that maps an individual to their utility for health state x.

For example, one could define the societal utility of health state x (i.e., $U(x)$) as a weighted average of $U_{avg}(x)$ and the minimum utility among the individual utility functions evaluated at that health state. In symbols, this would be $U(x) = \alpha U_{avg}(x) + (1 - \alpha) \min_i u_i(x)$, where $\alpha \in [0, 1]$ and $u_i(x)$ is the utility function of individual i evaluated at state x. In contrast to U_{avg}, this alternative U penalizes health states that leave one person with a low utility, even if everyone else is well-off. We explore the implications of such differences in the next section. Table 2.1 lists other normatively justifiable alternative forms for U.

HUI's cardinal fully comparability and adherence to the set of social choice axioms, demonstrated in the previous section, determines its normatively acceptable aggregation procedures. The procedure that the HUI system chose – averaging, U_{avg} – is normatively permissible. However, so are an infinite number of alternatives. That raises the question of how to choose among these possibilities. We frame the answers in terms of what is lost by relying on each.

Table 2.1: *Normatively justifiable aggregation procedures.* A variety of aggregation procedures (and their associated societal utility functions) when cardinal full comparability applies, i.e., when preferences are cardinal and when utility levels and differences in utility are comparable across people. The general form of any societal utility function in this context is $U(x) = U_{avg}(x) + g(u(x,\cdot) - U_{avg}(x))$, where $U_{avg}(x)$ is the average utility at health state x, g is a homogeneous function of degree 1, and $u(x,\cdot)$ denotes the function that maps an individual to their utility for health state x. See Roberts (1980, p. 431) for more detail on the first four of these functions, as well as for other examples.

Aggregation Procedure	*Societal Utility Function*	*Description*	*Features*
Mean	$U(x) = U_{avg}(x)$	Societal utility is average utility.	Insensitive to preference heterogeneity.
Minimum-adjusted	$U(x) = \alpha U_{avg}(x) + (1-\alpha)\min_i u(x,i)$, $\alpha \in [0,1]$	Adjust average utility to account for the minimum utility. The parameter α weights the contribution of the average and the minimum in defining the societal utility.	Health states with lower minimum utilities relative to their means – leaving some people much worse off than average – have lower societal utility. When $\alpha = 0$, it is concerned only with the worst off.
Maximum-adjusted	$U(x) = \alpha U_{avg}(x) + (1-\alpha)\max_i u(x,i)$, $\alpha \in [0,1]$	Adjust average utility to account for the maximum utility. The parameter α weights the contribution of the average and the maximum in defining the societal utility.	Health states with higher maximum utilities relative to their means – leaving some people much better off than average – have higher societal utility. When $\alpha = 0$, it is concerned only with the best off (for whom a treatment could be most cost-effective).
Variance-adjusted	$U(x) = U_{avg}(x) + k\sigma$, $k \in \mathbb{R}$, $\alpha \in [0,1]$	Adjust average utility by the standard deviation of the distribution. The parameter k allows control over how much the standard deviation affects the utility function.	When $k < 0$, treats health states with inequitable distributions of utility – where a given mean utility is attained by balancing those who are well off with many who do very poorly – as having lower societal utility. This is one way to operationalize equity.
Skew-adjusted	$U(x) = U_{avg}(x) + k\sqrt[3]{\text{skew}}$, $k \in \mathbb{R}$	Adjust average utility by the skew of the distribution. The parameter k allows control over how much the skew affects the utility function.	Can distinguish between health states based on their skew. For example, could penalize negatively-skewed distributions, whose health states leave many poorly off. Similarly, can favor health states that leave many well off.
nth moment	$U(x) = U_{avg}(x) + k\sqrt[n]{m^n}$, $k \in \mathbb{R}$	Adjust average utility by the nth root of the nth moment of the distribution (m^n). The parameter k allows control over how much the nth moment affects the utility function.	(Subject to the definition of m^n.)

2.3.3 Criteria for choosing among normatively acceptable aggregation procedures

As a way of illustrating the impact of alternative preference aggregation procedures, Figure 2.1 shows three hypothetical distributions of individuals' utilities for a health state. All three have the same mean value. The top distribution represents a health state with unanimity: everyone agrees on its utility. The middle distribution is bell-shaped. Utilities assigned to walking with a cane might have this shape, if they were elicited from individuals whose lifestyles range from sedentary to highly active. The bottom distribution is left-skewed. It might capture the utility assigned to imperfect but correctable vision, which is moderately high for most people, for whom glasses are only a minor inconvenience, but could be devastating for pilots who need perfect vision.

Because these three distributions have the same mean value, U_{avg} (equation (2.1)) treats them identically, thereby holding that the existence and nature of heterogeneity does not matter. Thus, the special needs of pilots might be washed out, when deciding what resources to allocate to vision research and treatment, just as the proportion of sedentary people in a population will affect the resources allocated to prime physical fitness. In these examples, the issue is *not* that members of these groups experience different health states (though they might) but that they value the health states differently, reflecting heterogeneity in preferences, which U_{avg} ignores.

As an example of a normatively acceptable aggregation function that addresses heterogeneity, consider

$$U(x) = U_{avg}(x) - k\sigma, k > 0.$$

It defines the utility of a health state as the mean of the distribution minus the standard deviation. As a result, it assigns a lower societal utility to health states with less societal consensus on their value (given a fixed mean). Rather than treating the three distributions in Figure 2.1 and their underlying health states equally, it would rank them $top > bottom > middle$. Table 2.1 and Roberts (1980) provide other aggregation options.

Figure 2.2 extends this logic to how the treatment of heterogeneous preferences can affect the allocation of resources across health states. It shows three sets of distributions of utilities for two health states, A and B. In the first (**a**) the distributions have the same shape but differ in their mean values. U_{avg} would assign a higher utility to B than to A. In the second (**b**), A and B have the same

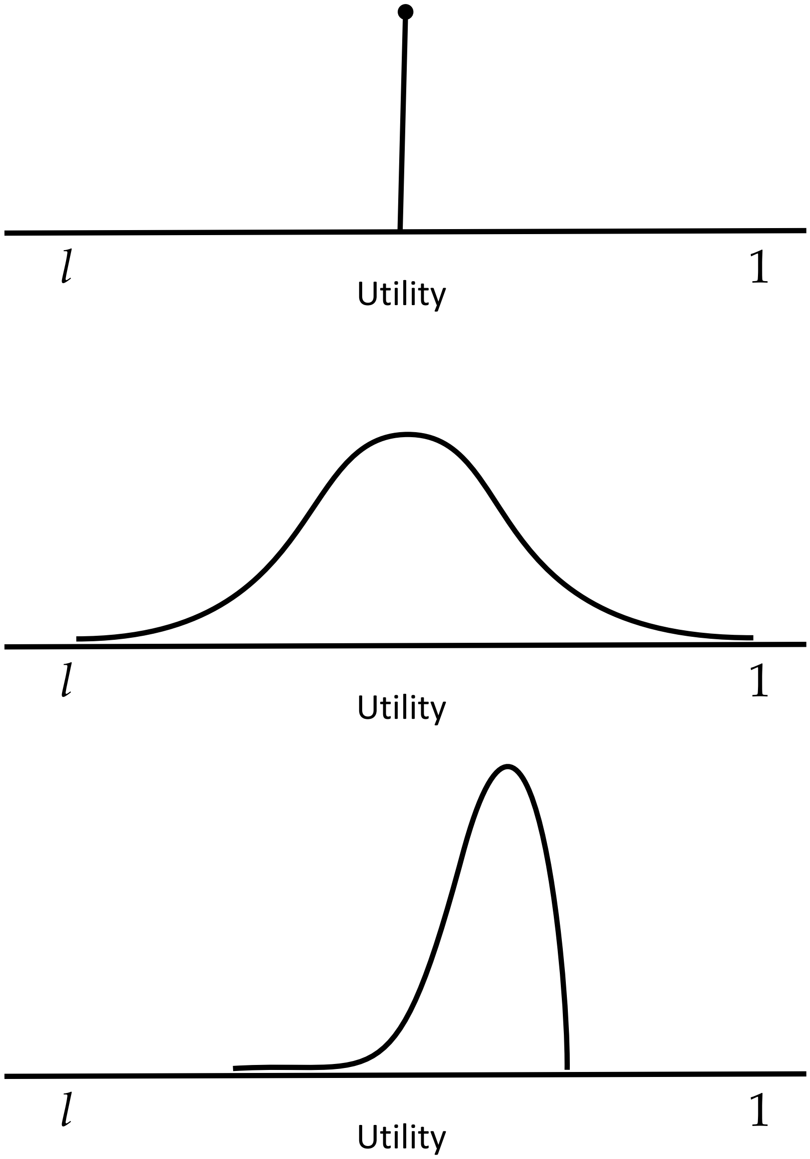


Figure 2.1: *Three distributions of utility.* Three distributions of utility corresponding to three hypothetical states of health. The maximum utility is 1, usually defined as the utility of full health, and the minimum utility is l, usually defined as the utility of some lower anchor state (e.g., dead). All three distributions have the same mean value of $\frac{1-l}{2}$. Thus, U_{Avg} would not be able to distinguish between them.

mean value, but B has a higher standard deviation, with some people valuing that state highly, whereas others are averse to it. A decision maker who valued equity – in the sense of being opposed to having some people with high utility, while others have low utility (for that same state) – would choose A over B; one who cared only about the average (U_{avg}) would be indifferent. A decision maker who wanted to ensure no one is too badly off would also prefer A to B, but for a somewhat different reason. In contrast, a decision maker focused on maximum values would choose B over A – as might happen when seeking medical treatments that make a big difference in some individuals' lives. The same social values might lead to preferring A to B given the distributions in **c**, which have similar means and variances, but differ in their skew. The fraction of people with outstanding health utility with A could outweigh the majority who fare somewhat worse than average.

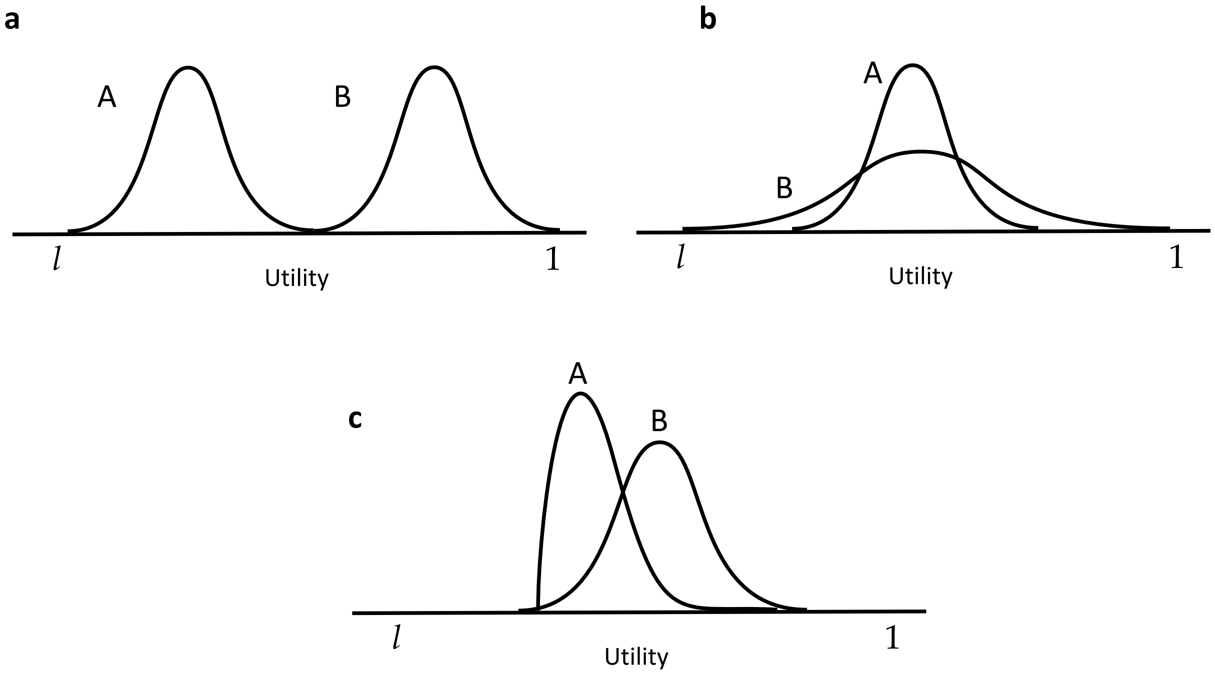


Figure 2.2: *Comparisons of distributions of utility.* Distributions of utility underlying hypothetical states of health. In **a**, *B* is simply a mean-shift of *A*. In **b**, *A* and *B* have the same mean, but different variances. In **c**, *A* and *B* have similar means and variances, but different skews. (Adapted from Fischhoff (1984, Figure 2).)

Relying on U_{avg} ignores such ethical concerns, treating each distribution of utilities as though it were the top distribution in Figure 2.1. Table 2.1 describes other potential aggregation procedures, along with the social values expressed by each. Choosing the aggregation procedure to use in analyses requires a preference – a *meta-preference* – over the set of normatively acceptable procedures (Sen, 1977). In the next section, we outline a method for deriving meta-preferences applicable to any societal preference-based measurement system.

2.3.4 A method for applying the criteria

In the absence of a dictator (Arrow, 1951), a socially acceptable approach is needed for selecting an aggregation procedure. We propose one that uses behavioral decision research methods (Edwards, 1954; Fischhoff & Kadvany, 2011) to implement an analytical-deliberative process, as advocated in *Understanding Risk* (Stern & Fineberg, 1996). A consensus report of the National Research Council, *Understanding Risk* proposes that defining the terms of analyses requires an iterative process, whereby analysts interact with decision makers to clarify the implications of alternative definitions (e.g., of societal utility functions and their associated aggregation procedures). We propose such a procedure for identifying socially acceptable societal utility functions. It has the following steps:

1. Select individuals with standing for making the choice.
2. Interview those individuals regarding the ethical principles that they wish to see in an aggregation procedure.
3. Select potential procedures from the (infinite) set of normatively acceptable procedures.
4. Develop materials for explaining the principles embodied in the procedures and their application to illustrative cases.
5. Elicit preferences (i.e., meta-preferences) among these procedures from individuals with standing.
6. Assess the construct validity of the elicited (meta-)preferences.
7. Repeat the process, as necessary.

1. Select individuals with standing for making the choice

By convention, societal preference-based HRQL scores reflect the preferences of the individuals who form society, depend on the healthcare system shaped by these scores, and pay its costs. That perspective could mean selecting a representative sample of the general public. However, one might also argue for disproportionate representation of individuals from groups such as insurers, regulators and providers, rather than members of the public. For example, one might justify that choice by claiming that such professionals are better informed about the "lifecycle" of medical conditions, and can put the public's interests above their own. Determining who has standing is outside social choice theory or any other mathematical formalism. Those individuals who have been chosen to have standing must be ensured the opportunity to articulate and express informed preferences (Fischhoff, 1991; Lichtenstein & Slovic, 2006).

2. Interview those individuals regarding the ethical principles that they wish to see in an aggregation procedure

Potential principles may come from philosophical analyses, legislation, or interviews, asking people to discuss allocation vignettes. The set of options should include the principles embodied in current approaches, in order to assess the social acceptability of the analytical conventions guiding them (Field & Gold, 1998). In addition to including widely discussed principles, such as differentially weighting end-of-life care or disease severity, the search should be broad enough to elicit principles that analysts might have neglected. Dolan, Shaw, Tsuchiya, & Williams (2005) review surveys that ask respondents to evaluate the relevance of such principles for various policy and personal decisions.

3. Select potential procedures from the (infinite) set of normatively acceptable procedures

Researchers should identify aggregation procedures that address the ethical concerns emerging from the previous step, screened to satisfy the axioms of social choice theory and categorized by what they assume about individual preferences (ordinal or cardinal) and the types of interpersonal comparability that they allow. Due to the possibly large number of potential aggregation procedures (as we saw with the HUI system), researchers may choose heuristics such as "absence of evidence is evidence of absence" in order to reduce the set of possible functions. For example, if no one mentions skew-related concerns, then researchers might reasonably ignore skew-sensitive aggregation procedures. Thus, there may be normatively acceptable procedures that express socially irrelevant (or unacceptable) principles, just as there may be principles that individuals endorse that violate the normative axioms or cannot be operationalized in a utility function. Principles of the last type could still play a role in the decision-making process, just not in the creation of the societal utility scores.

4. Develop materials for explaining the principles embodied in the procedures and their application to illustrative cases

To render informed preferences, participants need clear explanations of the procedures and their implications. For example, Wittenberg, Goldie, Fischhoff, & Graham (2003) used vignettes to explicate the principle of treating voluntarily and involuntarily incurred health effects differently. These vignettes presented scenarios about distributing finite medical resources among heterogeneous patient populations, illustrated with two examples (asthma, liver disease). There is an extensive empirical literature on methods for eliciting preferences for distributional justice (Charness & Rabin, 2002; Konow, 2003; Yaari & Bar-Hillel, 1984), some of which are used in studies eliciting utility scores for the Patient-Reported Outcomes Measurement Information System (PROMIS®) (Hanmer et al., 2015). As with any social research, careful development and pre-testing is needed to ensure that questions are interpreted as intended (Morgan, Fischhoff, Bostrom, & Atman, 2002).

5. Elicit preferences (i.e., meta-preferences) among these procedures from individuals with standing

Using the materials developed in the previous section, elicit preferences for defining societal preferences for health. Given the complexity of the task, an iterative process may be needed to ensure that participants understand the issues and the implications of their expressed (meta-)preferences. Following decision-analytic procedures, the protocol would have a skilled facilitator or use interactive internet-based methods to help participants articulate the implications of their basic values for these specific questions (Keeney & Raiffa, 2003). Instructions must ensure that participants understand the roles that gave them standing in the process (e.g., answering on behalf of their present or future selves, their families, the general public). That may mean acting as though they were behind a "veil of ignorance," not knowing how the procedures will affect them (Rawls, 1971), or applying some other principle from distributive justice that can guide them and the interpretation of their responses (Konow, 2003).

6. Assess the construct validity of the elicited meta-preferences

Construct validity assesses responses' internal and external consistency (Cronbach & Meehl, 1955). Internal consistency can be evaluated with tests such as scope sensitivity, namely, whether individuals prefer more of a valued outcome to less (when the comparison is not transparent). External consistency can be evaluated by tests such as whether individuals who self-identify as egalitarian also favor egalitarian aggregation procedures.

7. Repeat the process, as necessary

All empirical measures are imperfect. As a result, policy makers need to decide whether a set of expressed (meta-)preferences is good enough to guide social policy. One natural contrast may be whether the judgments of lay respondents are superior to those of the experts who would otherwise define social preferences. Experts are likely to have a better understanding of the technical issues, while lacking insight regarding lay concerns – unless they wish to claim that they know the public better than it knows itself.

2.4 Conclusion

Any societal utility function summarizes a distribution of individual utilities with a single number, necessarily making a value judgment about which features of the distribution matter. The procedure used in the HUI is a normatively justified aggregation procedure, in social choice theory terms. However, there are infinitely many other aggregation procedures that are normatively justifiable as well. Eliciting preferences for aggregation procedures, or meta-preferences, provides an empirical basis for choosing among those possibilities. This framework could be implemented during the construction of new societal preference-based measures of HRQL or adapted for pre-existing measures. Its logic applies to the design of discrete choice experiments in public policy domains, where similar preference aggregation is required.

Our framework makes explicit the ethical content of aggregating individual utilities into societal HRQL estimates, avoiding the potential for unintended ethical consequences created by policy choices made without normative analyses (Paulden, O'Mahony, Culyer, & McCabe, 2014; Culyer, 2006). Our approach complements other approaches that adjust HRQL estimates defined by conventional aggregation methods (e.g., the mean) by other factors (Asaria et al., 2015; Dolan et al., 2005; Nord et al., 1999), such as the severity of individuals' health states (Nord et al., 1999). It is also consistent with the *ethos* of societal preference-based HRQL measurement, which holds that societal utilities should be defined by social values (Gold et al., 1996), rather than determined *a priori* by some authority – although experts could suggest principles for societal representatives to consider.

Thus, we propose an analytical-deliberative process for addressing the well-documented heterogeneity in health-state utilities (Torrance et al., 1996; Hogg et al., 2013; Owens & Shekelle, 2013; Lilford et al., 2007; Basu & Meltzer, 2007; Kravitz, Duan, & Braslow, 2004; Volk et al., 2004). It enlists individuals with standing in evaluating normatively acceptable aggregation procedures, including the conventional averaging method. Who has standing for choosing aggregation procedures is a political-ethical question. The sample might be drawn from the general public, individuals with a condition, experts, patient advocates, etc. Unless the analytical-deliberative process produces a consensus, the distribution of (meta-)preferences that it elicits could be a source of inputs to sensitivity analyses. Fundamentally different ethical principles might still lead to the same choices, as has been found in risk perception studies (e.g., whether risks are incurred involuntarily and have delayed effects) (Fischhoff, 2015; Fischhoff & Kadvany, 2011; Fischhoff & Morgan, 2009). Of course, sensitivity analyses that reflect variation in the statistic used to summarize preferences are asking a very different question than sensitivity analyses that reflect disagreement about what summary statistic to use in the first place. Unlike the former source of uncertainty, the latter would remain even with error-free measurement of the preferences of every individual in the population.

We hope that our proposal will advance research into the choice of aggregation procedure and clarification of the ethical issues that it inevitably entails (Field & Gold, 1998; Lilford et al., 2007), by connecting the formal analyses of social choice theory and the empirical procedures of behavioral decision research. The ultimate goal is to ensure that analytical methods reflect the values of the individuals whose welfare they affect (Fischhoff, 2015).

Part II

The PROMIS-Preference (PROPr) Scoring System

3

Creation of the PROPr Scoring System

3.1 Introduction

Health-related quality of life (HRQL) is often assessed by using health-profile measures that capture one or several domains of health, such as physical function and mental health.[1] Such measures are used to evaluate health interventions, in epidemiologic studies, and to monitor the health of populations. Measures based on societal preferences for health-related quality of life provide summary scores to not only assess, track, and compare the health of populations, but also to conduct decision and cost-effectiveness analyses. Several measures are available including the EQ-5D-3L/5L, the Health Utilities Index (HUI) Mark 2 and Mark 3, the SF-6D, and the Quality of Well-Being Scale (Brazier, Roberts, & Deverill, 2002; EuroQol Group, 1990; Feeny et al., 2002; Herdman et al., 2011; Kaplan & Anderson, 1995; Torrance et al., 1996). The strengths and weaknesses of these generic measures have been widely discussed (Feeny, Krahn, Prosser, & Salomon, 2016; Fryback, Palta, Cherepanov, Bolt, & Kim, 2010; Kaplan et al., 2011; McNamee & Seymour, 2005; Tosh, Brazier, Evans, & Longworth, 2012).

Each such measure combines a *state space* of *health profiles* (i.e., a health classification system) with a *scoring function* that associates a number (*utility*) with each state. These numbers are treated as cardinal (interval-scale) *utilities*, representing preferences for health (Keeney & Raiffa, 2003; von

[1] The work presented here was overseen by the PROPr team, which includes Janel Hanmer (PI), David Feeny, Baruch Fischhoff, David Cella, Ron D. Hays, Rachel Hess, Paul A. Pilkonis, Dennis A. Revicki, Mark S. Roberts, Joel Tsevat, and Lan Yu. The PROPr website will have a link to the most up-to-date version of this chapter's content (http://janelhanmer.pitt.edu/PROPr.html).

Neumann & Morgenstern, 1944). By convention, a societal measure is created by aggregating the preferences of a sample of individuals (Dewitt et al., 2017; Torrance et al., 1982).

Since 2004, the National Institutes of Health has funded the development and dissemination of a health profile measurement system developed using item response theory (IRT), the Patient-Reported Outcomes Measurement Information System (PROMIS®) (Cella et al., 2010; Cella, Yount, et al., 2007; Cella, Gershon, Lai, & Choi, 2007). PROMIS has developed and evaluated item banks (Embretson & Reise, 2000; Reeve et al., 2007) for many HRQL domains (e.g., pain, physical function, sleep, social activity). The PROMIS item banks are freely available, customizable for individual studies, and comparable across studies (Collins & Riley, 2016; Gershon, Rothrock, Hanrahan, Bass, & Cella, 2010). Here, we apply decision theory methods to estimate the utility of health states for selected PROMIS domains for use in research, population health management, and policy analyses that require such estimates. We call the resulting scoring system the PROMIS-Preference (PROPr) scoring system.

PROPr is grounded in utility theory and designed to avoid the ceiling and floor effects sometimes observed with other measures (Fryback et al., 2010). Figure 3.1 provides an overview of our approach. From the left, PROMIS scores (A) are inputs to the PROPr single-attribute scoring functions (B) that yield utilities for each domain (C). PROPr applies a multi-attribute function to combine the single-domain scores (D), and produce a summary score (E). Hanmer and colleagues (2015) and the PROPr technical report (Hanmer & Dewitt, 2017), available in the Appendix, describe the process that led to the development of PROPr. The PROPr methods and single- and multi-attribute scoring functions are described here.

3.2 Methods

The PROPr scoring system uses the normative theory of preferences expressed in multi-attribute utility theory (MAUT) (Keeney & Raiffa, 2003; Luce & Suppes, 1965; Savage, 1972; von Neumann & Morgenstern, 1944). If underlying assumptions are met, MAUT procedures produce a utility function that can be treated as *cardinal*, meaning that it is measured on an interval scale that allows comparing utility differences (Coombs, Dawes, & Tversky, 1970). Cardinality is required to

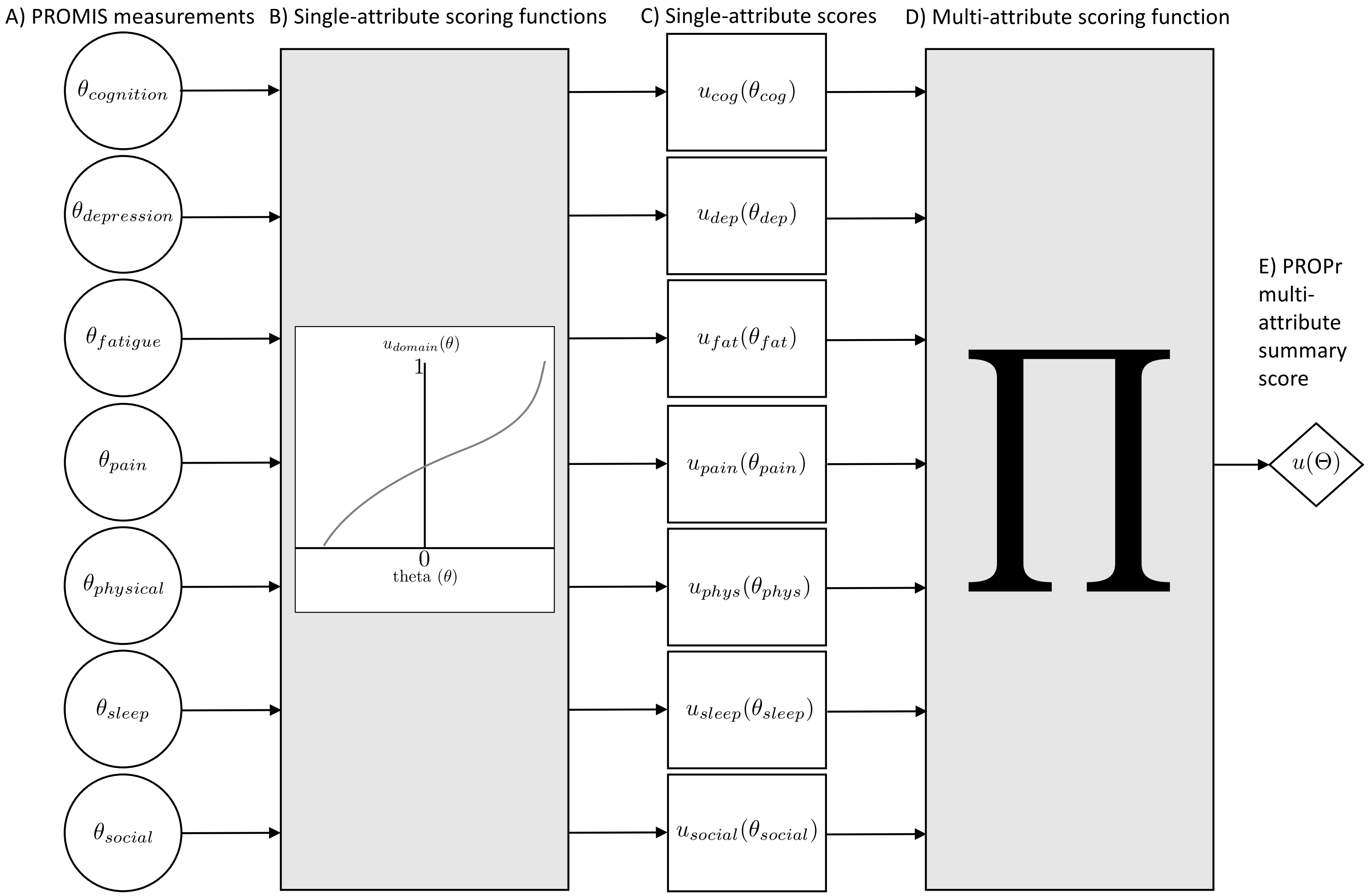


Figure 3.1: *The PROPr scoring system conceptual model.* In A), a measurement on one of the 7 PROMIS domains used in PROPr, denoted θ, is the input to its single-attribute scoring function u_{domain}. In B), the output of $u_{domain}(\theta)$ is a score on the scale where 0 is the utility of that domain's disutility corner state and 1 is the utility of full health. If we have all 7 PROMIS measurements, then we can take the outputs from the 7 single-attribute scoring functions (C) and use them as inputs to the multiplicative multi-attribute scoring function (D). The multi-attribute function produces a summary score, $u(\theta)$, for the entire vector Θ of 7 PROMIS measurements, on the scale where 0 is the utility of dead and 1 is the utility of full health (E).

combine the utility of morbidity and mortality. The PROPr scoring system applies the methodology of the Health Utilities Index Mark 2 and Mark 3 (Feeny et al., 2002; Furlong et al., 1998; Torrance et al., 1982, 1996) to preferences for PROMIS-defined health states, elicited from a U.S. nationally representative survey. The next 2 sections describe the health-state space used in PROPr and the preference survey. They are followed by descriptions of the analytical methods used to produce the 7 PROPr single-attribute scoring functions and the summary multi-attribute scoring function.

3.2.1 Health-state space

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

PROPr focuses on a subset of PROMIS domains, chosen to span the overall space, so that they would form a common set that would be of interest to the public, patients, and researchers. We also imposed the constraint, required by MAUT, that the domains be *structurally independent*, in the sense that all states could conceivably occur (Keeney & Raiffa, 2003). For example, physical function and depression are structurally independent so long as one can imagine a high score on one and a low score on the other, high scores on both, and low scores on both. Two domains can be structurally independent even if they are empirically correlated. Hanmer and Dewitt (2017) describe the procedure that identified the 7 PROMIS health domains in the PROPr state space: Cognitive Function – Abilities v2.0 (*cognition*); Depression v1.0 (*depression*); Fatigue v1.0, Pain – Interference v1.1 (*pain*); Physical Function v1.2 (*physical function*); Sleep Disturbance v1.0 (*sleep*); and, Ability to Participate in Social Roles and Activities v2.0 (*social roles*). All currently available physical function item bank versions (v1.0, v1.1, v1.2, and v2.0) and pain item bank versions (v1.0 and v1.1) can be used with PROPr. PROPr requires at least v2.0 of the cognition and social roles item banks – their 1.0 versions cannot be used. When new item banks become available, the PROMIS documentation will describe whether they are compatible with those used to develop PROPr; if so, they can be used with the PROPr scoring system.

Each domain in PROPr (and PROMIS) is treated as a continuous latent construct, called *theta* in

IRT. The domains are theoretically unbounded in both directions. They are constructed so that the population mean on theta is 50, with a standard deviation of 10, which is called a *T-score* scale. PROPr uses a z-score scale, a linear transformation of T-scores such that the mean is 0 and the standard deviation is 1. PROMIS scores rarely fall outside the range -4 to 4 on theta (a T-score of 10-90).

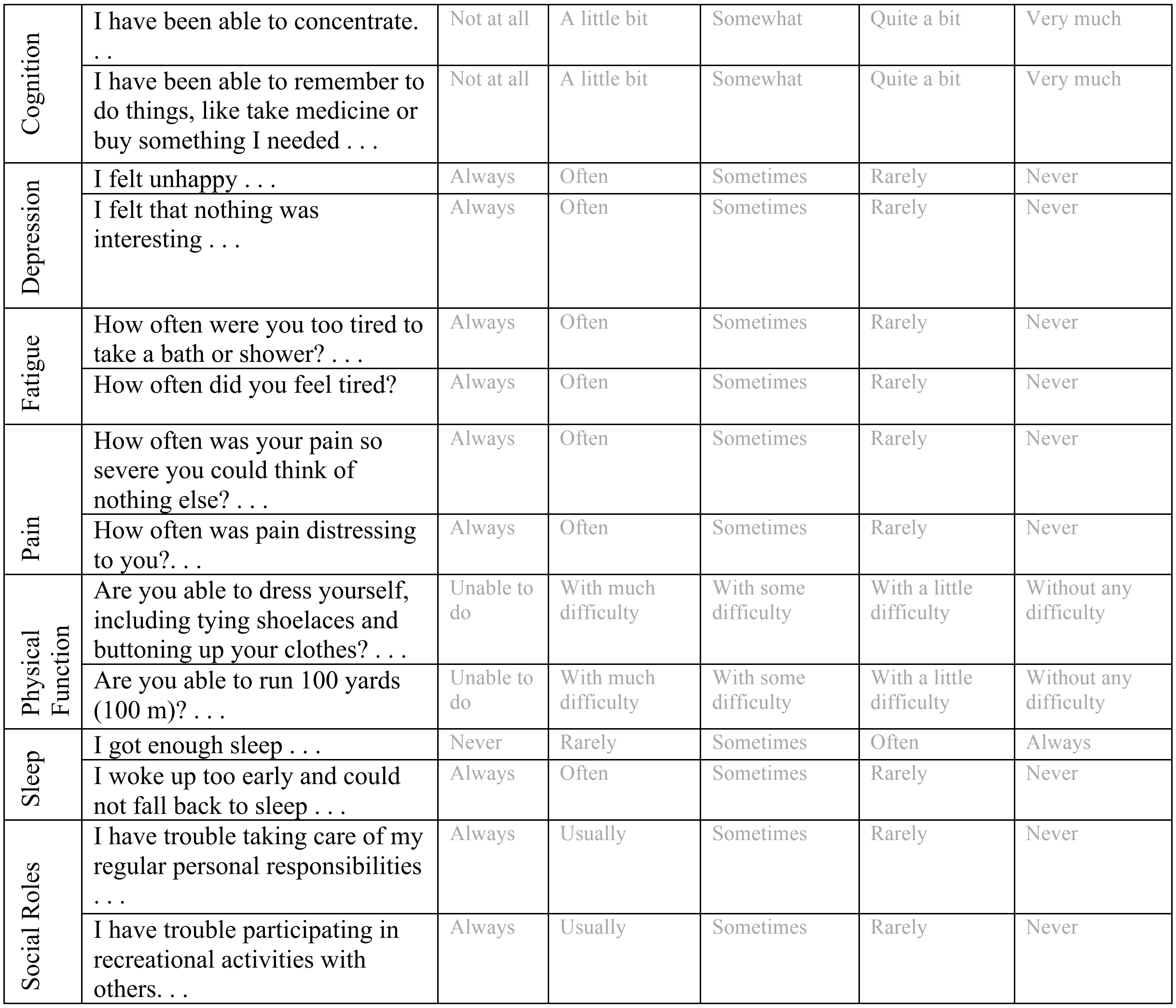
A functional capacity on a domain is called a *level* of theta. A health state (or profile) in PROPr is a vector with 7 elements, each corresponding to a level on 1 domain. Each domain was represented by the 2 items in Figure 3.2 (e.g., cognition was expressed as ability to concentrate and ability to remember). Levels of those items (e.g., not at all, a little bit) were chosen to represent 8 or 9 health states that spanned the space of theta values, with those levels selected to obtain the data necessary to construct the scoring system (Table 3.1; see Hanmer and Dewitt (2017) for fuller details).

Table 3.1: *PROMIS theta scores used in PROPr elicitation tasks.* The table shows the theta values corresponding to the health state descriptions valued in the PROPr survey. The levels between the unhealthiest and the healthiest correspond to the intermediate states valued in valuation set (i) of the elicitation task (see Part II or Part III). The unhealthiest levels, together, define the *all-worst state*, while the best levels, together, define *full health*. The *disutility corner state* for a domain corresponds to the state described by the worst level on that domain, and the best on all others.

PROMIS Domain	**Unhealthy**	· · ·		· · ·		· · ·		· · ·	**Healthy**
Cognition	-2.77	-2.00	-1.57	-1.20	-0.85	-0.50	-0.13	0.34	1.12
Depression	3.45	2.56	2.02	1.57	1.12	0.72	0.31	-0.19	-1.13
Fatigue	3.77	2.65	1.96	1.41	0.97	0.53	-0.07	-0.92	-2.27
Pain	3.20	2.42	1.92	1.57	1.25	0.96	0.65	0.28	-0.43
Physical functioning	-3.55	-2.66	-2.14	-1.64	-1.06	-0.67	-0.39	0.03	0.97
Sleep	3.45	2.44	1.13	0.60	0.05	-0.50	-1.15	-2.49	
Social	-2.40	-1.72	-1.38	-1.03	-0.70	-0.34	0.02	0.43	1.15

3.2.2 Survey overview

We collected preference data by using an online instrument administered by ICF (https://www.icf.com/services/research-and-evaluation) and SurveyNow (http://www.surveynowapp.com/). A description of the survey is available in the Appendix and



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

Figure 3.2: *Health-state descriptions in the PROPr survey.* Health-state descriptions were given as a table like the one above, with one answer selected for each item (row). For example, the health state describing the highest functional capacity on each domain would have the rightmost column selected for all items. The health state describing the lowest functional capacity on each domain (called the *all-worst state*) would have the leftmost column selected for all items.

technical report (Hanmer & Dewitt, 2017). The present analyses focus on the preference elicitation task, which was preceded in the survey by demographic questions and questions about the participant's health, including the PROMIS-29 inventory and the Cognition 4-item short-form (Gershon et al., 2010; PROMIS, 2015). In the preference elicitation task, participants evaluated the states spanning the range of 1 randomly chosen health domain from among the 7 included in PROPr, along with several other multi-domain health states, following the procedure below.

As compensation for completing the survey, participants could choose from 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 the authors received them.

In pre-testing, we found that participants could not thoughtfully complete the survey and read the essential introductory instructions in under 15 minutes. We therefore only used data from surveys completed in 15 minutes or longer.

3.2.3 Multi-attribute scoring function

PROPr associates a cardinal utility with each health state in PROPr's 7-domain state space, allowing each state to be compared with each other state on an interval scale.

MAUT identifies the normatively justified models for scoring functions, mapping states onto interval scales (Keeney & Raiffa, 2003). The 3 most common are the *linear additive, multiplicative,* and *multi-linear* models. They differ in their assumptions about interactions among preferences (i.e., how evaluations of states on one attribute depend on the state on other attributes). The linear additive model is the most restrictive; it assumes that preferences do not interact. The multi-linear model allows pairs of attributes (e.g., PROMIS domains) to be *preference complements* (e.g., if the disutility of being immobile and socially isolated is lower than the sum of the individual disutilities) or *preference substitutes* (e.g., if the disutility of being immobile and socially isolated is higher than the sum of the individual disutilities) (Feeny et al., 2002, p. 116). The multiplicative model allows all pairs of domains to be preference complements or substitutes, but not both (Furlong et al., 1998). The linear additive model is a special case of the multiplicative model.

Following the methods described by Furlong and colleagues (1998) and Feeny and colleagues

(Feeny et al., 2002), our preference elicitation survey collected responses needed to fit a multiplicative model. Although the multi-linear model is more flexible, and hence might provide a better fit, its data requirements were beyond the present project's capability. The PROPr procedures evaluate the appropriateness of the linear additive model (step 3, below).

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

$$u(\Theta) = \frac{1}{k}\left(\prod_{i=1}^{m}(1 + k \cdot k_i \cdot u_i(\theta_i)) - 1\right), \tag{3.1}$$

where,

$$\left(\prod_{i=1}^{m}(1 + k \cdot k_i)\right) - k - 1 = 0. \tag{3.2}$$

The k_i terms are utilities of the *corner states*, defined as ones with the best level on the ith attribute and the worst on all other attributes. The k term is the *global interaction constant*, which measures preference interactions among all the attributes; a negative value indicates that the domains are preference substitutes, and a positive value indicates that they are complements (Feeny et al., 2002; Keeney & Raiffa, 2003).

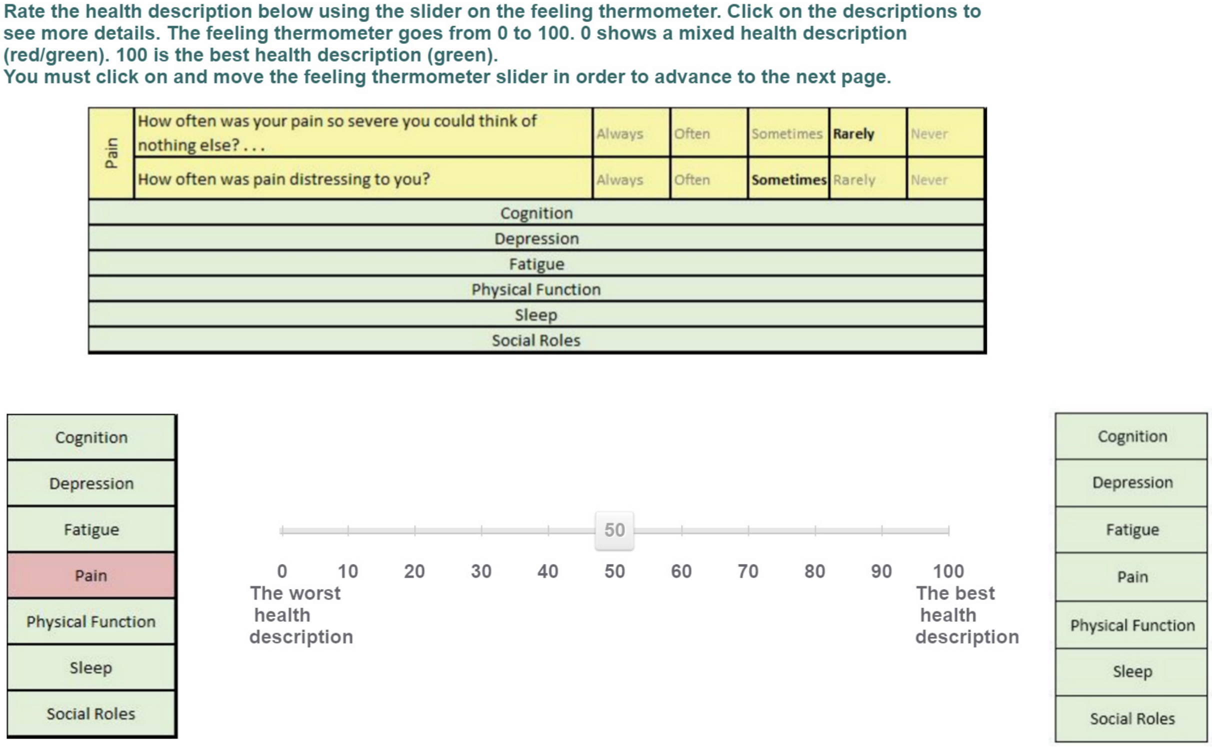
Following the method described by Feeny and colleagues (Feeny et al., 2002), the PROPr function is calculated in *disutility* terms, and then transformed to utility as *utility* = 1− *disutility*. This procedure asks participants to envision *disutility corner states*, with the unhealthiest level on the ith domain and the healthiest level on all other domains (Torrance et al., 1996).

3.2.4 Preference elicitation

Participants valued 2 sets of states, first using a visual analogue scale (VAS) and then a standard gamble (SG) (Gafni, 1994; Torrance, Feeny, & Furlong, 2001). The VAS task was intended to introduce the health states valued in the SG task (Torrance et al., 2001). The SG task was used for PROPr because of its grounding in expected utility theory (Keeney & Raiffa, 2003; von Neumann & Morgenstern, 1944).

The VAS had a 0-100 scale (sometimes called a Feeling Thermometer), where 0 is the value of a lowest health state and 100 the value of *full health*, the state with the highest functional capacity on all domains. Figure 3.3 shows the VAS, eliciting the impact of an intermediate state for pain. We asked participants to rate the health state that is perfect in all respects except on pain, where we asked them to consider rarely having pain so severe that they could think of nothing else and sometimes having pain that is distressing.

The SG task for the same intermediate health state poses a choice between (a) having this state with certainty and (b) a lottery with probability p of full health and $(1-p)$ for the bottom state (see below). The SG procedure offers a series of choices, changing the probability p until the participant is indifferent between the options. Following utility theory assumptions, this probability, $\hat{p}$, is the utility of the intermediate state. Figure 3.4 shows one gamble in such a sequence, with the intermediate state described at the top right (Choice B) and the gamble at the left (Choice A).



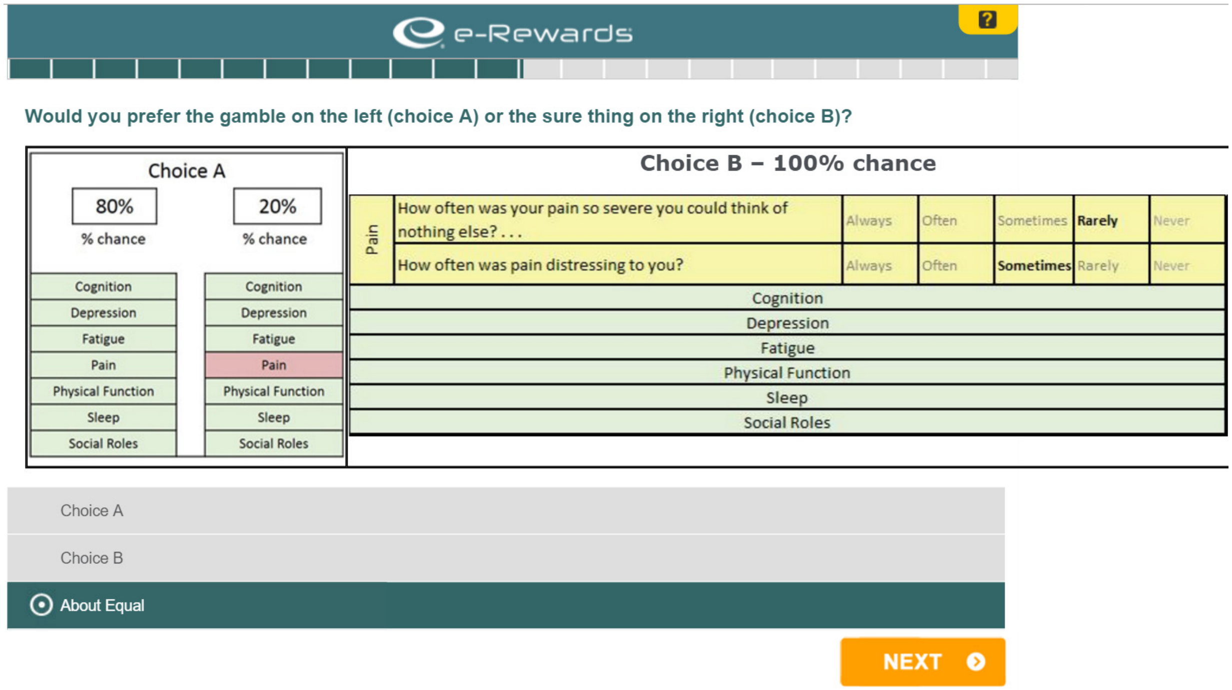
Rate the health description below using the slider on the feeling thermometer. Click on the descriptions to see more details. The feeling thermometer goes from 0 to 100. 0 shows a mixed health description (red/green). 100 is the best health description (green).
You must click on and move the feeling thermometer slider in order to advance to the next page.

Pain	How often was your pain so severe you could think of nothing else? . . .	Always	Often	Sometimes	**Rarely**	Never
	How often was pain distressing to you?	Always	Often	**Sometimes**	Rarely	Never
Cognition						
Depression						
Fatigue						
Physical Function						
Sleep						
Social Roles						

Figure 3.3: *The Visual Analogue Scale.* An example valuation, using the Visual Analogue Scale (VAS).

(a) Set (i)

We randomly assigned participants to assess 1 of the 7 health domains (e.g., cognition). Each participant evaluated the 6-7 states for that domain, selected to represent the intermediate theta



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

Choice A

80% % chance	20% % chance
Cognition	Cognition
Depression	Depression
Fatigue	Fatigue
Pain	Pain
Physical Function	Physical Function
Sleep	Sleep
Social Roles	Social Roles

Choice B – 100% chance

Pain	How often was your pain so severe you could think of nothing else? . . .	Always	Often	Sometimes	**Rarely**	Never
	How often was pain distressing to you?	Always	Often	**Sometimes**	Rarely	Never
Cognition						
Depression						
Fatigue						
Physical Function						
Sleep						
Social Roles						

Choice A

Choice B

● About Equal

Figure 3.4: *The Standard Gamble.* An example step in a Standard Gamble (SG) valuation. Choice A shows some gamble between the best and worst health states in the given domain – in this case, pain. Choice B shows the sure-thing of some intermediate health state.

values in Table 3.1 and described in verbal terms in Figure 3.2. The bottom state on these valuations was always the disutility corner state for the given domain (corresponding to the worst possible level in that domain and perfect health on all others, as in Figure 3.3 and Figure 3.4). Figure 3.5A illustrates this process, for the cognition domain.

(b) Set (ii)

Recognizing that participants may consider some states to be worse than dead (Feeny et al., 2002), we asked them whether they preferred the dead state or the state with the lowest level on each of the 7 domains (the *all-worst state*). We treated the option *not* chosen as the bottom state for the participant's valuations in this set. Participants then valued the disutility corner state for their assigned domain in set (i). They also valued 2 other states, randomly selected from the disutility corner states for the other domains, and 3 *marker states*, chosen to span the health state space (Feeny et al., 2002). Finally, participants valued either dead or the all-worst state, depending on which they had selected as better (Figure 3.5B and 3.5C).

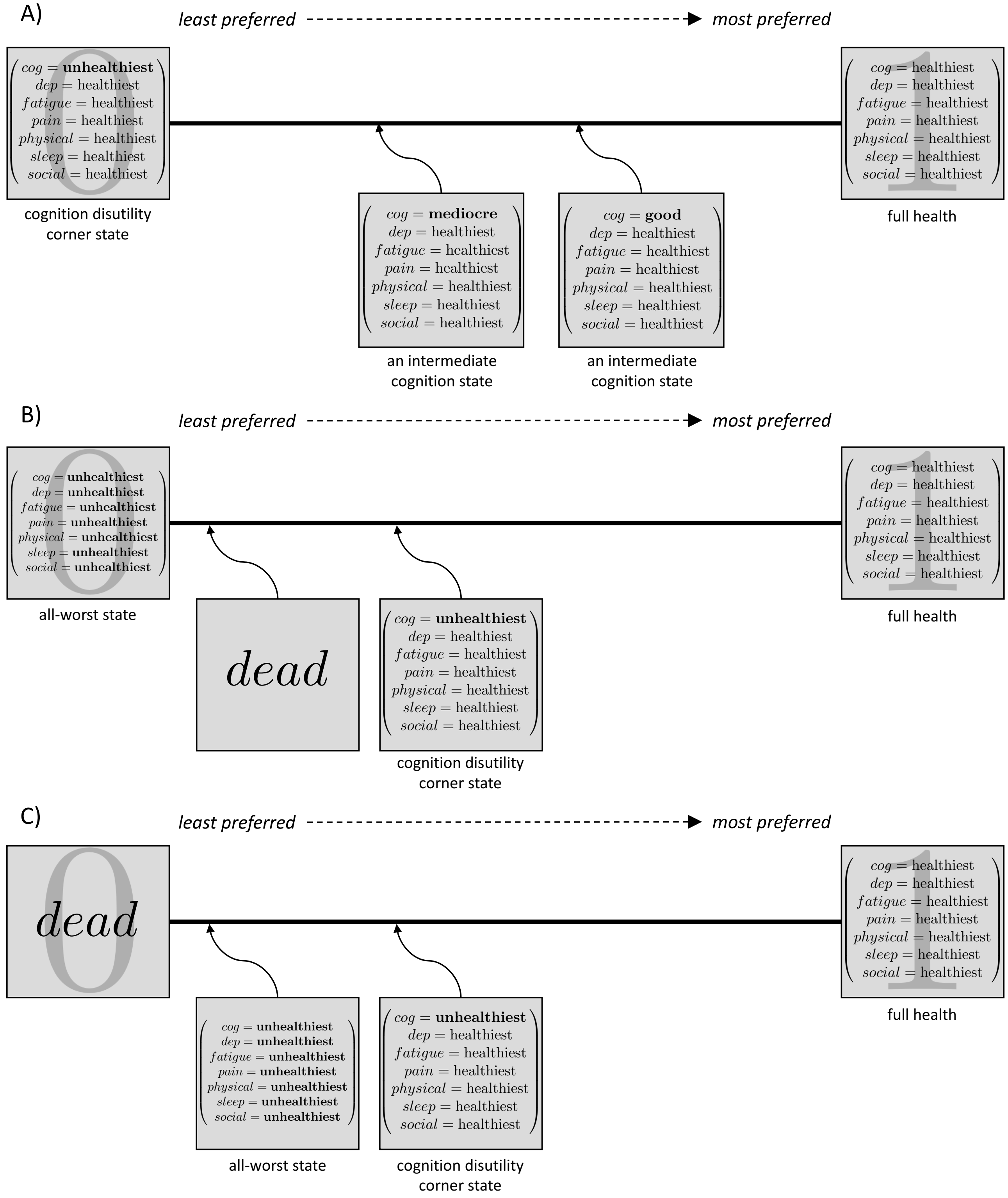


Figure 3.5: *Example state valuations.* An example, using the cognition domain, of the data produced by the preference elicitations, in utility terms. (The associated disutility scale is produced by taking 1?*utility*.) In A), the participant evaluates intermediate states of cognition on a scale from the unhealthiest level of cognition (the cognition disutility corner state) to full health. In B), a participant who prefers the state of dead to the all-worst state values dead and the cognition disutility corner state on a scale from the all-worst to full health; panel C) shows the output of someone who prefers the all-worst state to dead. Panel A) corresponds to set (*i*) in the main text, and panels B) and C) to set (*ii*).

3.2.5 Calculating the PROPr scoring system

To create the PROPr scoring system, we first calculate a single-attribute scoring function for each PROMIS domain, with 1 equal to the utility of full health and 0 equal to the utility of that domain's disutility corner state. The 7 single-attribute functions are combined to produce a multi-attribute summary scoring function, where 0 is the utility of dead and 1 is the utility of full health, with scores less than 0 corresponding to states judged worse than dead. Specifically, the creation of the PROPr scoring system follows these steps:

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

Following prior work (Feeny et al., 2002; Furlong et al., 1998), we excluded the highest and lowest 5% of elicited utilities (10% trimming) for each health state.

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

We chose the functional form for each domain via k-fold cross-validation, comparing polynomial models with a smooth spline, predicting the conditional mean disutility for each theta value (Shalizi, n.d.). We compared the predictive power of smoothing spline models with that of simpler parametric models in terms of root-mean-squared-error (rMSE) on the split data sets used for cross-validation. Having no prior beliefs about whether the form of the single-attribute functions varied by domain, we used the same functional form for all domains. We wanted the rMSE difference to be below 0.01 – 1% of the utility scale, too low to have any theoretical or practical importance (Grootendorst, Feeny, & Furlong, 2000; Samsa et al., 1999), and to have similar curvature to the smooth spline.

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

We calculated the mean values of the disutility corner states separately for participants who preferred the all-worst state to dead and for participants who preferred dead to the all-worst state. We used an affine transformation to translate the values produced by the former group to the scale of the latter. We then combined estimates from the 2 groups, weighting each by its size. Thus, the resulting values are on the all-worst to full health disutility scale.

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

MAUT determines the relative fit of the multiplicative and linear additive models by the sum of the k_is, with the linear additive being superior only if that sum equals 1 (Keeney & Raiffa, 2003). When the linear additive model is superior, the global interaction constant is 0. Otherwise, the global interaction constant for the multiplicative model is determined by solving equation (3.2), using the disutility corner state values calculated in Step 2. That equation is a polynomial in the global interaction constant, hence can have several real roots. Which root is the constant is determined by theorems from MAUT (Keeney & Raiffa, 2003, Appendix 6B).

4. Create the multi-attribute disutility function

We used the values derived from Steps 1-3 to specify the multi-attribute disutility function. If step (3) indicated that the multiplicative model is the better functional form, then the multi-attribute disutility function uses equation (3.1), written in disutility terms:

$$\bar{u}_{AW}(\Theta) = \frac{1}{c}\left(\prod_{i=1}^{7}\left(1 + c \cdot c_i \cdot \bar{u}_i(\theta_i)\right) - 1\right). \tag{3.3}$$

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

$\bar{u}_i$, each multiplied by its respective c_i.

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

By definition, the utility function $u_{AW}(\Theta)$ equals $1 - \bar{u}_{AW}(\Theta)$. Following the transformation procedure of step 2, we rescaled the disutility function to a utility function, $u(\Theta)$, where 0 equals the utility of dead and 1 equals the utility of full health:

$$u(\Theta) = 1 - \frac{\bar{u}_{AW}(\Theta)}{\bar{u}_{AW}(dead)}, \tag{3.4}$$

where $\bar{u}_{AW}(\Theta)$ is equation (3.3), and $\bar{u}_{AW}(dead)$ is the mean disutility value of dead on the all-worst to full health scale.

6. Perform sensitivity analyses

Societal health utility measures aggregate the preferences of individuals. Those in the present sample were selected to represent the U.S. adult population capable of using the survey (i.e., those with sufficient vision and English literacy). However, not all data provided were of equal quality. As mentioned, we excluded both participants who spent less than 15 minutes on the task and those with the highest and lowest 5% of utility values. There is no obvious reason for those who rushed through the task to have distinctive health state preferences. However, the highest and lowest values might reflect either noisy responses (from individuals who responded casually or without fully understanding the task) or thoughtfully produced, but unusual ones, reflecting individuals for whom a health domain is particularly important or unimportant (e.g., physical function for someone who is athletic or sedentary). We did not exclude "out-of-bounds" responses, where the SG estimated a health state utility below 0 or above 1, reflecting violations of dominance (Keeney & Raiffa, 2003), but rounded them to 0 and 1, respectively. We conducted 4 sensitivity analyses to estimate the effects of these choices, repeating the analysis with:

i. No minimum completion time threshold, with 10% trimming

ii. 15-minute completion threshold, without 10% trimming

iii. 15-minute completion threshold and 10% trimming, excluding "out-of-bounds" responses (rather than adjusting them to 0 or 1)

iv. A "stringent criteria" subsample.

Case (iv) excluded participants who met any of the following exclusion criteria: spent less than 15 minutes on the survey; violated dominance more than twice; used less than 10% of the scale for all valuations; rated understanding of the survey questions as less than 2, on a scale of 1 = "Not at all" to 6 = "Very much"; had a numeracy score of less than 2.5, on a scale from 1-6 (McNaughton, Cavanaugh, Kripalani, Rothman, & Wallston, 2015); or, rated dead or the all-worst state as equal to full health. Similar exclusion have been used in other studies (Engel, Bansback, Bryan, Doyle-Waters, & Whitehurst, 2016).

We calculated the multi-attribute scoring function for our core sample (10% trimming and 15-minute completion threshold) and for the samples defined by cases (i)-(iv). We next estimated the health utility of each participant in the core sample with each of these 5 functions, using that individual's health profile as reported on the survey's PROMIS-29 inventory and Cognition 4-item short-form. We then calculated linear correlations between these 5 utility scores (for participants' current health state). As the disutility corner state values determine the weighting of the single-attribute functions in the final summary scoring function (equation (3.3)), we also calculated the linear correlations between the disutility corner state values defined by the core sample and those defined by the 4 alternative cases.

3.3 Results

Of the 2,026 individuals who were invited to the survey, 1,779 completed the consent form (87.8%) and 1,164 (57.5%) completed the entire survey. Of the 615 people who completed the consent form but did not complete the survey, 331 dropped out before the health state valuation section. The median survey completion time was 25 minutes, with 983 participants spending at least 15 minutes – defining the core sample. The 10% trimming procedure removed individual *responses*, but not entire participants. Overall, 630 (64.1%) participants chose dead to be better than the all-worst state, and the remainder (353) the opposite.

3.3.1 Sample demographics

The sample's demographic characteristics largely match the U.S. 2010 Census except that the core sample was slightly older, more educated, reported higher income, and had a larger proportion of White individuals than the U.S. population (Table 3.2). In the core sample, reported overall health status was excellent for 12.5%, very good for 39.4%, good for 33.8%, fair for 12.4%, and poor for 1.9%.

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

Figure 3.6 shows the 7 single-attribute disutility functions, where the x-axis is the construct measured on the PROMIS z-score scale and the y-axis is disutility. For example, the upper left graph shows utilities of the PROMIS cognition domain. The curves for cognition and physical functioning slope downwards because they improve as theta increases, while the rest improve as theta decreases. Our modeling procedure indicated a cubic polynomial for each domain, constrained to go through 0 and 1 at the domain endpoints (Hanmer & Dewitt, 2017).

2. Calculate the mean value of the disutility corner states

The corner states had a range of disutility values (Table 3.3). The greater the magnitude of the number, the more weight that domain had in the final (dis)utility calculation.

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

The sum of the disutility corner states was 4.45. For the linear additive MAUT model to have been a more appropriate fit, the sum would have had to have been 1 and the interaction constant would have been 0 (Keeney & Raiffa, 2003; Torrance et al., 1996). Using the disutility corner state values and equation (3.2), and following Appendix 6B in (Keeney & Raiffa, 2003), the global interaction constant for the multiplicative model is -0.999. The value of the global interaction constant indicates that the domains are preference complements, as has been the case in all versions of the HUI (Feeny et al., 2002).

Table 3.2: *Participant demographics.* The first column shows the expected demographic characteristics based on the U.S. 2010 Census. The second column shows the demographic characteristics of the participants who completed the survey, and the final column shows the demographic characteristics of the core sample used in producing the scoring system.

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

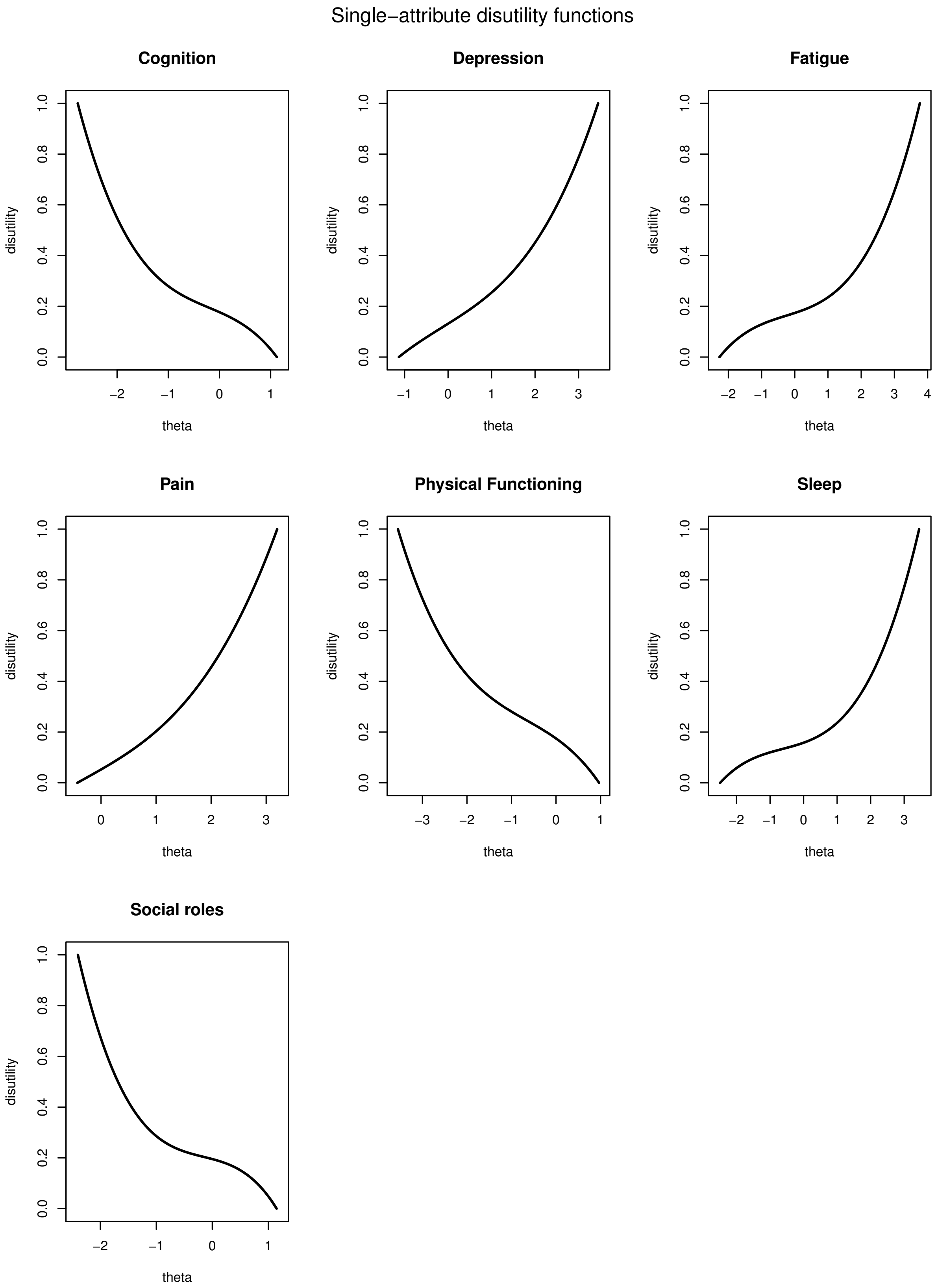


Figure 3.6: Single-attribute disutility functions.

Table 3.3: *Disutility corner state values.* The disutility corner state values of each domain, on the scale where 1 is the disutility of the all-worst state, and 0 is the disutility of full health. They are listed in decreasing order.

Domain	**Disutility**
Physical function	0.688
Depression	0.666
Pain	0.653
Fatigue	0.639
Cognition	0.635
Social roles	0.611
Sleep	0.563

4. Create the multi-attribute disutility function

Using these estimates for the disutility corner states, the global interaction constant, and the single-attribute disutility functions, we calculated the multi-attribute disutility function $\bar{u}_{AW}$ on the all-worst to full health scale with equation (3.3).

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

The mean utility value of dead on the all-worst to full health scale is 0.021. Using that value and the function $\bar{u}_{AW}(\Theta)$ (equation (3.3)) from step (4), the **PROMIS-Preference (PROPr) multi-attribute scoring function** is given by $u(\Theta)$ in equation (3.4). After rescaling so that dead has a utility of 0, the all-worst state has a utility of -0.022, which is the lowest possible utility of the PROPr multi-attribute scoring function. By construction, 1 is the highest possible score.

6. Perform sensitivity analyses

We repeated the process of steps 1-5 for each of the 4 samples, created with the alternative exclusion rules. The 5 resulting multi-attribute scoring functions were then applied to the health states that all 983 participants described in their responses to the PROMIS-29 plus Cognition 4-item short form. We found that these utilities were highly correlated with one another (linear correlations all ≥ 0.98, with p-values all < 0.001). The disutility corner state values defined by the alternative cases were correlated above 0.90 with the values defined by the core sample (p-values all <0.01) except for

Case (iii) (removal of out-of-bounds responses) where the correlation was 0.76 (p-value = 0.046).

3.4 Discussion

This paper describes the development of the PROMIS-Preference (PROPr) scoring system, which generates 7 single-attribute scoring functions and a multiplicative multi-attribute summary scoring function for 7 PROMIS domains: Cognitive Function – Abilities; Depression; Fatigue; Pain – Interference; Physical Function; Sleep Disturbance; and Ability to Participate in Social Roles and Activities. Each single-attribute scoring function can be used on its own to compare groups or to track a group through time, while the multi-attribute function provides a summary score that combines all 7 domains. For the multi-attribute scoring function, 0 is the utility of dead and 1 the utility of full health. For the single-attribute scoring functions, 0 corresponds to the utility of the state with the unhealthiest level on a domain and the healthiest levels on all other domains (i.e., the disutility corner state of that domain), and 1 corresponds to the utility of full health.

The 7 single-attribute functions suggest utility is a nonlinear function of the PROMIS scores (Figure 3.6). Furthermore, the type of nonlinearity, reflected in the curvature of the single-attribute functions, varies by domain, even though the states on each domain cover a similar range of functional capacity. For example, the single-attribute function for social roles is almost constant in the mid-range of theta, while the single-attribute function for pain has close to constant curvature.

The disutility corner state values are all similar (Table 3.3). A methodological interpretation of this result is that participants had enough difficulty with the SG task to blur distinctions among these states. A substantive interpretation of this result is that participants believe that the disutility corner states would affect their overall HRQL similarly. That similarity could reflect the success of our attempt to choose the most important domains and to represent each with values that spanned its range (Table 3.1). Moreover, the range of health states is so large and disutility corner states describe such low levels of functioning that the utilities assigned to them plausibly could be very close.

Several limitations should be considered when interpreting the findings of this study. First, only 57% of invited participants completed the entire survey. Participants were recruited from an online panel; while the representativeness of online panels for the general population is a perennial concern (Tourangeau & Plewes, 2013), they provide affordable access to a demographically diverse sample. The survey company used in this study released invitations in waves to ensure the final sample's demographic characteristics matched the 2010 U.S. Census. Thus, the final sample's demographic characteristics do not impact the generalizability of the results by design – although it is an open question whether the preferences of those who did not complete the survey are systematically different from those who share their demographic characteristics and did complete the survey.

Second, we excluded subsets of participants based on their response behavior. Our core sample excluded participants who took less than 15 minutes and responses in the top and bottom 5% of the utility distribution for each health state. These exclusion criteria sought to balance external validity (having a more representative sample) and internal validity (having better quality responses). Sensitivity analyses found that the multi-attribute scoring function for this sample produced similar utility estimates for participants' self-described health states as the scoring functions created with 4 other samples, based on other exclusion criteria. Nonetheless, further research on exclusion criteria is warranted, especially to identify cases where trimmed responses reflect thoughtful, but uncommon valuations (Engel et al., 2016; Wittenberg & Prosser, 2011).

One feature of the current procedure is asking participants whether they prefer the all-worst state or dead, and then using the worse of the 2 as the origin for some valuation tasks. In order to place all responses on a common scale, some calculations for the 2 groups were done separately, and then combined, weighting by group size. An alternative approach is to transform each participant's valuations individually. We did not use that approach because it would exclude more responses and impose a scale transformation that participants might not endorse (Dewitt et al., 2017). Future research might seek to provide empirical guidance of what kinds of comparisons participants believe are reasonable.

The PROMIS measures provide greater granularity than other descriptive systems (Fryback, 2010). PROPr inherits this granularity when producing utilities. Previous descriptive systems have shown substantial ceiling effects in the general population or floor effects in unhealthy populations (Fryback, 2010). The PROMIS item banks do not have ceiling or floor effects (Reeve et al., 2007), and the range of the PROMIS domains included in PROPr were chosen to avoid these effects as well; future work should verify this assumption. By using health states that represent the range of PROMIS scores, PROPr should be applicable to studies using these domains, whatever their specific design. An important future evaluation of the PROPr score will be to compare the preference scores derived from surveys composed of different sets of PROMIS items for the PROPr health domains.

Another benefit of PROPr is that it can quantify the statistical uncertainty underlying its measurements, a task which has been difficult in previous systems (Chan, Xie, Willan, & Pullenayegum, 2016). Based on the number and type of PROMIS questions answered by an individual, his or her measurement on the underlying health domain(s) under consideration has known precision, derived from IRT. This can be propagated into the PROPr score, providing the quantification of uncertainty needed for incorporating PROPr into stochastic models.

Standardized code is available, at no cost, for users of R and SAS for calculating PROPr scores (Hanmer & Dewitt, 2017). In the spirit of PROMIS, PROPr seeks to make health valuation as easy-to-use as possible for researchers, clinicians, and policy-makers.

Based on this study, a generic societal preference-based scoring system based on 7 selected PROMIS health domains is now available. Clinical, population, and health services research studies that include short-form or computerized adaptive test scales covering these health domains can now use the PROPr scoring algorithm to estimate preference-based scores. This flexibility of measurement effectively links modern measurement based scores with those derived from utility theory, allowing for a more unified assessment of health outcomes for clinical and health policy studies.

Part III

Exclusion

Introduction

Utility-based measures of health-related quality of life (HRQL) provide quantitative estimates of preferences for health. They are used as metrics for compactly representing the value of health states in cost-effectiveness and cost-utility analyses, decision analyses, clinical trials, population health studies and population health management (Wilson & Cleary, 1995). Many of these applications use estimates meant to represent *societal preferences*, produced by eliciting individuals' utilities and then aggregating them (Neumann, Goldie, & Weinstein, 2000).

Dewitt, Fischhoff, Davis, and Hanmer (2017) (Part I) detail the ethical issues underlying the choice of aggregation procedure. Part II describes the creation of a generic societal preference-based measure of HRQL, using conventional procedures including the aggregation rule, namely, relying on the mean utility (Feeny et al., 2002; Hanmer et al., 2015; Hanmer & Dewitt, 2017). That measure of HRQL is called the Patient-Reported Outcomes Measurement Information System (PROMIS®) Preference scoring system (the PROPr scoring system). Here, we examine one essential aspect of the conventions embodied in those procedures – the selection of the sample of preferences that are aggregated to create societal utilities. Specifically, we consider the *exclusion criteria*, whereby some responses are disqualified, based on some property of their content (e.g., unusually high values) or the response process (e.g., unusually low survey completion time) deemed to indicate that the responses do not faithfully represent participants' preferences.

In a recent review, Engel et al. (2016) found large variation in the exclusion criteria used in 76 analyses. In this study, we offer a systematic approach to characterizing these conventions and their practical and ethical implications for analyses relying on them. We begin with a theoretical discussion, building on others in the literature (Devlin, Hansen, Kind, & Williams, 2003; Devlin, Hansen, & Selai, 2004; Engel et al., 2016; Lamers, Stalmeier, Krabbe, & Busschbach, 2006; Lancsar & Louviere, 2006) and adding perspectives from decision-making research (Edwards, 1954; Fischhoff, 2010; Fischhoff & Kadvany, 2011). We then re-analyze the PROPr data set to reveal the implications of exclusion criteria considered in the review. These results address the call by Engel et al. (2016) for better understanding of the implications of exclusion criteria and for better methods, reducing the potential need for exclusion.

We use two methods from quantitative and mathematical psychology to examine the empirical relationships among possible exclusion criteria and their implications for analyses applying them: *multidimensional scaling* and *regression analysis*. Multidimensional scaling (MDS) (Baird & Noma, 1978; Borg & Groenen, 2005; Shepard, 1980) allows us to compare exclusion criteria in terms of their agreement about which responses to exclude (and which to include). That comparison can reveal the extent to which conceptually independent criteria produce empirically similar results (Fischhoff, Slovic, Lichtenstein, Read, & Combs, 1978), revealing, for example, cases where criteria that are ethically acceptable in principle achieve results that are ethically unacceptable in practice (e.g., by virtue of disproportionately excluding members of one group), or vice versa. Regression analysis allows us to estimate the differences between the preferences of those excluded and those not excluded by a criterion. That analysis reveals the practical and ethical effects of exclusions on the responses used in analyses that base decisions on societal HRQL values. Because utility scores are double-bounded variables, ranging from 0 to 1 (inclusive), we use *beta regression*, a recently developed approach (Cribari-Neto & Zeileis, 2010; Smithson & Verkuilen, 2006) that imposes weaker theoretical constraints than the regression methods used in Part II.

This part of the dissertation is structured as follows. First, we provide an overview of the PROPr data and exclusion criteria, repeating some material from earlier chapters for the sake of completeness. Second, we introduce MDS, apply it to the PROPr data, and discuss the implications of the results for the validity of criteria as detectors of low-quality data. Third, we introduce beta regression, apply it to the PROPr data, and discuss the implications of those results. Finally, we address the implications of these analyses for using the PROPr data in policy analyses and for designing future preference elicitation studies. An appendix contains detailed methodological results and discussions not presented in the main text.

4

PROPr & Exclusion Criteria Overview

4.1 The online survey for the Patient-Reported Outcomes Measurement Information System (PROMIS) Preference (PROPr) Scoring System

The PROPr scoring system focuses on a subset of 7 PROMIS domains, chosen to span the overall space of generic health states, so that all health states of interest (to the public or researchers) are either in the space or correlated with states that are. Hanmer and Dewitt (2017) describe the procedure that identified those PROMIS health domains: 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*).

Each domain in PROPr (and PROMIS) is treated as a continuous latent construct, called *theta* in Item Response Theory (IRT) terminology. The domains are unbounded in both directions. PROMIS uses a *T-score* metric, constructed so that the population mean is 50, with a standard deviation of 10. PROPr uses a *z-score* metric, a linear transformation of the T-score, such that the mean is 0 and the standard deviation is 1. PROMIS measurements rarely fall outside the range -4 to 4 on theta (equivalent to a T-score of 10-90).

A functional capacity on a domain is called a *level* of theta. Individuals are characterized by their health state, equal to their theta scores (levels) for the seven domains. These states are inputs to the *PROPr scoring system*, which produces seven single-attribute scores (utilities) and one multi-attribute summary score. Those scores were derived by eliciting preferences for six or seven levels of theta, characterized by qualitative descriptions. Table 3.1 shows the theta values used. The PROPr technical report and Part II describe the procedures used to elicit the utilities for these theta values and for estimating continuous utility functions spanning the range of possible values (Hanmer & Dewitt, 2017).

We collected the data online, with an instrument administered by ICF (`https://www.icf.com/services/research-and-evaluation`) and SurveyNow (`http://www.surveynowapp.com/`). The survey outline is available in the PROPr technical report (Hanmer & Dewitt, 2017). The present analyses focus on the subset of the survey's tasks used to define the exclusion criteria and that define participant characteristics used as regressors in the modeling section of this part of the dissertation:

- Two preference elicitation tasks: one for a set of health states described by changes on a single PROMIS domain (e.g., cognition), the other for a set of health states described by changes on multiple PROMIS domains (e.g., the state with the worst level on each domain).
- An assessment of participants' numeracy, using the short form of the Subjective Numeracy Scale (Fagerlin et al., 2007; McNaughton et al., 2015).
- Participants' self-assessed understanding of the survey.
- The PROMIS-29 inventory and the PROMIS Cognition 4-item short-form (Gershon et al., 2010; PROMIS, 2015), used to characterize participants' health states in PROMIS terms.

As compensation, participants could choose from 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 the authors received them.

In the preference elicitation task, participants valued two sets of states, as described below, first using a visual analogue scale (VAS) and then a standard gamble (SG) (Gafni, 1994; Torrance et al., 2001). We used the VAS mainly to introduce the health states valued in the SG version of the task (Torrance et al., 2001). SG responses were used for the PROPr scoring system because of their grounding in expected utility theory (Keeney & Raiffa, 2003; von Neumann & Morgenstern, 1944), hence are our focus. As discussed later, the present approach could be used to compare empirical properties of SG and VAS responses (Engel et al., 2016).

The VAS had a 0-100 scale (sometimes called a Feeling Thermometer), where 0 represents the value of a *lowest health state* (see below) and 100 represents the value of the highest, called *full health*

– the state with the highest functional capacity on all domains. Participants evaluated an intermediate health state corresponding to a theta value in Table 3.1.

The SG presented one of the intermediate health states, offering a choice between (a) having it with certainty and (b) a lottery with probability p of full health and $(1 - p)$ for the lowest state. The SG procedure offered a series of such choices, varying the probability until the participant was indifferent between the two options. Following utility theory assumptions, this probability, $\hat{p}$, is the utility of the intermediate state.

(a) Set (i)

Each participant was randomly assigned to one of the 7 health domains (e.g., cognitive functioning). Each evaluated the 6 or 7 states corresponding to the theta values for that health domain, each described in verbal terms (as in Part II, Figure 3.2). The lowest health state in each set of valuations was the *disutility corner state* for that domain (corresponding to the worst possible level in that domain and perfect health on all others).

(b) Set (ii)

Recognizing that participants may view some states as worse than dead (Feeny et al., 2002), we asked them whether they preferred the state of being dead or the state with the unhealthiest level on each of the seven domains (the *all-worst state*). We used the one *not* chosen as the bottom state for participant's valuations of states in this set. Participants valued the disutility corner state belonging to the domain they valued in set (i). They also valued two other states, randomly selected from the set of other domain disutility corner states and three *marker states* (chosen to span the health state space) (Feeny et al., 2002). Finally, participants valued dead or the all-worst state, whichever had been selected as better.

The next section identifies a set of exclusion criteria that are then applied to these responses.

4.2 Selecting criteria for study

We considered exclusion criteria that 1) were used by the community of researchers involved with health state valuation and 2) reflected the range of theoretical arguments (Engel et al., 2016; Wittenberg & Prosser, 2011).

Table 4.1 lists common exclusion criteria in health state valuation studies, which can be divided into *preference-based* and *non-preference-based*. Preference-based exclusion criteria remove responses based on features of their content that cast doubt on their quality, either because they appear to indicate inattentive responses, protest responses, or contradict the theory of such preferences (Wittenberg & Prosser, 2011). Non-preference-based exclusion criteria remove responses based on features of the process producing them that cast doubt on their quality. An example of a preference-based criterion is removing participants who provide the same utility for all health states, under the assumption that they were either ignoring the task or communicating something other than their utilities, such as a protest response. Two examples of non-preference based criteria are excluding the responses of participants who complete the survey too quickly or whose score on a numeracy scale falls below a threshold, interpreted as indicating that they could not have grasped the question.

As mentioned, criteria reflecting different rationales might end up excluding similar individuals. MDS examines that question. The regression analyses that follow it ask how application of any criterion would affect societal utilities.

Table 4.1: *List of common exclusion criteria.* 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. (2016), and adds others that are based on the two categories used in that table – "lack of understanding/engagement" and "model requirements" – including criteria from specific studies (e.g., (Devlin et al., 2003; Feeny et al., 2002)). Unshaded rows indicate *preference-based criteria*, shaded rows indicate *non-preference-based criteria*.

Exclusion criterion	**Description of criterion**	**Rationale for exclusion**	**Notes**
Providing 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.	
Using 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 three.
Violates 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. 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 assignments.
Dead or the all-worst state better than all or some health states	Assigned a utility to dead or the all-worst state as 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 violating dominance. Makes certain modeling tasks impossible or uninterpretable, depending on the modeling strategy.	
Valuing a lower anchor (e.g., dead) the same as 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.	
Did not value dead, all-worst state, or full health	Missing data for one of these three states.	Makes certain modeling tasks impossible or uninterpretable, depending on the modeling strategy.	
Valuations are 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 are 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".
Low numeracy	Scored too low on a numeracy scale.	Low numeracy implies the participant could not understand the elicitation task.	
Low self-assessed understanding	Rated themselves too low on a rating of their own ability to perform the survey.	Participant is admitting inability to use the task to communicate their preferences.	
Completed survey too quickly	Completed the survey below a minimum time threshold.	Completing the survey too quickly implies careless responses.	

5

Multidimensional Scaling

Multidimensional scaling (MDS) provides a holistic picture of the similarities (and differences) between objects. Originating in psychophysics (Baird & Noma, 1978), MDS reduces pairwise comparisons among a set of objects to a low-dimensional graphical space in which the *distance* between objects is taken as a proxy for their similarity (Borg & Groenen, 2005). Here, the objects are the exclusion criteria and the similarity metric is a measure of their agreement on which responses to exclude and which to include.

A common example of MDS lets the objects be cities, and the proximity index the true distance between them. A MDS solution produces a 2-dimensional scaling (map) out of those pairwise distances. For example, Dewitt et al. (2015) use MDS to produce a geometric representation from judgments of the similarity of weather scenes in terms of their tornadic potential. The proximity index was the proportion of participants who selected one picture as more tornadic, with proportions closer to 50% taken as indicating greater similarity.

Here, we treat each exclusion criterion as a *binary classifier*, either excluding or including (i.e., not excluding) a response. The similarity of two binary classifiers can be summarized in a *confusion matrix* like that in Table 5.1.

There are many ways to extract a proximity index from a confusion matrix (Borg, Groenen, & Mair, 2012; Gower & Legendre, 1986; Warrens, 2008). For example, one could divide the number of

Table 5.1: *Confusion matrix*. A confusion matrix (or *2-by-2 contingency table*) shows the possible outcomes of two 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 one 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.

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

joint exclusions (entry a in Table 5.1) by the total number of cases:

$$\frac{a}{a+b+c+d}.$$

Or, one could consider both the joint exclusions and inclusions:

$$\frac{a+d}{a+b+c+d}.$$

The first of these indices is called the *simple matching coefficient*, or *S2* (Gower & Legendre, 1986; Warrens, 2008). The second is the observed *proportion of agreement*, or *S4*. The first (S2) treats two exclusion criteria that exclude the same 10 people (i.e., $a = 10$) as less similar than two criteria that exclude the same 100 ($a = 100$), with the same overall number of cases. S4 for these two cases would reflect their agreement on exclusions (a) *and* inclusions (d).

There is no way to capture every aspect of the confusion matrix in a single index (Gower & Legendre, 1986; Warrens, 2008). We chose the following index, which incorporates each cell in the confusion matrix:

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

Known as *S14*, the *phi* (ϕ) *coefficient*, and the *Matthews Correlation Coefficient* (Gower & Legendre, 1986; Matthews, 1975; Warrens, 2008; Yule, 1912), it is also the Pearson correlation for two binary variables. The Appendix contains more detail on the choice of proximity index.

Given n exclusion criteria, the basic input to MDS is the n-by-n matrix p, whose (i, j)-th entry

p_{ij} is the phi value for the proximity of exclusion criteria i and j. The diagonal of the matrix is 1 (i.e., $p_{ii} = 1$ for all i), and the matrix is symmetric (i.e., $p_{ij} = p_{ji}$ for all i and j), by construction.

MDS is used to reveal relationships that accommodate the interdependencies among a set of variables. Phi has limited range, relative to the Pearson correlation for continuous variables, hence should be interpreted in relative, rather than absolute values – for those accustomed to the Pearson (see the Appendix). Proximal objects in the resulting space are similar in the terms of the construct defined by the metric. Here, such clusters indicate criteria that exclude and include similar responses, even if they express different principles. Objects distant from one another have different exclusion and inclusion patterns, even if they express seemingly similar principles. In these ways, MDS configurations can reveal unintuitive relationships or affirm expected ones.

MDS algorithms produce m-dimensional configurations or plots (where $m < n$), in which the distance between objects (here, the exclusion criteria) best approximate the values in the matrix (here, the phi values), as measured by a goodness-of-fit (or *stress*) value. For interpretability sake, we let $m = 2$. Higher-dimensional configurations are always possible, and necessarily provide a better fit because of the extra degrees of freedom they provide to the MDS algorithm (see the Appendix). However, they are more difficult to interpret. For that reason, most applications use two dimensions (Borg & Groenen, 2005). We do that here as well, considering additional dimensions as sensitivity analyses. The Appendix shows *Shepard plot* assessments of how well the MDS has represented the phi values as distances.

5.1 Choosing exclusion criteria for MDS

MDS begins by deciding which objects to include. We sought exclusion criteria that represent diverse principles, spanning the space of those used in the research. Examining the exclusion criteria identified by Engel et al. (2016) revealed that many express the same rationale, but vary in their stringency, such that the sets of excluded responses are *nested*. For example, criteria might exclude the highest 1% or 5% of responses. As the nested criteria are necessarily proximal, MDS can achieve goodness-of-fit by placing them close to one another, revealing little, while neglecting relationships among criteria that might differ. Thus, we represent each such set by one criterion (with one exception, described below), selecting the most stringent of the set, so as to capture the shared logic most clearly.

5.2 MDS sensitivity analysis

To test the robustness of our results, we perform the following sensitivity analyses:

a) *Dimensionality*: We compare the focal 2-dimensional MDS solution with a 3-dimensional one, asking whether qualitatively similar patterns emerge.

b) *MDS jackknife*: We remove each criterion and repeat the analysis (de Leeuw & Meulman, 1986).

c) *MDS algorithm*: We use MDS algorithms other than *ordinal* MDS, which was chosen for our focal analyses because it makes the fewest assumptions.

d) *Clustering*: We apply *k-means* clustering to look for structure we might have missed (Bishop, 2006). *K*-means provides a computational approach to revealing structure in an MDS configuration, in which an algorithm examines the *spatial* properties of the configuration just like a researcher does with the human eye.

The Appendix describes these procedures and the results from applying them. Briefly, none reveal patterns substantively different than those in the primary analysis.

5.3 Description of the PROPr dataset

The PROPr online survey was completed by 1,164 participants. Table 3.2 in Part II displays the sample's demographic characteristics, which largely match the U.S. 2010 Census. The sample was slightly older, more educated, reported higher income, and had a larger proportion of White individuals than the U.S. 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%.

The preference-based exclusion criteria reflect participants' preferences on the single-domain valuation tasks using the SG (set (i) above), unless otherwise indicated.

5.4 Results of selecting criteria for scaling

We began the selection process with 19 criteria (Table 5.2), chosen to implement the rationales listed in Table 4.1. We reduced that set to the 10 *core criteria* in Table 5.3 by choosing the most stringent implementation of the nested criteria. For example, we represented the general principle of "excluded for violating dominance" with the specific example of excluding participants for even one violation, of any size (*violates-SG* in Table 5.3).

The one exception to this selection process is that we included both the *low-range* and *no-variance* criteria. 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 another criterion: someone excluded by *no-variance* cannot also be excluded by *violates-SG*, and vice versa. That is because giving the same utility for every health state does not violate dominance. However, a participant who violates dominance must have valued two states differently (and in the "wrong" way), hence cannot be excluded by *no-variance*. We explain the usefulness of this comparison below.

Table 5.4 shows how many PROPr participants are excluded by each criterion from Table 5.3, which ranges from 7.8% (*numeracy*) to 84.7% (*violates-VAS*). Table 5.5 shows that breakdown by domain. Note that some criteria, such as *numeracy*, remove entire participants (and all their responses), while others, such as *lower-tail*, treat each of a participant's responses separately.

5.5 MDS results

Table 5.6 shows the 10-by-10 matrix of phi values correlating exclusion criteria in the core set (Table 5.3). Each criterion is a row and a column, with the (i, j)-th entry equal to the phi value for exclusion criteria i and j. As phi ignores the order of the criteria, the matrix is symmetric. The greatest agreement (0.98) is for (*no-variance*, *low-range*); the greatest disagreement (−0.56) is for (*no-variance*, *violates-SG*).

The matrix in Table 5.6 is the input to an MDS algorithm. We used the *ordinal* MDS algorithm because of its flexibility and minimal assumptions (Borg & Groenen, 2005). (See the Appendix for details.) We use the **smacof** package for the statistical software R (de Leeuw & Mair, 2009). As no algorithm can identify the global optimum configuration analytically, we ran hundreds of iterations, choosing the one with the best fit (the lowest stress).

Figure 5.1 shows the 2-dimensional MDS solution for the core set of exclusion criteria. Although it is common to interpret the dimensions in an MDS configuration (Baird & Noma, 1978;

Table 5.2: *Methods of implementing exclusion criteria in PROPr.* Examples of how to implement the exclusion criteria from Table 4.1 using the PROPr data. Not every criterion from Table 4.1 is represented, because some would exclude no one by virtue of the design of the PROPr survey (e.g., there is no missing data). Unless otherwise indicated, valuations refer to the valuations of the single-attribute states (i.e., set (i) described in Part II and the first part of Part III). Unshaded rows indicate preference-based criteria, shaded rows indicate non-preference-based criteria.

Exclusion criterion	**Requirements for exclusion**
Violates dominance on the SG	A participant, using the standard gamble (SG), violates dominance at least once.
Violates dominance on the SG, more than twice	A participant, using the standard gamble (SG), violates dominance at least twice.
Violates dominance on the SG by more than 10% of the scale	A participant, using the standard gamble (SG), is considered to have violated dominance only if they do so by a difference of more than 0.1 on the utility scale.
Violates dominance on the SG by more than 10% of the scale, more than twice	A participant, using the standard gamble (SG), is considered to have violated dominance only if they do so by a difference of more than 0.1 on the utility scale, more than twice.
Violates dominance on the VAS	A participant, using the visual analog scale (VAS), violates dominance at least once.
Violates dominance on the VAS, more than twice	A participant, using the visual analog scale (VAS), violates dominance at least twice.
Violates dominance on the VAS by more than 10% of the scale	A participant, using the visual analog scale (VAS), is considered to have violated dominance only if they do so by a difference of more than 10 on the 0-100 VAS scale.
Violates dominance on the VAS by more than 10% of the scale, more than twice	A participant, using the standard gamble (SG), is considered to have violated dominance only if they do so by a difference of more than 10 on the 0-100 VAS scale.
Valued the all-worst state or dead as the same or better than full health.	A participant is excluded if they rated the all-worst state or dead as the same or better than full health, using the standard gamble (SG).
Used less than 10% of the utility scale	A participant is excluded if their valuations, using the standard gamble (SG), represent less than 10% of the range of the utility scale.
Provided the same response to every SG	A participant is excluded if they valued every state the same, using the standard gamble (SG).
In the top 5% of responses for an SG	A response is excluded if it falls in the bottom 5% of responses for that health state, using the standard gamble (SG).
In the bottom 5% of responses for an SG	A response is excluded if it falls in the bottom 5% of responses for that health state, using the standard gamble (SG).
Score on the Subjective Numeracy Scale of less than 2.5	A participant is excluded if they scored less than 2.5 on the short form of the Subjective Numeracy Scale (McNaughton, Cavanaugh, Kripalani, Rothman, & Wallston, 2015).
Self-assessed understanding equal to 1, on a scale of 1 = "Not at all" to 5 = "Very much"	A participant is excluded if they rated themselves a "1" on the self-assessed understanding question, which occurred after the preference elicitations.
Self-assessed understanding equal to 1 or 2, on a scale of 1 = "Not at all" to 5 = "Very much"	A participant is excluded if they rated themselves a "1" or a "2" on the self-assessed understanding question, which occurred after the preference elicitations.
15-minute time threshold	A participant is excluded if they completed the PROPr survey in under 15 minutes.

Table 5.3: *Core exclusion criteria for study*. The subset from Table 5.2 of criteria studied in Part III, included in the core MDS analysis. Their shorthand names are included in the parentheses, and will be used throughout Part III. Note that, following Table 4.1, *upper-tail* and *lower-tail* combine to form *10% trimming*. Unshaded rows indicate preference-based criteria, shaded rows indicate non-preference-based criteria.

Exclusion criteria (*shorthand*)
Violates dominance on the SG (***violates-SG***)
Violates dominance on the VAS (***violates-VAS***)
Valued the all-worst state or dead as the same or better than full health (***dead-all-worst***)
Provided the same response to every SG (***no-variance***)
Used at less than 10% of the utility scale (***low-range***)
In the top 5% of responses for an SG (***upper-tail***)
In the bottom 5% of responses for an SG (***lower-tail***)
Score on the Subjective Numeracy Scale of less than 2.5 (***numeracy***)
Self-assessed understanding equal to 1 or 2, on a scale of 1 = “Not at all” to 5 = “Very much” (***understanding***)
15-minute time threshold (***time***)

Table 5.4: *Number of participants flagged by each criterion*. The number of participants in the PROPr data flagged by the each criterion from Table 5.3. The total number of participants in the sample is 1,164. Unshaded rows indicate preference-based criteria, shaded rows indicate non-preference-based criteria.

Exclusion criterion	**Number excluded (total sample** n = 1164**)**
Violates dominance on the SG (***violates-SG***)	833 (71.6%)
Violates 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 at less than 10% of the utility scale (***low-range***)	142 (12.2%)
In the top 5% of responses for an SG (***upper-tail***)	117 (10.1%)
In the bottom 5% of responses for an SG (***lower-tail***)	214 (18.4%)
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-minute time threshold (***time***)	181 (15.6%)

Table 5.5: *Proportion of participants flagged by each criterion, per domain*. The proportion of participants in the PROPr data flagged by each criterion from Table 5.3, broken down by domain. Each column label is one of the seven PROPr domains, with the number of participants assigned to value that domain in parentheses, with the a total of sum 1,164. Each row is labeled by one of the core criteria (Table 5.3), with the percentage of all participants excluded by each criterion in parentheses (those values are from Table 5.4). Unshaded rows indicate preference-based criteria, shaded rows indicate non-preference-based criteria.

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)
understanding (14.3%)	17.5%	10.8%	14.5%	14.5%	15.1%	12%	15%
time (15.6%)	12%	17.4%	16.9%	16.3%	17.5%	13.9%	15%
numeracy (7.8%)	8.4%	9.0%	9.0%	12.7%	4.2%	5.4%	6.0%
no-variance (11.8%)	12.0%	6.6%	14.5%	15.1%	9.0%	13.3%	12.0%
low-range (12.2%)	12.7%	7.2%	15.1%	15.7%	9.6%	13.3%	12.0%
lower-tail (18.4%)	16.9%	21.0%	19.3%	19.9%	20.5%	17.5%	13.8%
upper-tail (10.1%)	9.6%	9.0%	10.2%	12.7%	9.6%	10.2%	9.0%
violates-SG (71.6%)	72.3%	74.9%	72.3%	71.1%	77.7%	64.5%	68.3%
dead-all-worst (28.0%)	28.9%	25.7%	26.5%	30.7%	24.7%	28.9%	30.5%
violates-VAS (84.7%)	85.5%	80.8%	88.6%	80.1%	89.8%	85.5%	82.6%

Table 5.6: *Proximity matrix.* The proximity matrix for the core criteria: each entry is the phi (phi) value of the exclusion criteria in the row and column. The shaded row and column titles indicate the non-preference-based criteria.

	understanding	*time*	*numeracy*	*no-variance*	*low-range*	*lower-tail*	*upper-tail*	*violates-SG*	*dead-all-worst*	*violates-VAS*
understanding	1									
time	0.04	1								
numeracy	0.06	0.02	1							
no-variance	0.10	0.09	0.05	1						
low-range	0.10	0.08	0.06	0.98	1					
lower-tail	0.04	0.02	-0.01	-0.13	-0.13	1				
upper-tail	-0.005	-0.002	0.02	0.01	0.01	-0.14	1			
violates-SG	0.02	0.03	-0.04	-0.58	-0.56	0.11	0.01	1		
dead-all-worst	0.06	0.13	0.06	0.25	0.24	-0.03	0.12	-0.08	1	
violates-VAS	0.02	0.02	0.03	-0.05	-0.05	0.04	0.02	0.09	0.03	1

Dewitt et al., 2015), any figure that is a translation, rotation, flip, or any combination of those transformations (collectively called *Procrustean transformations*) is an equivalent solution (Borg & Groenen, 2005). As a result, one can interpret any direction in the figure, and not just the two axes. However, regions of the configuration, with clusters of objects, are invariant under Procrustean transformations. As a result, our discussion of the results (next section) focuses on them.

5.6 Discussion of MDS results

5.6.1 Non-preference-based criteria

The three non-preference-based criteria – *numeracy*, *time*, and *understanding* – are based on three rationales: low-numeracy implies difficulty understanding the SG, completing the survey quickly implies inattentiveness, and admitting to not understanding the tasks implies poor quality data for those reasons or others. Their relative proximity in Figure 5.1 suggests that these rationales are related, potential expressions of some underlying construct of inability to use the SG to express one's preferences.

5.6.2 Preference-based criteria

Violates-SG and *violates-VAS*

Violates-SG and *violates-VAS* are preference-based criteria (Table 4.1 and Table 5.3), reflecting the *structure* of participants' responses. Each entails rating a state with lower functional capacity as better than a state with higher functional capacity. Both criteria exclude many participants (Table 5.4). The fact that the VAS excluded more may reflect its greater precision: the VAS allows increments of 0.01, whereas the SG only offers probability (hence utility) increments of 0.05.

More generally, the axis in the figure with *violates-SG* at one end and *no-variance* at the other might represent the precision in participants' responses. As mentioned, *violates-SG* and *no-variance* must have a large distance between them because they cannot exclude the same participants.[1] The

[1] *Violates-VAS* is the only criterion based on responses using the VAS. Thus, although the VAS has more precision than the SG (more decimal points), *violates-VAS* does not have the same relationship with *no-variance* as does *violates-SG*, because *no-variance* is defined using the SG responses. In contrast, the other preference-based criteria are all defined by the SG, and thus their position relative to *violates-SG* and *no-variance* are based on the responses to the same part of

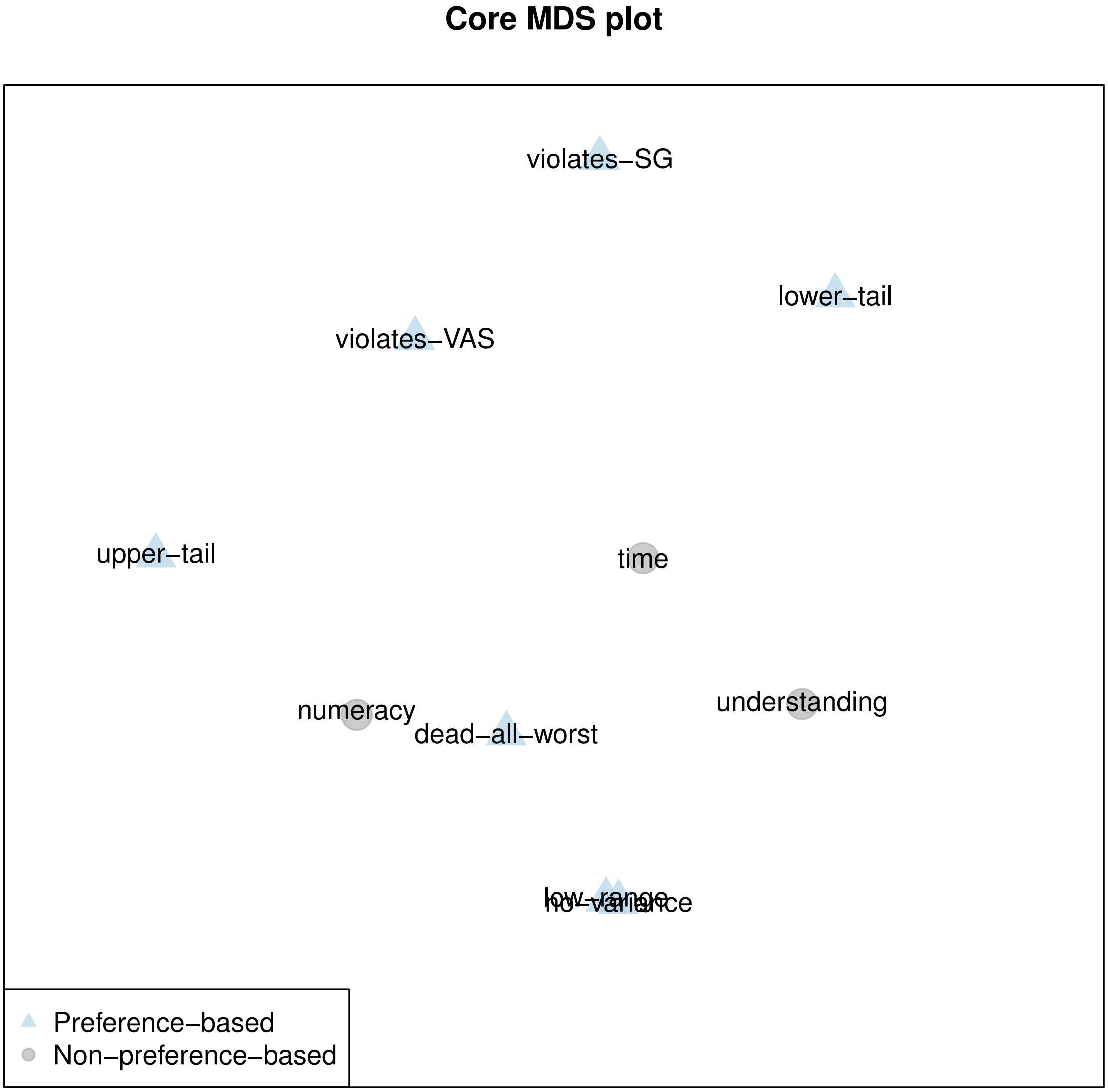


Figure 5.1: *Core 2-dimensional MDS configuration.* The core 2-dimensional MDS configuration.

figure shows that *no-variance* and *low-range* are close to one another, which is expected given that they are nested criteria (constant responses (*no-variance*) are a subset of nearly-constant responses (*low-range*)). Both are close to the non-preference-based criteria (*numeracy*, *understanding*, *time*), which appear to reflect low quality responses (rather than true indifference). The fact that *violates-SG* is distant in the figure suggests that it might remove many participants who are trying to express well-considered utilities but cannot do so without violating dominance, perhaps reflecting problems with the task experienced by relatively hard-working participants who have something to say. If so, then perhaps only participants flagged by *violates-SG* are providing valid responses.

In research practice, *violates-SG* is relaxed to allow participants some number of illogical responses, up to 11 in one study (Engel et al., 2016). Box 1, below, suggests some of participants' challenges in using this response mode, which might justify such practices. Such practices suggest that researchers who prefer the SG on theoretical grounds may be forced to accept imperfect data from participants who cannot meet its cognitive demands. The VAS rating-scale task appears less demanding than considering hypothetical gambles among health states. As mentioned, its higher rate of violations of dominance may reflect the more precise response mode.

***Upper-tail* and *lower-tail* (collectively *10% trimming*)**

These two criteria were combined in constructing the PROPr scoring system (Part II) and the widely-used Health Utilities Index Mark 3 (Feeny et al., 2002). However, the distance between *upper-tail* and *lower-tail* (Figure 5.1) indicates that they exclude quite distinct groups of participants. Participants who give unusually low responses are most similar to participants removed for violations of dominance, which reflects the *structure* of participants' preferences, and not their absolute value, as implied by *lower-tail*. The similarity of *lower-tail* and *violates-SG* might mean that, unless participants give a very low utility only on the worst state they rate, the low utility that caused their exclusion by *lower-tail* will cause a violation of dominance unless *all* their other utilities are lower still. Based on the distributions of responses for almost every state in every domain, a

the PROPr survey. If we were to repeat the analysis with *no-variance* defined on the VAS responses, then it would be maximally far apart in the resulting MDS configuration, just as *violates-SG* and *no-variance* are in Figure 5.1. The smaller distance between *violates-VAS* and *no-variance*, as well as the non-overlap between *violates-VAS* and *violates-SG*, indicates that participants did not respond the same way using the two techniques.

participant would have to give a utility of 0 to be excluded by the *lower-tail* criterion.[2] Doing so would, then, lead to violating dominance – unless the participant provided all 0s, which would mean being excluded by *no-variance*.

In contrast, the *upper-tail* criterion is most similar to the *numeracy* criterion, which reflects the ability to understand numerical information, including probability statements. That similarity suggests that participants with low numeracy had difficulty with the SG task, which uses probability. Box 1 offers a speculative interpretation of how that difficulty may have led to unusually high responses, rather than unusually low ones. Thus, *10% trimming*, which is typically justified as eliminating atypical responses, or "noise," may remove two very different groups of participants: those who are confused by the task and inadvertently produce unusually high utilities, and those who understand the task and deliberately express unusually low utilities.

Box 1. One possible account for high valuations is that the first question in any SG valuation presents a degenerate gamble with a 100% chance of receiving full health, with the other option the usual sure-thing of the intermediate state whose utility is being estimated. The participant can choose the "gamble," the sure-thing, or indifference. Taking the gamble leads to another choice between a (real) gamble and the sure-thing. Taking the sure-thing or expressing indifference completes the task, and is interpreted as a utility estimate of above 1 and 1, respectively. Therefore, making either of those choices leads to a response in the upper tail of the utility distribution. Thus, the mechanics of the procedure could mean that confused or inattentive responses sometimes are recorded as utilities in the upper tail of any distribution of responses. Choosing one of the two options leading to a utility of 1 or greater could be likely among the less numerate, who might have more difficulty than others understanding the probability statements in the SG.

Dead-all-worst

Dead-all-worst captures participants providing a response that is necessarily extreme with respect to other participants' SG values for dead or the all-worst state (i.e., the highest possible responses), as

[2] Of all 48 states across the seven domains (seven states for every domain except sleep, which has six), the 5th quantile is 0 except for one state in each of depression and fatigue, two states in physical functioning, and three in sleep. For all 48 states, the 95th quantile is 1.

well as a response that is a violation of dominance (unless the participant provided constant responses). Its proximity to *numeracy* suggests that the mechanism in Box 1 could explain what causes some participants to be flagged by *dead-all-worst*: an error, made more likely by low numeracy. Its distance from *violates-SG* indicates that those violating dominance by rating, e.g., the all-worst state as equal to or better than full health, are different from participants who violate dominance in other ways. That difference could define one group (*violates-SG*), who are trying to express themselves but produce violations due to the difficulty of the SG response mode, while another (*dead-all-worst*) defines a group who are not communicating their preferences, either because they cannot (e.g., Box 1) or because they are not engaged with the task. Said differently, *dead-all-worst* could be considered the most extreme violation of dominance, and thus be more likely to capture participants not providing their responses than *violates-SG*, which flags participants for any violation, no matter the distance between the states on the underlying construct (theta). We conjecture that the larger the theta-distance of two states ordered the "wrong" way in utility space, the higher probability that violation belongs to a participant who is not engaged with the task or is otherwise unable to use the SG.

Low-range* and *no-variance

Eighty-seven percent of those flagged by *no-variance* had 1 or above 1 as their constant response, which represents 84% of those captured by *low-range*. That could be because, given the SG procedure, two of the three choices on the first screen of the SG would lead to those values, so that participants repeating that choice *for each state* could be trying to get through the task with the fewest clicks possible, be inattentive or confused (Box 1), or providing protest responses. Those providing values below 1 as their (nearly) constant response could be substantively different from their co-participants, or simply have found a different pattern of choices to repeat for each question. There are too few of those participants to know more using the current set of analyses.

5.7 Conclusion

The preference-based criteria (Table 4.1 and Table 5.3) take normative stances on what count as legitimate responses. The non-preference-based criteria posit mechanisms that cannot produce true preferences: *time* claims that working too fast implies inattention; *numeracy* claims that low numeracy makes the SG too difficult; and, *understanding* claims that participants who say that they do not understand the task are admitting that their data is of low quality. The clustering of the three non-preference-based criteria in Figure 5.1 suggests that they pick up related aspects of poor performance that could be independent, *a priori*.

In contrast, the distribution of preference-based criteria suggests that they reflect disparate mechanisms. As we have interpreted them, those mechanisms are sometimes seemingly at odds with the stated rationales for using them (Table 4.1). For example, the large distance between *upper-tail* and *lower-tail*, the former's relative proximity to *numeracy* and the latter's distance from any non-preference-based criteria provide a plausible explanation for the mechanism driving *upper-tail* (Box 1). In contrast, *lower-tail* defines responses as low. However, there is no obvious empirical evidence of why they should be treated as *too* low. Thus, although the two trimming criteria, *upper-tail* and *lower-tail*, are conventionally treated similarly (as parts of *10% trimming*), they are, in fact, quite different. *Lower-tail*'s grouping with *violates-SG* and *violates-VAS* indicates that it has more in common with preference-based criteria defined by how a participant's response on a state relates to that participant's responses on other states, despite the rationale for *lower-tail* considering how a participant's response on a state relates to the responses of other *participants* for *that* state alone.

The present analyses signal that the SG has design problems. As implemented, its first screen provides an easy way to provide a response equal to or above 1. Our analyses suggest some of these are likely chosen because of low numeracy, but it is difficult to parse those from the inattentive or protest responses. The large number of dominance violations, and our evidence that many flagged by *violates-SG* could be *trying* to discriminate between each state suggest it is a noisy tool for mapping the health states (theta) to utilities, if so many cannot perform that mapping while preserving the assumed order of the states. The large number of those flagged by *violates-VAS* implies that providing more gradations in the utility scale is not the answer. Rather, providing the participant in real-time with the ordering implied by their previous and current valuations could provide the feedback necessary to use the SG more accurately. That feedback might induce those engaged with the survey – but who find the SG task difficult – to provide something other than constant responses, while we would expect the inattentive and those protesting the survey not to respond to the feedback. The VAS, combined with that feedback, might be even better, given that it does not require understanding of probability and thus might be less burdensome to the less numerate. Discarding the normative properties of the SG (see Part II) could be worthwhile if doing so allows the inclusion of more responses and more confidence in the exclusions we do implement.

In the next section, we estimate the impacts of the exclusion criteria on the distribution of utilities. That analysis complements our MDS results, as comparing the preferences of those excluded with those not excluded tells us the practical and ethical effects of implementing criteria of varying ability to capture low-quality data.

6

Modeling Utilities as a Function of Exclusion Criteria

MDS revealed the similarities and distinctions among exclusion criteria, with respect to how they classify participants. Our selection process for winnowing Table 4.1 and Table 5.2 to Table 5.3 aimed to represent the variety of rationales for excluding responses. The core MDS configuration, Figure 5.1, showed that only two criteria classified participants in almost *exactly* the same way, the nested criteria of *low-range* and *no-variance*.

This pattern is an empirical result. *A priori*, we could not know if any non-nested criteria would make nearly coincident classifications nor if any would be opposites, with one excluding if and only if the other included (which we did not find). If two criteria classified responses identically, then either could represent the pair in health-related policy analyses, and produce similar conclusions. Absent such agreement, the practical impact of each criterion must be considered separately. In this section, we demonstrate a general method for analyzing the effects of exclusion criteria on the utilities used to represent the sampled population.

Below, we introduce our main tool for analyzing the effect of exclusion criteria on utilities: *beta regression*. Then, we present our models, the results of applying those models to the PROPr data, and their implications for the use of exclusion criteria on societal utilities.

6.1 Introduction to beta regression

By convention, utilities are elicited on scales that are doubly bounded at 0 and 1. Such variables can make conventional regression analysis difficult. They may have substantial and varying skewness. They may have heteroskedasticity, because the variance of a bounded random variable depends on the mean. (A variable with mean value near a boundary must have low variance, whereas one with a middle mean value need not.)

Rather than ignoring such heteroskedasticity and skewness, or trying to remove them with a transformation, *beta regression* includes them in its models (Smithson & Verkuilen, 2006). Beta regression assumes that, conditional on any regressors (i.e., independent variables) – that is, holding the regressors at fixed values – a dependent variable is distributed according to a beta distribution $Beta(\omega, \tau)$. Two *shape* parameters, $\omega > 0$ and $\tau > 0$ define a beta probability density function over $(0, 1)$. 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. (As the beta distribution is bounded, it can become bell-shaped but not normal, which has some probability density across the whole real line.) In general, ω pulls the density towards 1 and τ towards 0, producing skewed distributions when the two parameters are unequal.

The probability density function of a random variable $y \sim Beta(\omega, \tau)$ 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

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

We will focus on the mean and a modification of the variance, which have more direct implications for societal utility estimates than do the shape parameters.

Paolino (2001) was the first to provide the alternative parametrization of the beta that has now become the standard (Cribari-Neto & Zeileis, 2010; Smithson & Verkuilen, 2006; Smithson, Budescu, Broomell, & Por, 2012). If we let $\mu = E(y)$ and $\phi = \omega + \tau$, then $\omega = \mu\phi$ and $\tau = \phi - \mu\phi$. Therefore, $Var(y) = \frac{\mu(1-\mu)}{(\phi+1)}$, 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. (Note that, although ϕ (phi) is also the name of the proximity index used in our MDS analysis, it should always be clear by context to which "phi" we are referring.)

Thus, beta regression models both the mean and the precision, as well as variance and skewness – both functions of the mean and precision. In our case, the health states (theta) and exclusion criteria are the regressors, in models predicting utility values.

Beta regression models follow a strategy similar to that of generalized linear models (e.g., logistic regression), where a transformation of μ and ϕ is modeled as a linear function (called a *linear predictor*, usually denoted η) of the regressors through a *link function* (Shalizi, n.d.). The link function connects the statistic being modeled (e.g., the mean μ) with the linear combination of the regressors (the linear predictor), so that both are unbounded. For the mean (μ), the standard link function is the logit (i.e., $\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$), so that coefficients in a model for μ are changes in log-odds. For ϕ, the link function is usually the logarithm (i.e., $\log(\phi)$).

The boundedness of the beta distribution not only means that the conditional variance depends on the mean, but also that its maximum value is constrained. As

$$Var(y) = \frac{\mu(1-\mu)}{\phi+1},$$

the variance is maximized when the numerator is as large as possible and the denominator is as small as possible. It is straightforward to show that $\mu(1-\mu) = \mu - \mu^2$ is maximized when $\mu = 0.5$, and $\phi + 1$ is minimized when $\phi \to 0$, as $\phi = \omega + \tau > 0$. Thus, $Var(y) < \frac{0.5(1-0.5)}{1} = \frac{1}{4}$. For a fixed μ, the largest variance is given by the smallest precision ϕ, and is bounded above by $\mu - \mu^2$.

One limit to beta regression is that values of the dependent variable – in our case, utility values – cannot equal 0 or 1. This is because the link function maps the random variable to the *entire* real line, and in order to do so smoothly (i.e., so that it is sufficiently differentiable) the link function is not defined for 0 and 1. For data sets with 0s and 1s, the convention is to *squeeze* the data (Smithson & Verkuilen, 2006), 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. This transforms *all* the data, not just the 0s and 1s. As n increases, the impact of the transformation decreases. Unlike a transformation of only the endpoints – say by adding $\epsilon > 0$ to 0s and subtracting it from 1s – the squeeze transformation seeks to avoid introducing bias. By applying a uniform linear transformation to all data (Smithson & Verkuilen, 2006), it preserves the ratios of the distances between every pair of data points, treating the data as interval-scaled, as is assumed for utility (Ellsberg, 1954; Koebberling, 2006; Torrance et al., 1982).

A procedure that avoids squeezing the data is *zero-one inflated beta* (ZOIB) regression (Liu & Kong, 2015). ZOIB regression models the data as a mixture of three models: a beta distribution for data in the open interval $(0, 1)$; a binomial distribution for the outcomes $\{0, \text{ not } 0\}$; and, a binomial distribution for the outcomes $\{1, \text{not } 0 \text{ or } 1\}$. Thus, the data are treated as being generated by these three distributions, with some probability of coming from each one. The first is modeled with beta regression. The second and third are modeled with logistic regressions. More detail is presented in the Appendix. Although ZOIB avoids squeezing the data, the extra models sacrifice parsimony. As a result, we primarily use *squeezed models*, using ZOIB for sensitivity analysis and for one subset of the primary models, described below.

6.2 Beta regression models for health state utilities

Our goal is to model the utilities as a function of the exclusion criteria, so that we can estimate the difference between the utilities associated with the included and excluded responses. As described earlier, PROPr participants assigned utilities to one of the seven health domains that form the health-state space of the PROPr scoring system: cognition, depression, fatigue, pain, physical functioning, sleep and social roles. For each domain, our dependent variable is the utilities assigned to it, as extracted from participants' responses. These states were valued on a scale where 0 is the utility of a domain's disutility corner state (with the unhealthiest level on that domain and the healthiest level on all other domains) and 1 is the utility of full health.

For each domain, we estimate the effects of each exclusion criterion on the mean and precision of the distributions of responses that would be used to represent societal utilities.[1] For a given health domain (whose levels are expressed as theta) and a given exclusion criterion,[2] we use these beta regression models:

$$\text{logit}(\mu_{criterion,domain}) = \beta_0 + \beta_1 theta_{domain} + \beta_2 criterion + \beta_3 theta_{domain} : criterion, \tag{6.1}$$

and,

$$\log(\phi_{criterion,domain}) = \beta_0 + \beta_1 theta_{domain} + \beta_2 criterion + \beta_3 theta_{domain} : criterion, \tag{6.2}$$

Here, μ and ϕ are the mean and precision parameters for the beta distribution; $theta$ is a continuous variable representing health states in the domain (see Table 3.1 and Part II); and $criterion$ is a dummy variable equal to 1 if a response is excluded and 0 otherwise.

In the model for the mean, β_0 (the intercept or constant) gives the mean log-odds utility for included responses, when theta is 0 (equal to the mean population health status on the domain[3]); β_1 gives the change in log-odds utility for a one unit (one population standard deviation) change in theta for included responses; β_2 gives the difference in the intercept for excluded responses; $\beta_3 + \beta_1$ gives the change in log-odds utility for a one-unit change in theta for excluded responses, so that β_3 is the change in slope (on the log-odds scale) when going from the included group to the excluded. Any coefficient involving a *theta* term defines the slope of a best-fit line on the log-odds scale (and the curvature of the line on the utility scale). That slope (or curvature) shows how sensitive predicted utilities are to theta (i.e., how much they change for a given change in health status).

As these estimates are for log-odds utility (i.e., the logit of the mean utility), the estimate for mean utility is

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

[1] If the responses are truly conditionally beta distributed, then by assumption the mean and precision characterize the entire distribution of responses.

[2] Of the nested criteria, *low-range* and *no-variance*, we chose *low-range* to represent the pair in our regression analyses, as we know that those excluded by *no-variance* will necessarily have a completely horizontal mean utility curve, while those excluded by *low-range* could show some small perturbations from the horizontal. (Perturbations that are, *a priori*, not predictable.)

[3] By construction, a theta of 0 on any PROMIS domain is the mean score for the PROMIS reference population and a change of 1 on the theta scale is a change of one standard deviation. The PROMIS reference population has been shown to be equivalent to the general population (H. Liu et al., 2010).

where $\eta = \beta_0 + \beta_1 theta_{domain} + \beta_2 criterion + \beta_3 theta_{domain} : criterion$.

In the model for the precision parameter ϕ, the parameters are similar, except that ϕ is transformed via the log link function, rather than the logit.

Although many features of these distributions might be relevant to understanding the effects of exclusion criteria, we focus on the mean μ, the estimate used in almost all analysis involving estimates of societal utilities (see Part I).

Equation (6.1) models the parameters as a linear function of theta. Recall that the utilities were elicited for six or seven values of theta, corresponding to the states in Table 3.1. To test this linearity assumption, we compare models that treat theta as a linear variable and as a factor (categorical) variable.

To estimate the coefficients of these models, we use the **betareg** package in R (Cribari-Neto & Zeileis, 2010). The Appendix describes the estimation methods used for sensitivity analyses, including the ZOIB models.

6.3 Incorporating participant health status

Individuals' health status can affect their preferences for health states (Neumann et al., 2000). For example, Mulhern et al. (2014) find that patients with epilepsy provide higher values for health states describing severe epilepsy-related impairment than the values provided by the public. Here, we ask how participants' health status on a domain is related to their valuations *for that domain* and to the classification of their responses by the exclusion criteria. We represent participant health by their responses to the PROMIS-29 and Cognition 4-item short form questionnaires, which provide PROMIS scores on the seven PROPr domains. How societal preferences should weight the preferences of individuals with different health states is a moot point, reflecting competing principles (Sanders et al., 2016). The present analyses show how important its resolution is in practical terms, by estimating how sensitive utilities are to values of theta for individuals with varying personal experience as well as their interaction with exclusion criteria.

The beta regression model we use to incorporate participant health status modifies equation

(6.1) to include participants' PROMIS scores:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 theta + \beta_2 criterion + \beta_3 PROMIS + \beta_4 theta : criterion \\ + \beta_5 theta : PROMIS + \beta_6 PROMIS : criterion + \beta_7 theta : PROMIS : criterion, \quad (6.3)$$

where PROMIS is a participant's PROMIS score on the domain. Equation (6.3) will show, for example, how any difference between the height of the utility curve for the included (β_0) group and the excluded group ($\beta_0 + \beta_2$) changes as a function of the health of the included (β_3) and the excluded (β_6). It also shows us how any difference in sensitivity (i.e., slope on the log-odds scale or curvature on the utility scale) between the utility curve for the included (β_1) and the excluded ($\beta_1 + \beta_4$) varies as a function of the health of the included (β_5) and the excluded (β_7).

6.4 Beta regression results

6.4.1 Results of the beta regression models

Table 5.4 shows the number and percentage of participants (out of the 1,164 total) excluded by each criterion. Every domain was valued by 166 or 167 participants. The proportion excluded by each criterion is similar across the domains (Table 5.5), although there is some variation (e.g., 12.7% of those valuing pain were flagged by *numeracy*, compared to 4.2% in physical functioning and 7.8% across the whole sample). Figures 6.1-6.9 show the curves for models predicting the mean (equation (6.1)) as they vary over theta, for both the excluded and included groups. The error in estimating the difference between the utility curves of the excluded and included depends on the number of responses in each group and their variability. A later section discusses the work needed to produce uncertainty estimates.

For expository purposes, we focus initially on the model for one domain, sleep, and one exclusion criterion, *numeracy*, which excludes all responses of participants who score below a threshold on the numeracy test (Table 5.3). In Figure 6.10, the solid black dots are conditional means estimated by the factor model for individuals excluded by the numeracy criterion; the solid black line is the best-fit beta regression for those individuals, treating theta as a continuous variable

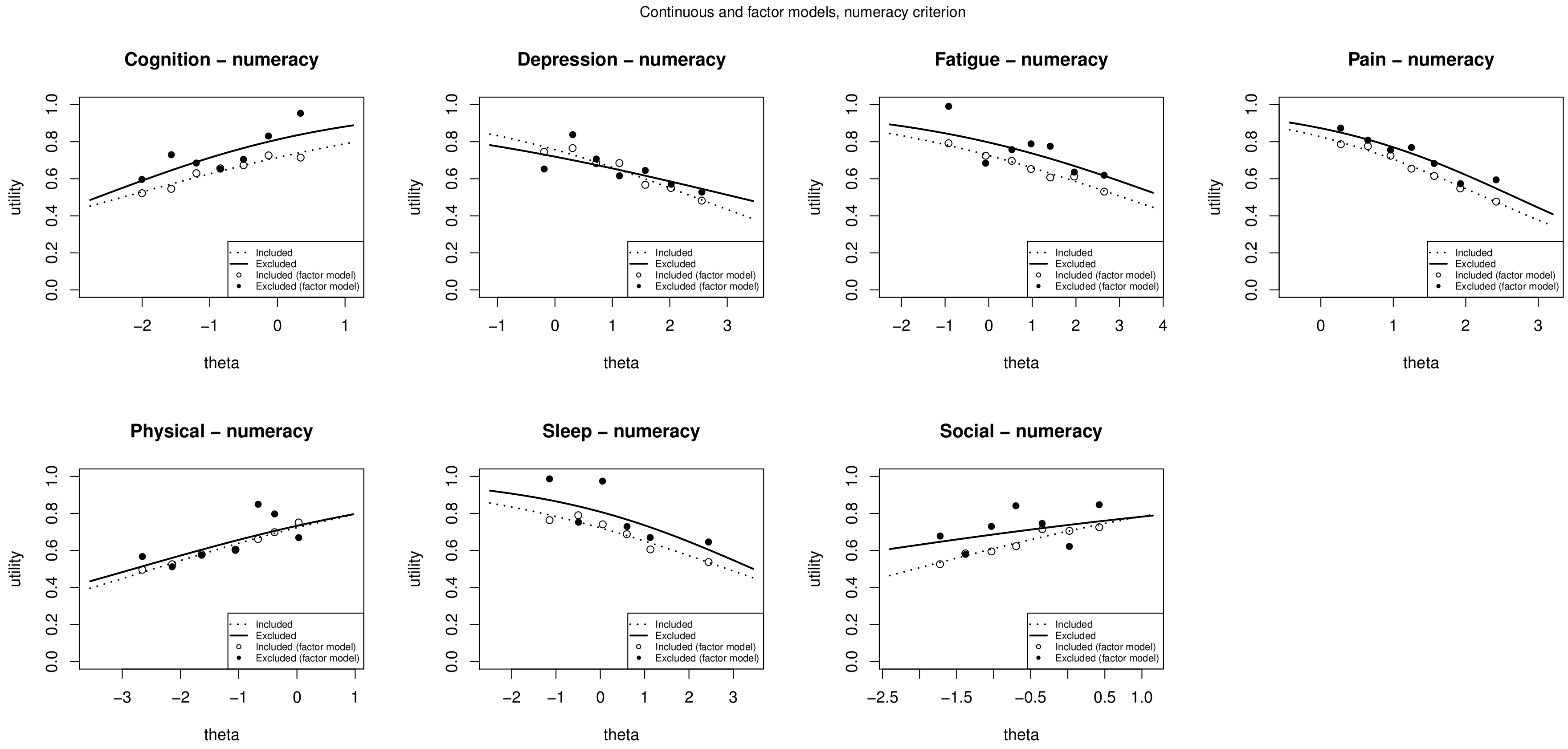


Figure 6.1: *Beta regression models for numeracy.* Modeling mean utilities for each domain as a function of theta and the *numeracy* criterion.

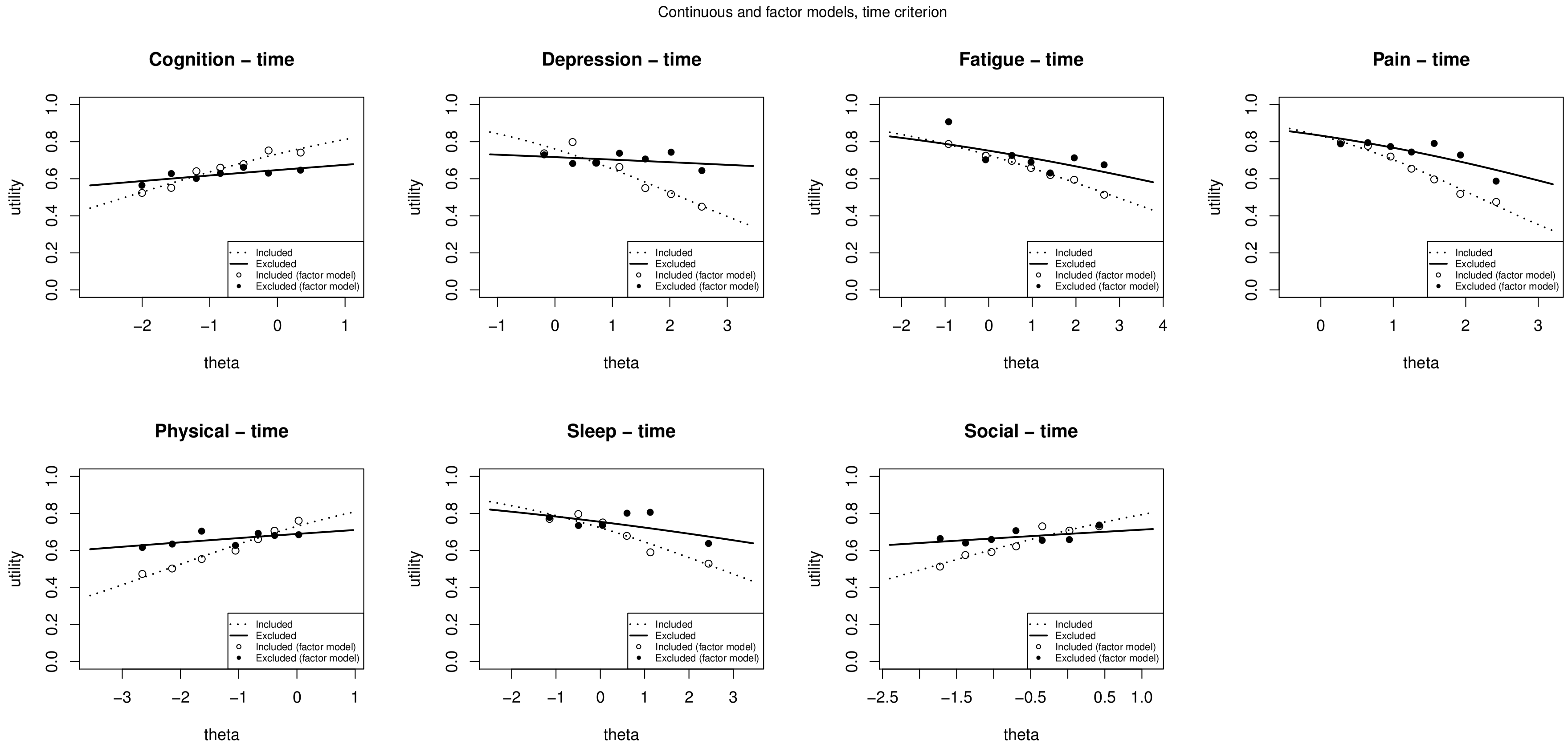


Figure 6.2: *Beta regression models for time.* Modeling mean utilities for each domain as a function of theta and the *time* criterion.

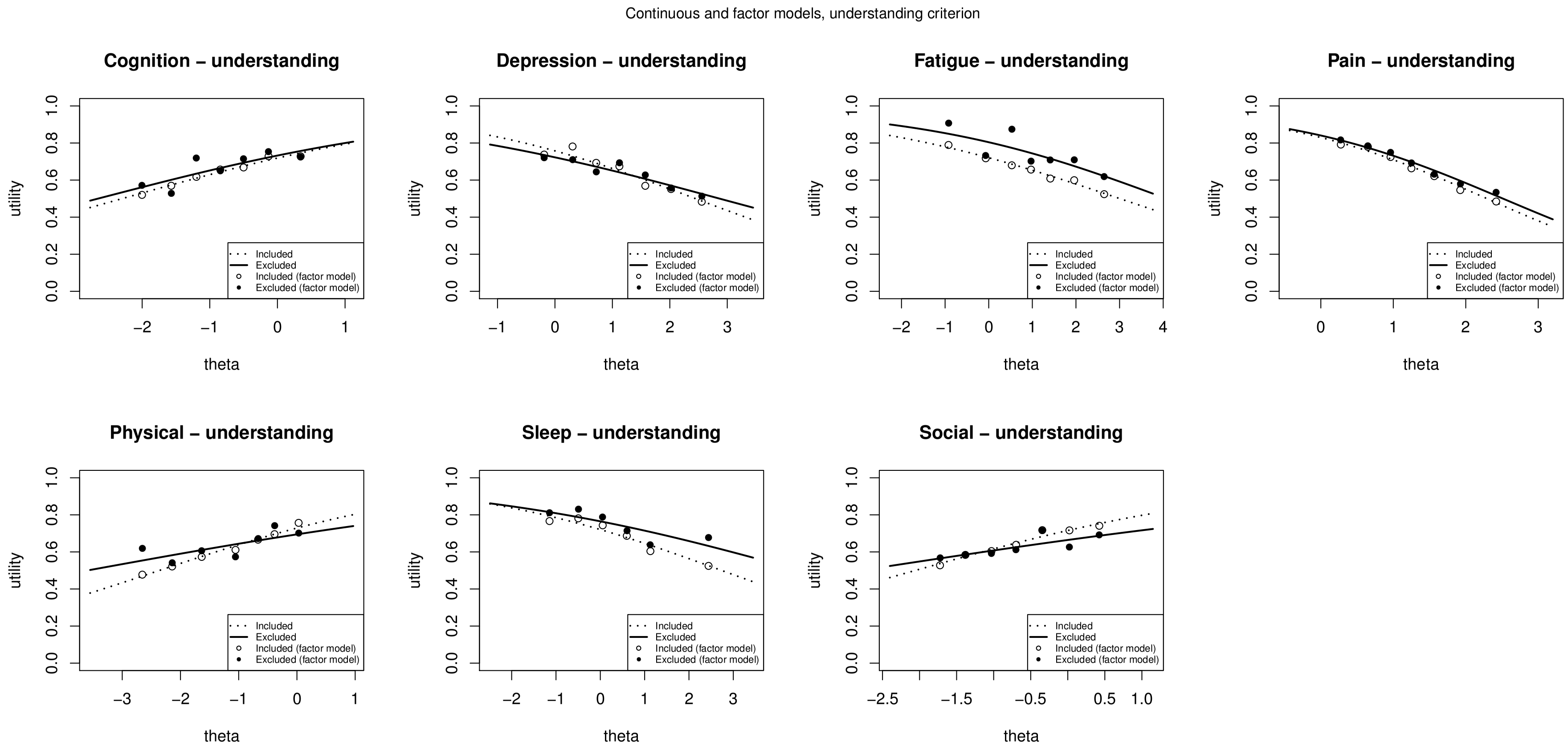


Figure 6.3: *Beta regression models for understanding*. Modeling mean utilities for each domain as a function of theta and the *understanding* criterion.

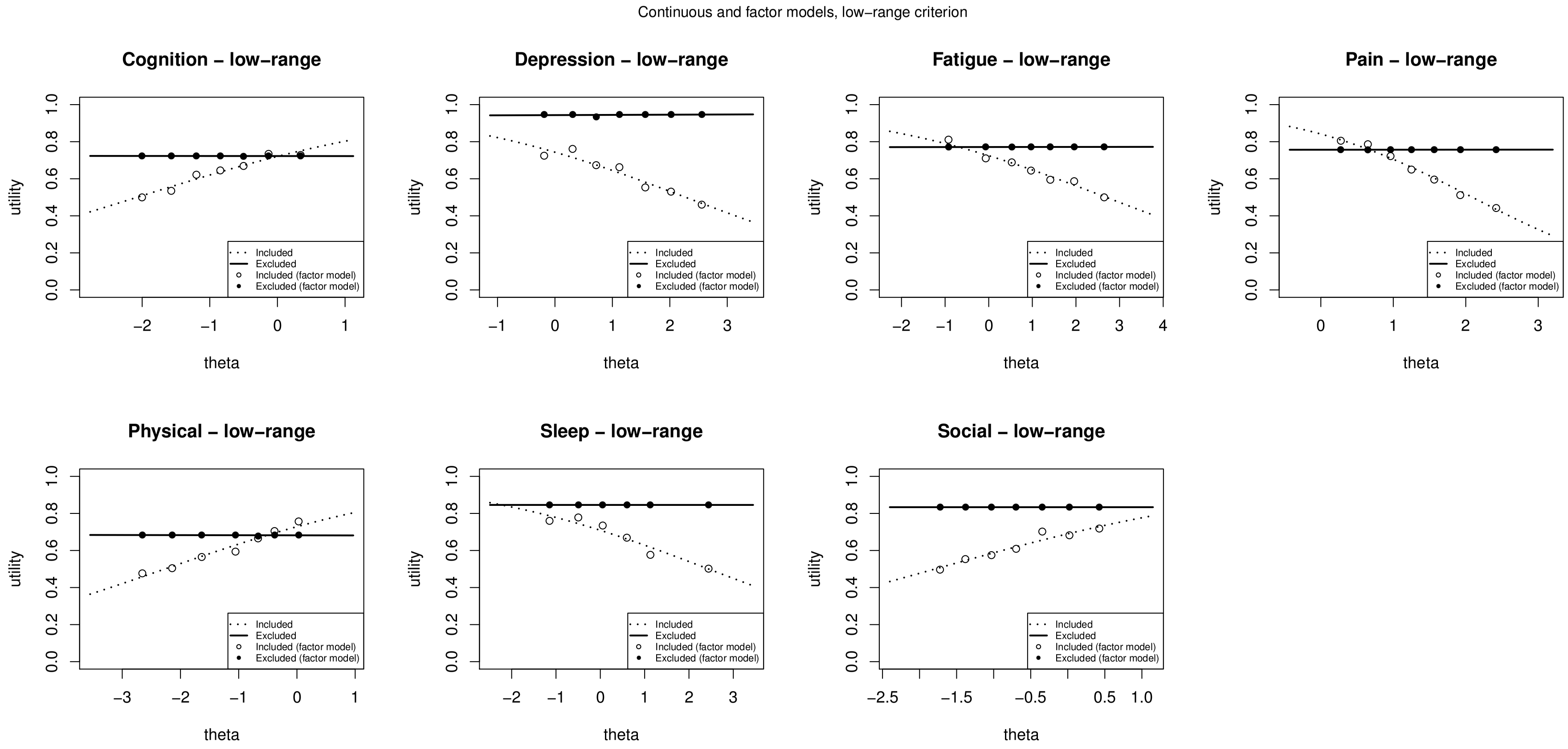


Figure 6.4: *Beta regression models for low-range.* Modeling mean utilities for each domain as a function of theta and the *low-range* criterion.

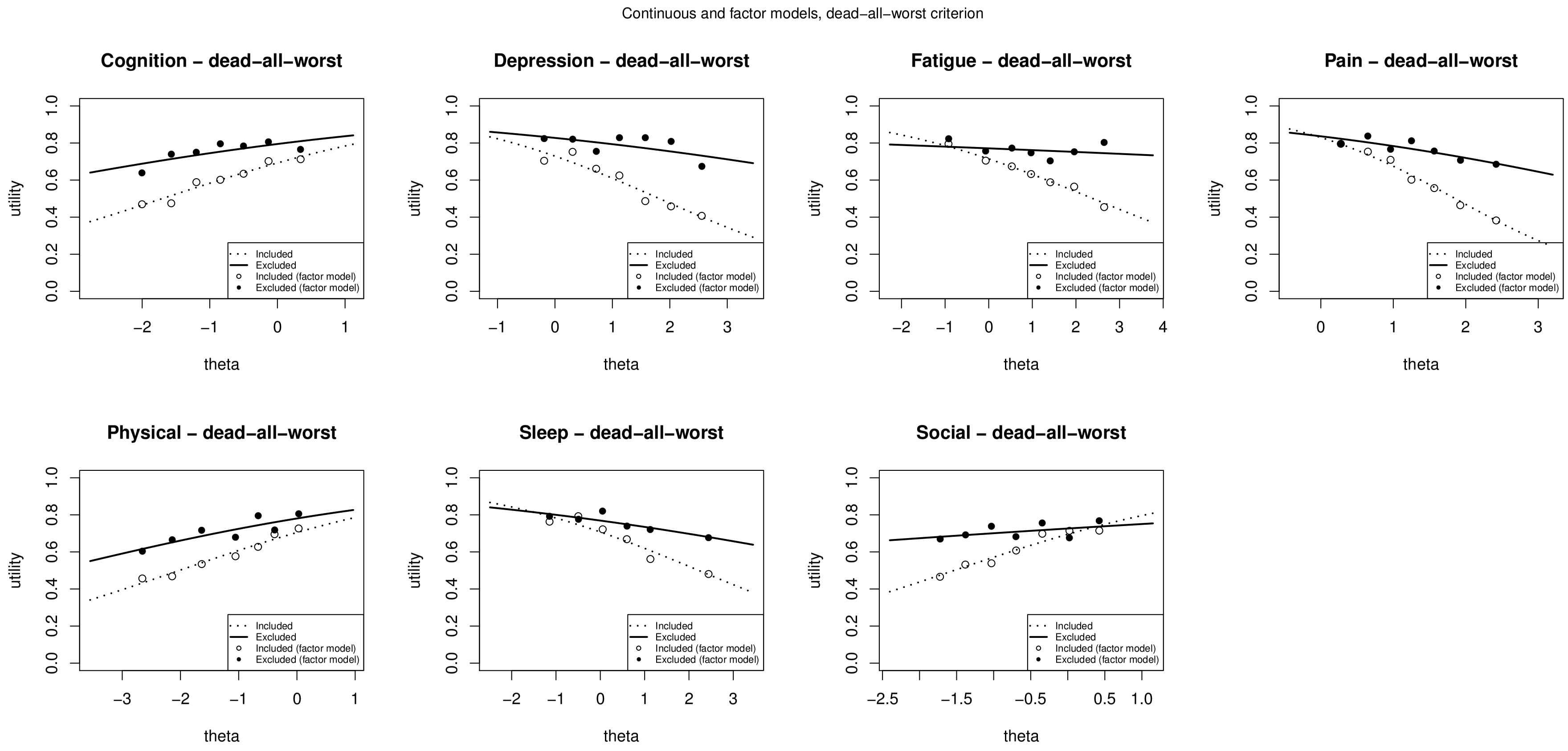


Figure 6.5: *Beta regression models for dead-all-worst.* Modeling mean utilities for each domain as a function of theta and the *dead-all-worst* criterion.

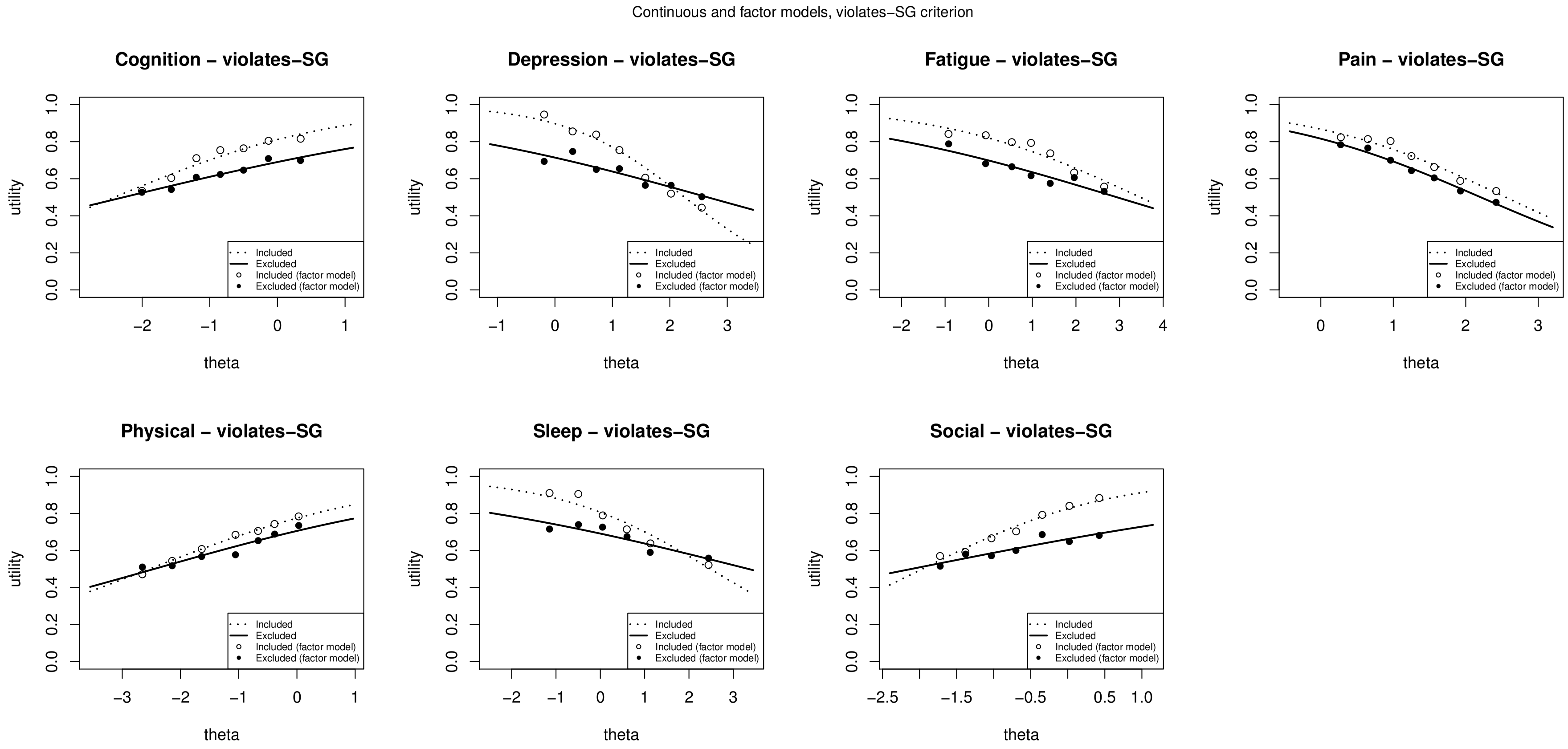


Figure 6.6: *Beta regression models for violates-SG*. Modeling mean utilities for each domain as a function of theta and the *violates-SG* criterion.

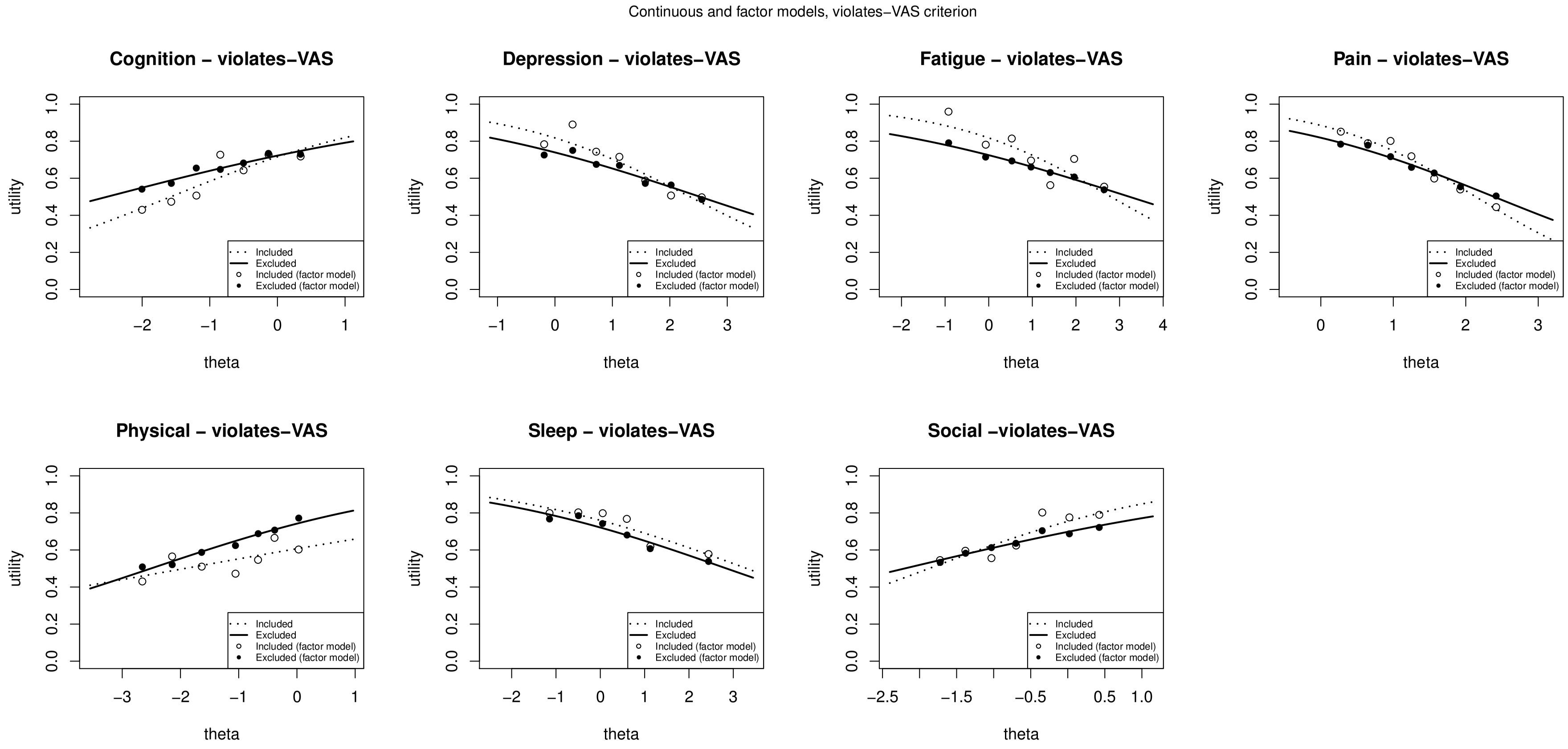


Figure 6.7: *Beta regression models for violates-VAS*. Modeling mean utilities for each domain as a function of theta and the *violates-VAS* criterion.

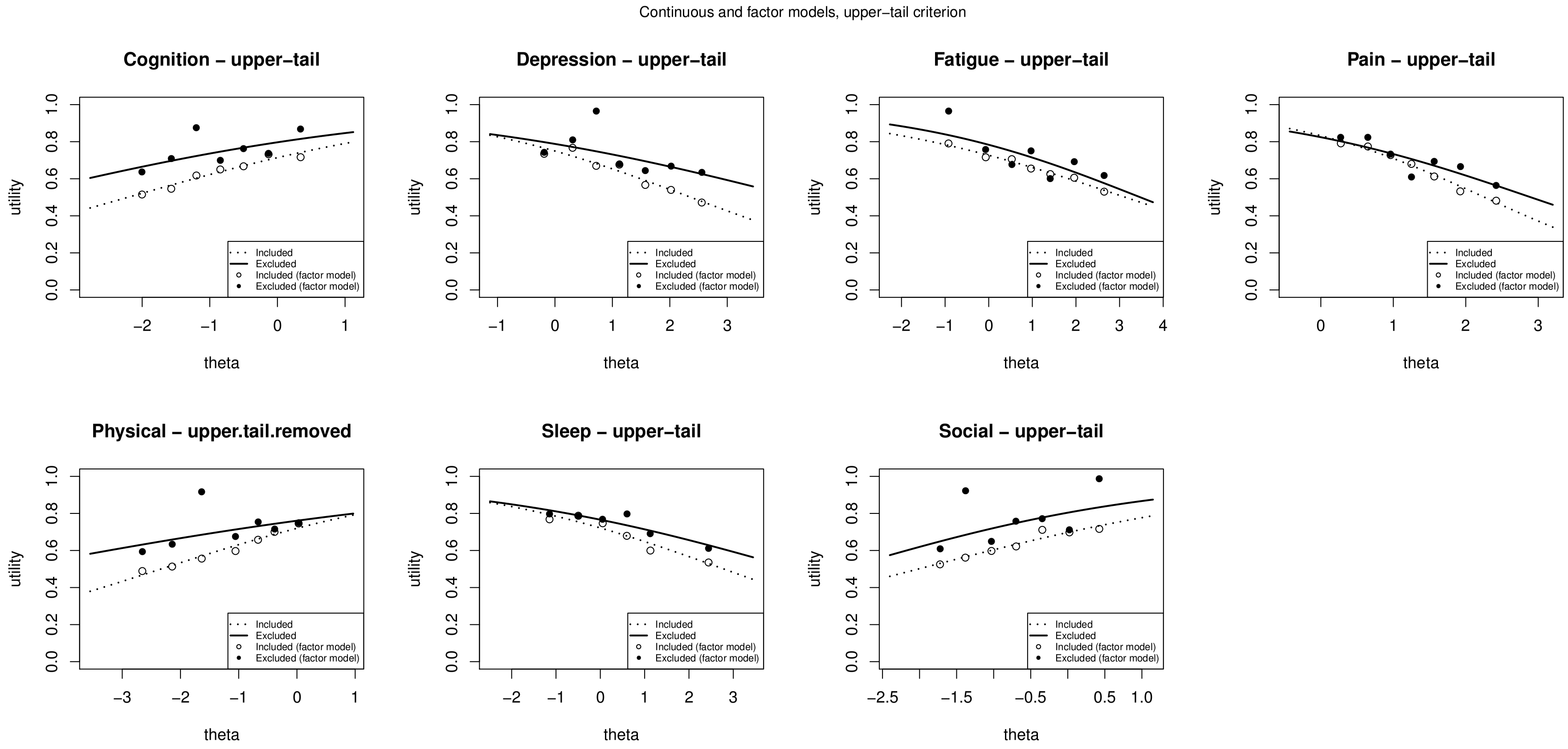


Figure 6.8: *Beta regression models for upper-tail.* Modeling mean utilities for each domain as a function of theta and the *upper-tail* criterion.

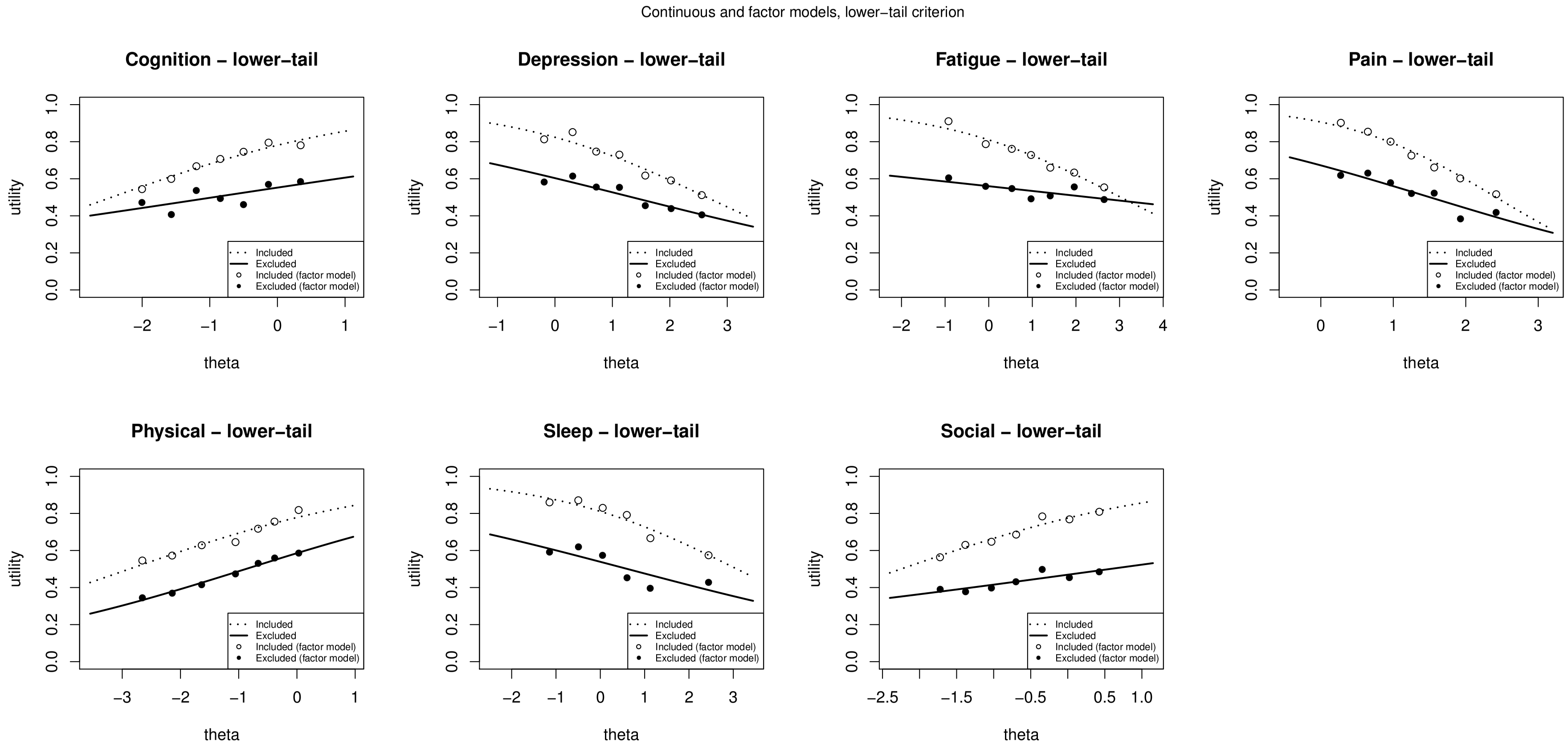


Figure 6.9: *Beta regression models for lower-tail.* Modeling mean utilities for each domain as a function of theta and the *lower-tail* criterion.

(dummy variable = 1). The open circles and dotted line are for individuals included by the numeracy criterion (dummy variable = 0). The figure shows that, with the continuous model (the curves), those excluded by *numeracy* have higher utility for sleep, for all theta, than those not excluded. For the factor model (the dots), the same result holds, except for the state corresponding to a theta of −0.50, where the open circle is above the solid dot. We focus on the continuous model here, considering the factor model in the sensitivity analyses (see Appendix). Finally, the solid blue curve shows the conditional mean from estimating the continuous model on the full sample, with no exclusion criterion applied (i.e., equation (6.1) with the $criterion$ variable removed).

Table 6.1 shows the regression coefficients for the mean model (equation (6.1)) treating theta as a continuous variable. The entries are on the logit (log-odds) scale, so an entry of value x has a value of $\frac{e^x}{1+e^x}$ on the utility scale.

Table 6.1: *Coefficients for the beta model of mean sleep utilities as a function of theta and numeracy.* The beta regression model for mean sleep utilities as a function of theta and the *numeracy* criterion. Note that the coefficients are on the log-odds scale, and that *numeracy* = 1 indicates exclusion.

	Dependent variable:
	log-odds utility
constant (intercept)	0.954^{***}
	(0.052)
theta	-0.332^{***}
	(0.040)
numeracy	0.487^{**}
	(0.245)
theta:numeracy	−0.085
	(0.180)
Observations	996
pseudo-R^2	0.080
Log Likelihood	1,562.959
Note:	*p<0.1; **p<0.05; ***p<0.01

The constant corresponds to the utility of a theta score of 0 for the group not excluded by *numeracy* (dummy = 0). By construction, PROMIS assigns theta of 0 to the mean population health

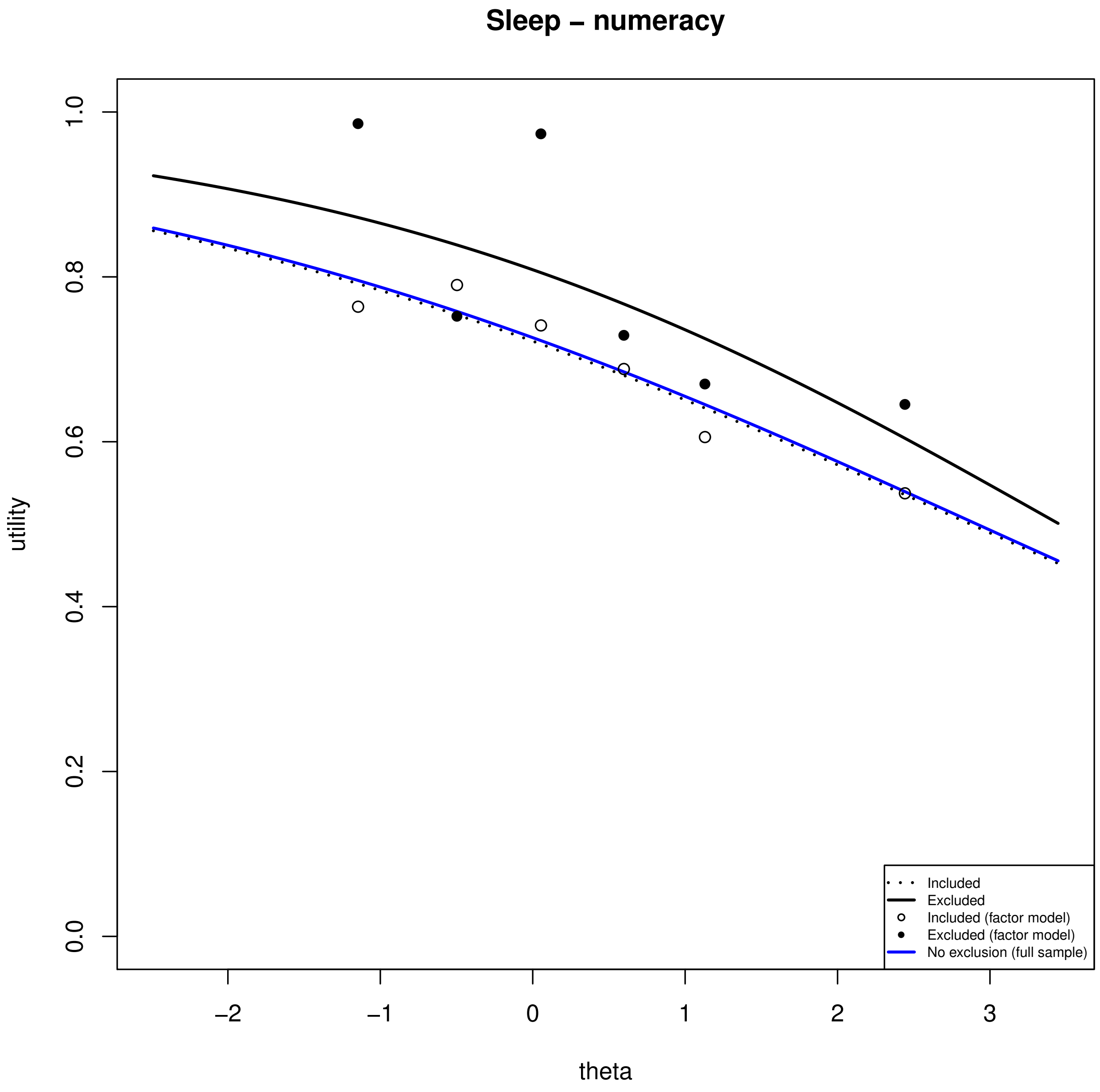


Figure 6.10: *Beta regression of sleep as a function of theta and numeracy.* The curves show the conditional means for the model of utilities for sleep as a function of the interaction of the exclusion criterion *numeracy* with sleep's theta construct, treated as a continuous variable (and specified as a linear term in the model). The points show the predicted means from the equivalent factor model, where theta is treated as a factor (categorical) variable. The solid black curve (solid black points) show the predicted means for those excluded by *numeracy*, the dotted black curve (empty points) show the predicted means for those not excluded. The solid blue curve shows the mean estimated on the entire sample (i.e., with no exclusion criterion implemented).

status on a domain. For sleep, it corresponds to sleep of moderate quality (e.g., rarely waking up and having trouble falling back to sleep). The value of 0.954 (constant in Table 6.1) on the log-odds scale equals 0.722 on the 0-1 utility scale. The coefficient on *theta* is the predicted change in the mean utility (on the log-odds scale) for a unit increase in theta, for the group not excluded. As seen in Figure 6.10, the utility of sleep decreases for those not excluded (the solid line) as sleep disturbance worsens (i.e., as theta increases). The value of the theta coefficient on the log-odds scale, -0.332, says that, for example, moving from a theta of 0 to a theta of 1 (one standard deviation of worse sleep disturbance in the general population) reduces predicted mean utility from $\text{logit}^{-1}(0.954) = 0.722$ to $\text{logit}^{-1}(0.954 - 0.332 \times 1) = \text{logit}^{-1}(0.954 - 0.332) = \text{logit}^{-1}(0.622) = 0.651$.

The *numeracy* coefficient in Table 6.1, indicates the difference in estimated utility of theta = 0 for the excluded group (solid line in Figure 6.10). The positive sign means participants with low numeracy scores assigned higher values to sleep quality (so that excluding them reduces the societal utility of sleep quality). The log-odds utility for the excluded group is the *numeracy* coefficient plus the constant, which on the utility is the inverse logit of $(0.954 + 0.487) = 1.441$, equal to a utility of 0.809 (i.e., $\text{logit}^{-1}(1.441)$), compared to the non-excluded group's predicted value of 0.722.

The coefficient on 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 6.10, the curvature (sensitivity) of the excluded group (solid line) is only slightly more pronounced than that of the non-excluded group (dotted line).[4] The closeness of the model with no exclusion (solid blue) to the non-excluded group (dotted line) is caused by the relatively small number excluded by *numeracy* (Table 5.5) and the size of its effect.

By way of contrast, Table 6.2 and Figure 6.11 show the model with the *violates-SG* criterion. Unlike *numeracy*, where the utility curves for the included and excluded groups are roughly parallel (the theta-criterion interaction coefficient is small), the interaction effect *theta:violates-SG* is -0.619 on the log-odds scale (0.35 on the utility scale). The curve for the excluded group (solid

[4] The value of the interaction term in Table 6.1 means that if we go from a theta of 0, with a predicted mean utility for the excluded group of $\text{logit}^{-1}(0.954 + 0.487) = \text{logit}^{-1}(1.441) = 0.809$, to a theta of 1, the predicted utility is $\text{logit}^{-1}(1.441 + (-0.332 - 0.085) \times 1) = 0.736$. The utility is changing in the expected direction, with a slightly larger change than for the non-excluded group (0.073 versus 0.071).

black) is around 0.15 higher for high sleep quality (negative values of theta), but decreases much more quickly than the curve for the included group as sleep quality decreases, eventually intersecting and moving below the curve for the included group. In contrast to the previous example, the model estimated with the full sample (solid blue) is different in both sensitivity and elevation from the included (dotted line), because of the large number excluded by *violates-SG* (Table 5.5) and the large interaction term (Table 6.2). That means implementing *violates-SG* would have a large effect on utility calculations based on the final utility curve, whereas implementing *numeracy* would not. Figure 6.12 shows the included (dotted) curves for every criterion when applied to the sleep domain, as well as the curve for the full sample (no exclusion).

Table 6.2: *Coefficients for the beta model of mean sleep utilities as a function of theta and violates-SG.* The beta regression model for mean sleep utilities as a function of theta and the *violates-SG* criterion. Note that the coefficients are on the log-odds scale, and that *violates-SG* = 1 indicates exclusion.

	Dependent variable:
	log-odds utility
constant (intercept)	1.426^{***} (0.093)
theta	-0.574^{***} (0.067)
violates-SG	-0.619^{***} (0.111)
theta:violates-SG	0.333^{***} (0.082)
Observations	996
pseudo-R^2	0.11
Log Likelihood	1,580.685
Note:	*p<0.1; **p<0.05; ***p<0.01

The goodness-of-fit statistic, the pseudo-R^2 reported in Table 6.1 and Table 6.2, also known as the proportional reduction of error (PRE), compares the log-likelihoods of the null model (estimating the global mean) and of the model under consideration (Smithson & Verkuilen, 2006). The value in Table 6.1 is 0.08. To provide context for that fit, we conducted simulations using data

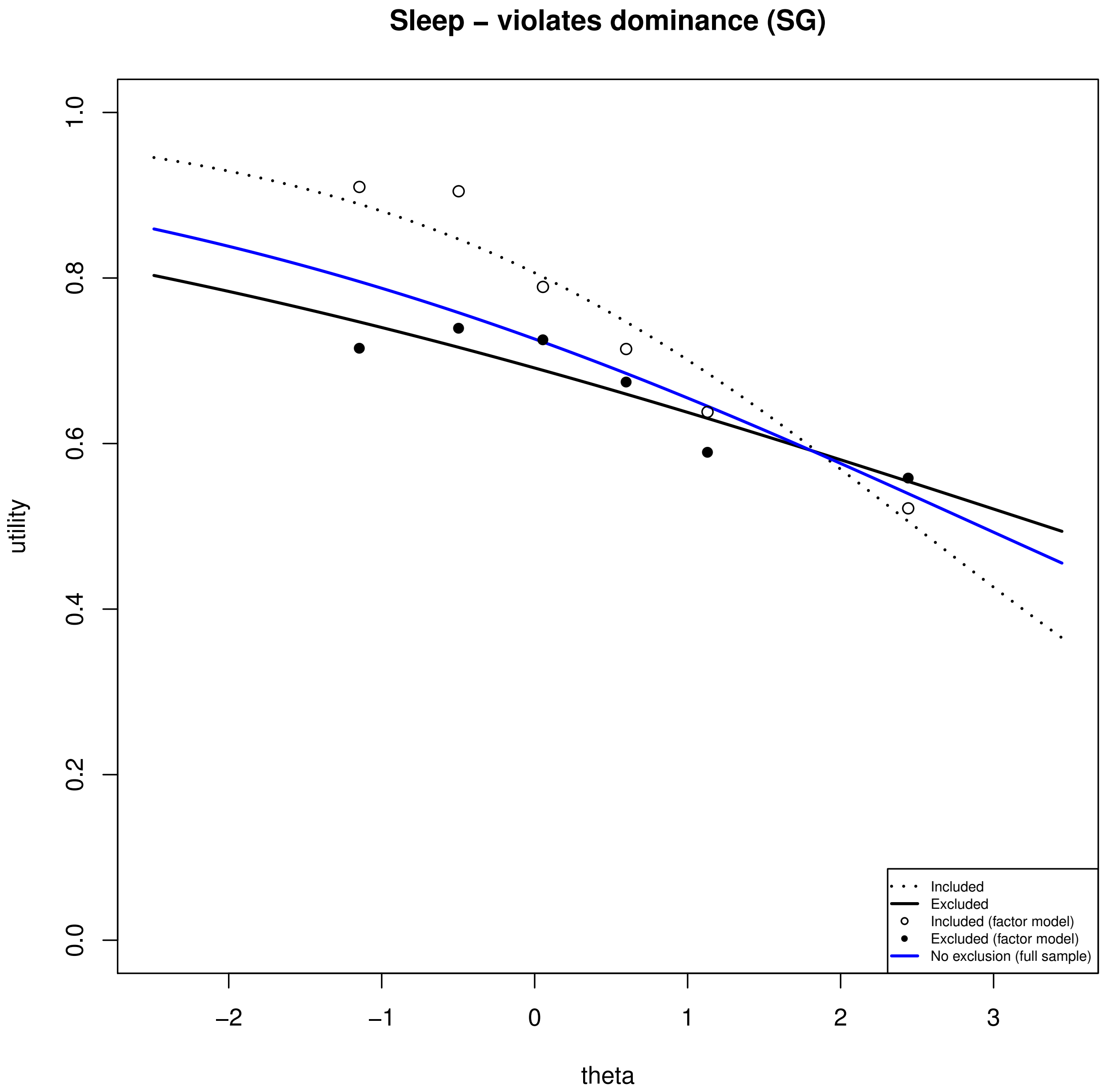


Figure 6.11: *Beta regression of sleep as a function of theta and violates-SG.* The curves show the conditional means for the model of utilities for sleep as a function of the interaction of the exclusion criterion *violates-SG* with sleep's theta construct, treated as a continuous variable (and specified as a linear term in the model). The points show the predicted means from the equivalent factor model, where theta is treated as a factor variable. The solid black curve (solid black points) show the predicted means for those excluded by *violates-SG*, the dotted black curve (empty points) show the predicted means for those not excluded. The solid blue curve shows the mean estimated on the entire sample (i.e., with no exclusion criterion implemented).

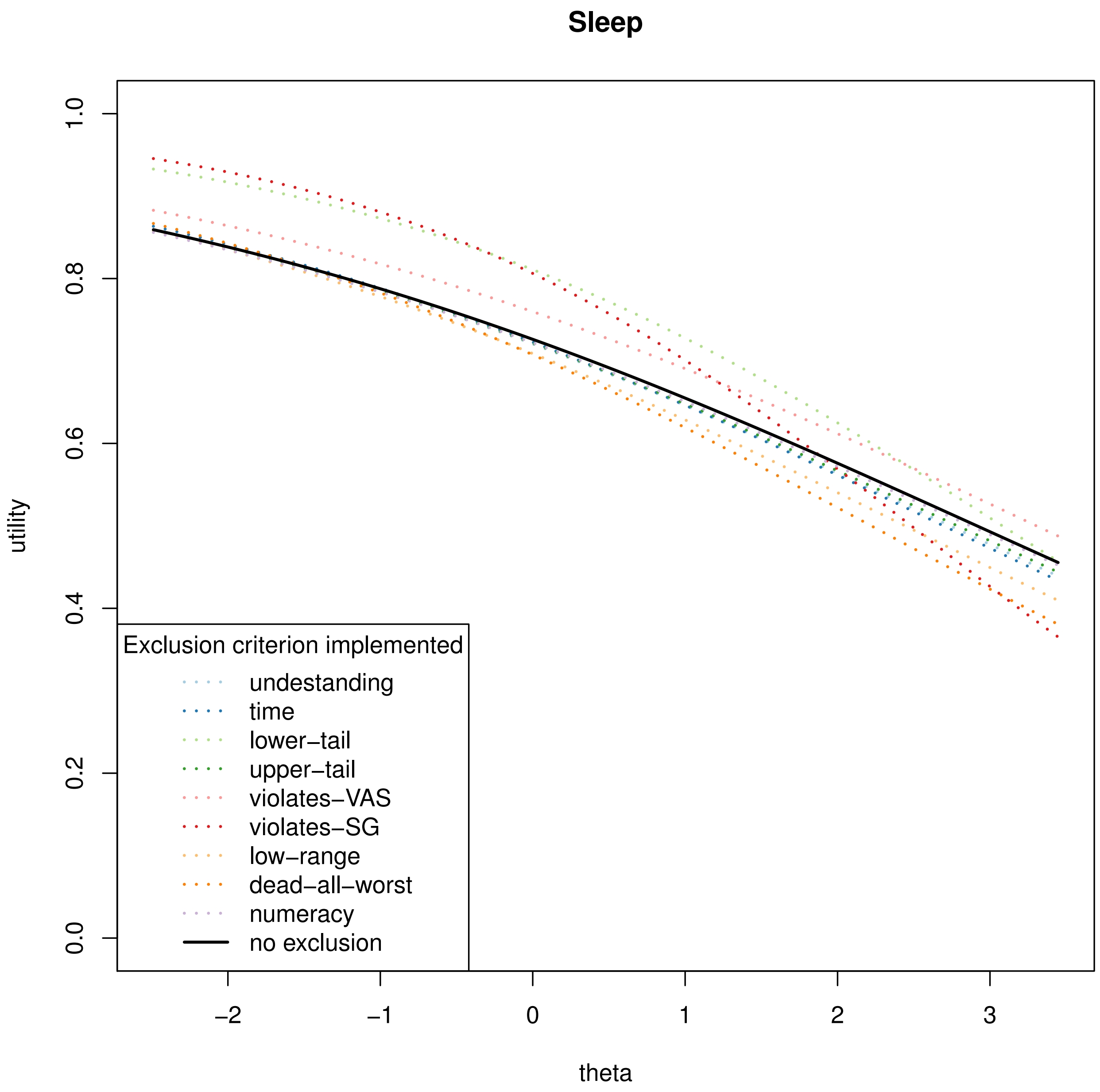


Figure 6.12: *Utility curves for sleep, after each exclusion and with no exclusion.* The estimated conditional mean utility curve for sleep, after applying the indicated exclusion criterion (or no exclusion).

created to be conditionally beta distributed, with two key properties of the present data: (a) a continuous regressor discretized to have six values matching the theta values used in the sleep domain and (b) the slope estimated in Table 6.1. The pseudo-R^2 of that model is 0.09 (compared to 0.08 for the actual data). We repeated the simulation, for *violates-SG* (Table 6.2). The pseudo-R^{2}is 0.22, compared to the observed value of 0.11 in Table 6.2.

Figure 6.13 shows the results of analogous analyses for all seven domains, comparing the curve for all responses (no exclusion) with the curves for those remaining after applying each exclusion criterion.

6.4.2 Results of incorporating participant health status

Figure 6.14 shows the distributions of PROMIS scores for participants valuing each domain. Most are skewed, with modes at or near the highest functional capacities. Thus, in this representative sample, good health is modal. The exceptions are fatigue and sleep, both of which are closer to bell-shaped. Because the more symmetric distribution of sleep scores simplifies including it as a regressor, we continue initially with that example from the previous section.[5] We look first at the *violates-SG* exclusion criterion, because of its large effect on the included preferences (Figure 6.12).

We created a three-way interaction model with the regressors of health (theta), exclusion (*violates-SG* or not) and *participant* health (PROMIS score), following equation (6.3). To make the results more easily interpretable, we discretized participants into those with better sleep health (sleep PROMIS score < 0) and those with poorer sleep health (sleep PROMIS score ≥ 0).

For the entire data set, the conditional distributions for the squeezed model were too far from being beta distributed to allow reliable estimation. Removing the 0s and 1s made the beta assumption more suitable, without squeezing. As a result, we use the zero-one inflated beta (ZOIB) models.[6] They model the untransformed non-0/1 data as a beta regression, and the 0 and 1 responses with separate logistic regressions. There were 107 responses of 0 and 553 responses of 1, among the 996 responses.

[5] We do not have to worry about the possibility of little variation in health, which could be the case for those with a spike at a utility of 1.

[6] The Appendix provides a detailed account of the results leading to the use of the ZOIB models as the main modeling strategy for this set of analyses.

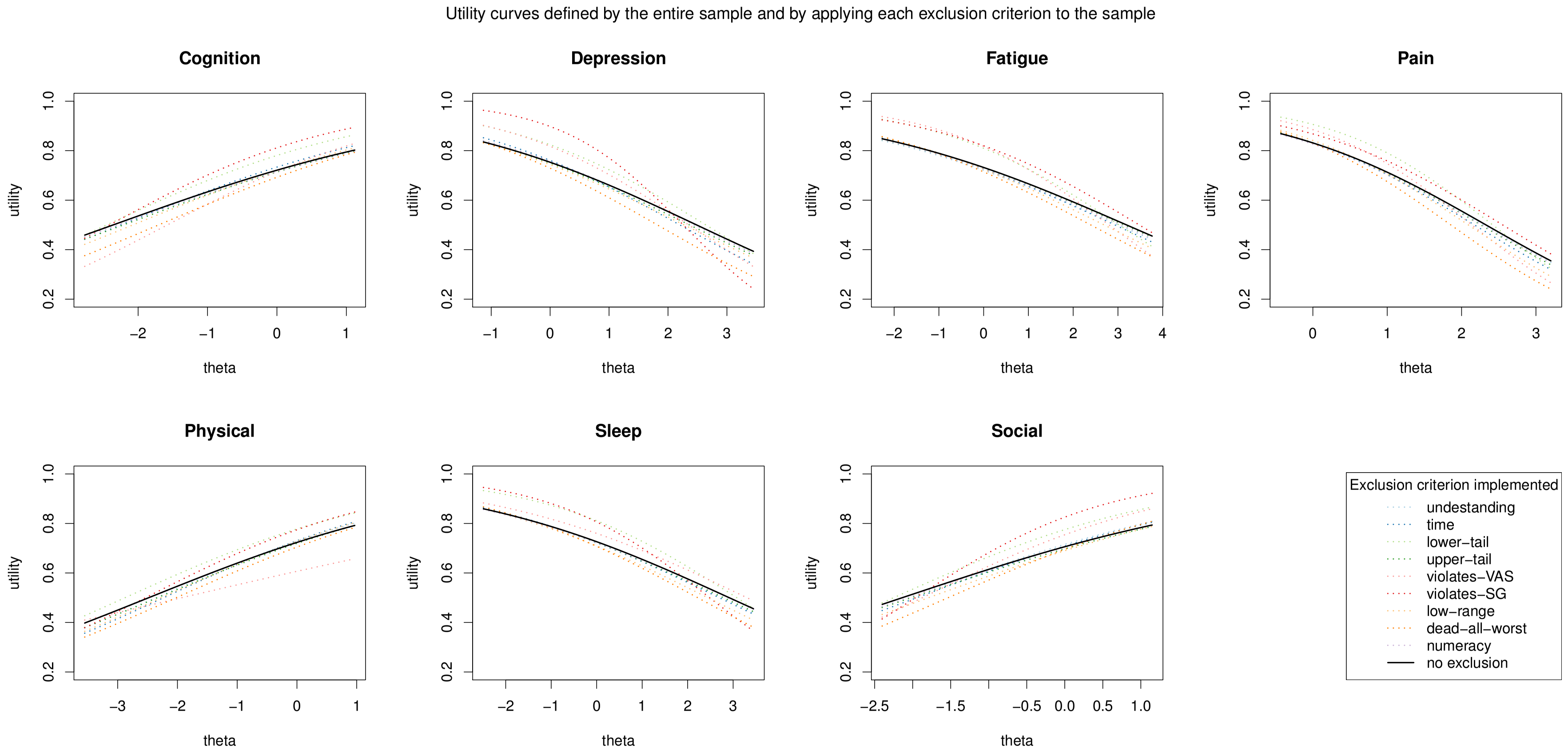


Figure 6.13: *Utility curves for each domain, after each exclusion and with no exclusion.* The estimated conditional mean utility for each domain, after applying each exclusion criterion (or no exclusion). (Note y-axis starts at 0.2, to magnify the curves.)

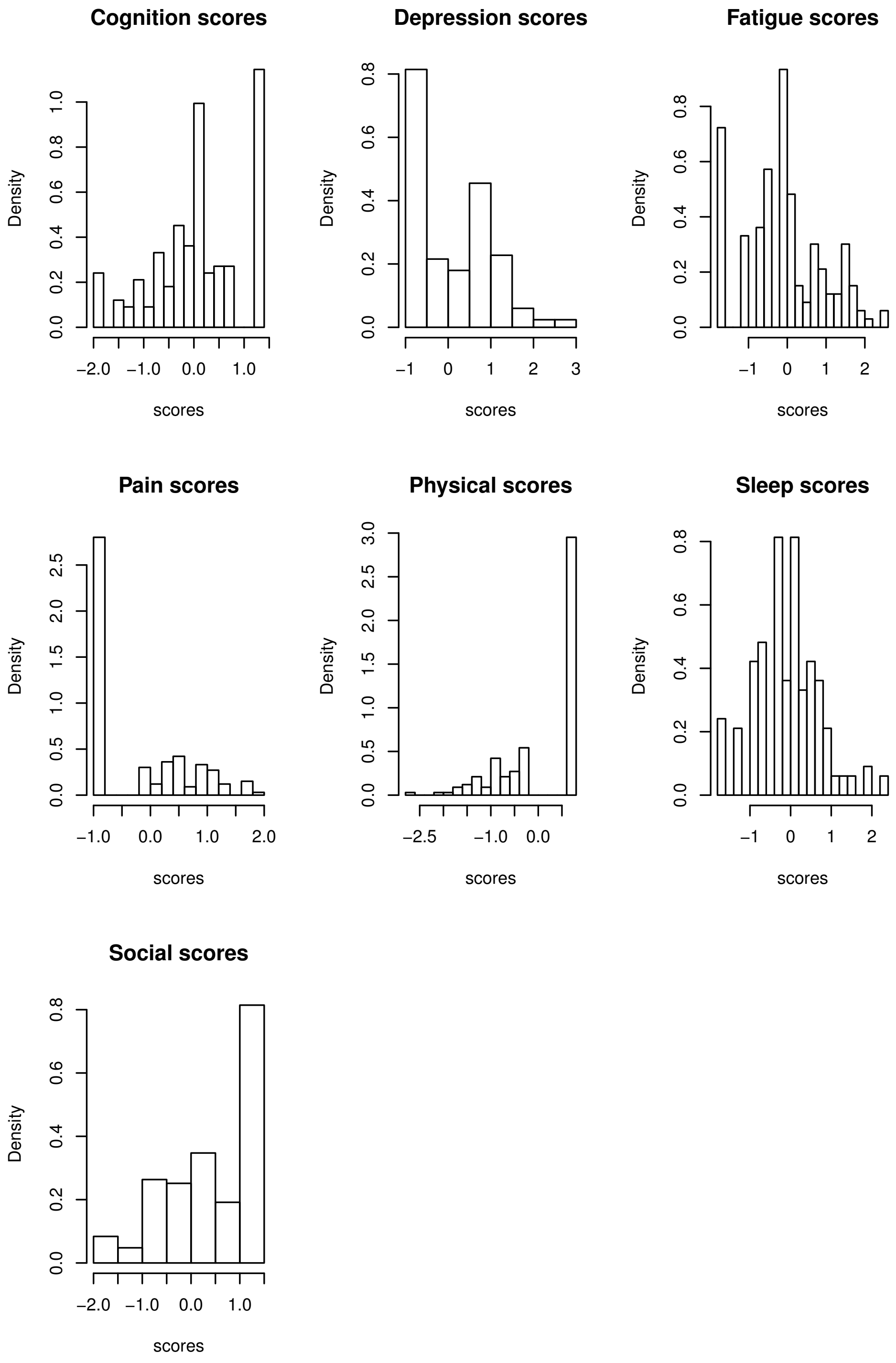


Figure 6.14: *Histograms of participants' PROMIS scores.* Each histogram shows the participant PROMIS scores for the same domain that they were randomly assigned to value in the PROPr survey.

We first present the three constituent parts of the ZOIB model, and then the final model that combines them. Figure 6.15 shows best-fit lines for the beta portion of the model, estimated on the 336 of 996 responses that were not 0 or 1, for the four groups defined by sleep quality and exclusion by *violates-SG*. All four groups assign lower utilities as sleep disturbance gets worse, with excluded participants who sleep poorly showing the least sensitivity to changes in states of sleep,[7] whereas their included counterparts show the most.[8] The predicted population mean level of sleep (theta of 0) is highest for included participants who sleep poorly[9] and lowest for excluded participants who sleep well.[10]

Comparing Figure 6.11 and Figure 6.15 shows that adding the PROMIS term (the *sleep-quality* variable in Table 6.3) has increased the difference between the curves of the included and excluded when sleep quality is poor. Comparing the models (Table 6.2 and Table 6.3), the pseudo-R^2 of the model with participant PROMIS scores increases from 0.11 to 0.15.

The logistic regression predicting responses of 0 (Figure 6.16) shows that excluded participants (solid lines) were much less sensitive to the level of sleep quality that they were evaluating (theta) than were the included participants (dotted lines), as reflected in their predicted propensity to assign a value of 0. That insensitivity was similar for those who slept well (green) and poorly (red). In contrast, included participants were no more likely to assign 0 utility to good sleep states, but much more likely to do so for poor sleep states – with that change even more pronounced for those who slept poorly (dotted red).

The logistic regression predicting responses of 1 (Figure 6.17) shows that included participants are more likely to assign 1 to the best sleep states (dotted lines). Participants who sleep better (green) are more sensitive to changes in theta than those who sleep poorly, whether included or excluded. Those who sleep poorly and are excluded (solid red) were almost as likely to assign 1 to all levels of sleep, suggesting indifference to the health state and insensitivity to the task.

Figure 6.18 combines the three models, by weighting each following equation 6 in Liu & Kong (2015). The composite curves show that included participants (dotted lines) are more sensitive to

[7] The sum of all coefficients involving *theta* in Table 6.3.

[8] Removing the contribution of the three-way interaction *theta:sleep-quality:violates-SG* as well as the two-way interaction *theta:violates-SG* in Table 6.3.

[9] The sum of the constant and the coefficient on *sleep-quality* in Table 6.3.

[10] The sum of the constant and the coefficient on *violates-SG* in Table 6.3.

Table 6.3: *Three-way interaction beta model for sleep (with no 0s or 1s), as a function of theta, violates-SG, and participant sleep health.* The beta regression model, excluding the 0s and 1s from the data, for mean sleep utilities as a function of theta, the *violates-SG* criterion, and participant sleep health (*sleep-quality*). Note that the coefficients are on the log-odds scale, and that *violates-SG* = 1 indicates exclusion and *sleep-quality* = 1 indicates poor sleep health. This model is the beta portion of the associated ZOIB model.

	Dependent variable:
	log-odds utility
constant (intercept)	0.632***
	(0.155)
theta	−0.553***
	(0.137)
sleep-quality	0.605**
	(0.261)
violates-SG	−0.322*
	(0.184)
theta:sleep-quality	−0.433*
	(0.228)
theta:violates-SG	0.170
	(0.157)
sleep-quality:violates-SG	−0.575*
	(0.297)
theta:sleep-quality:violates-SG	0.644**
	(0.254)
Observations	336
pseudo-R^2	0.149
Log Likelihood	40.443
Note:	*p<0.1; **p<0.05; ***p<0.01

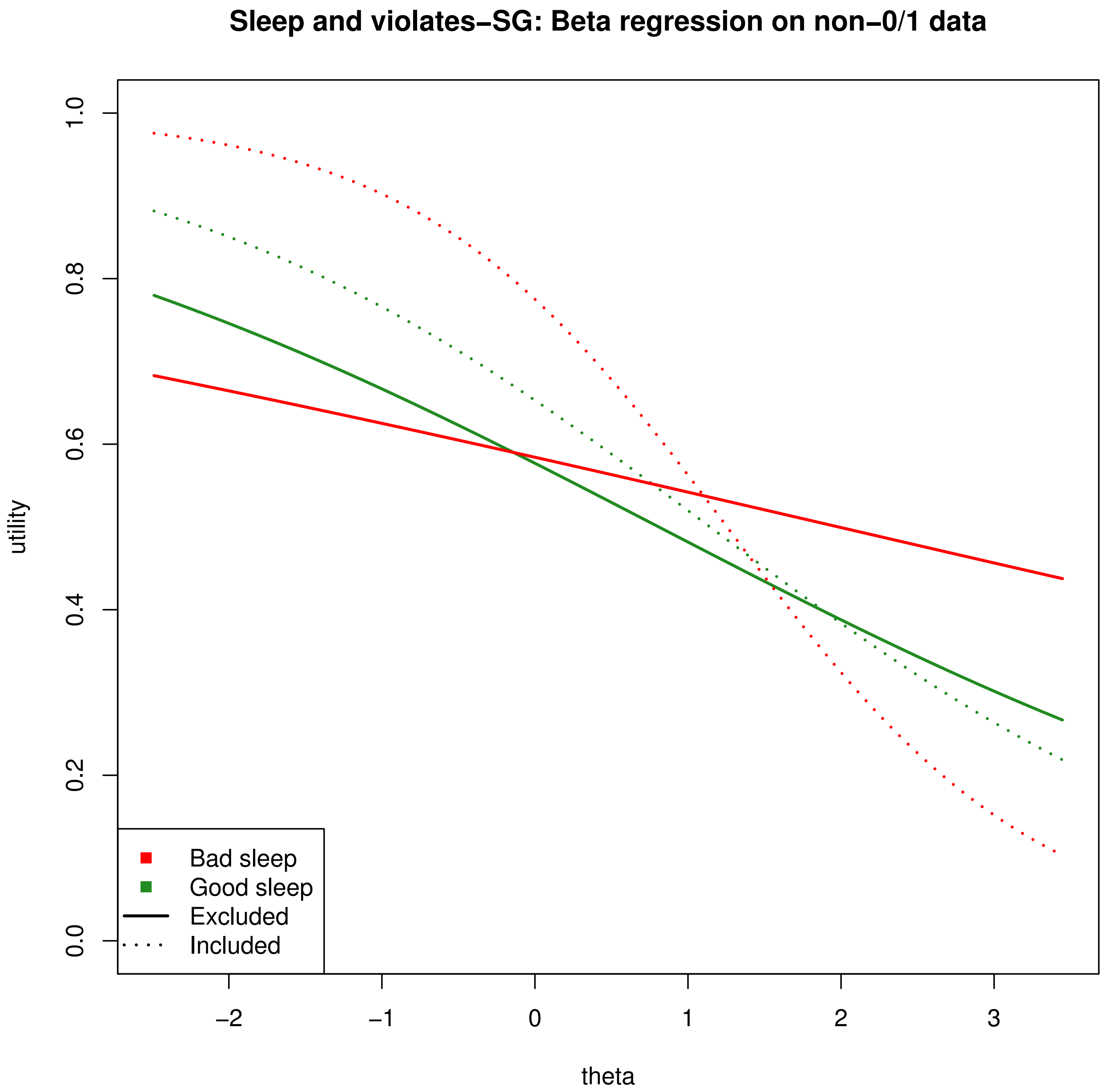


Figure 6.15: *Utility curves from beta part of ZOIB model for sleep with violates-SG, incorporating participant sleep health.* The curves show the conditional means from the beta portion of the ZOIB model of utilities for sleep as a function of *violates-SG*, theta, and participant sleep health. As it is the beta part of a ZOIB model, the model is estimated on the data without the 0s and 1s. These curves correspond to the model in Table 6.3.

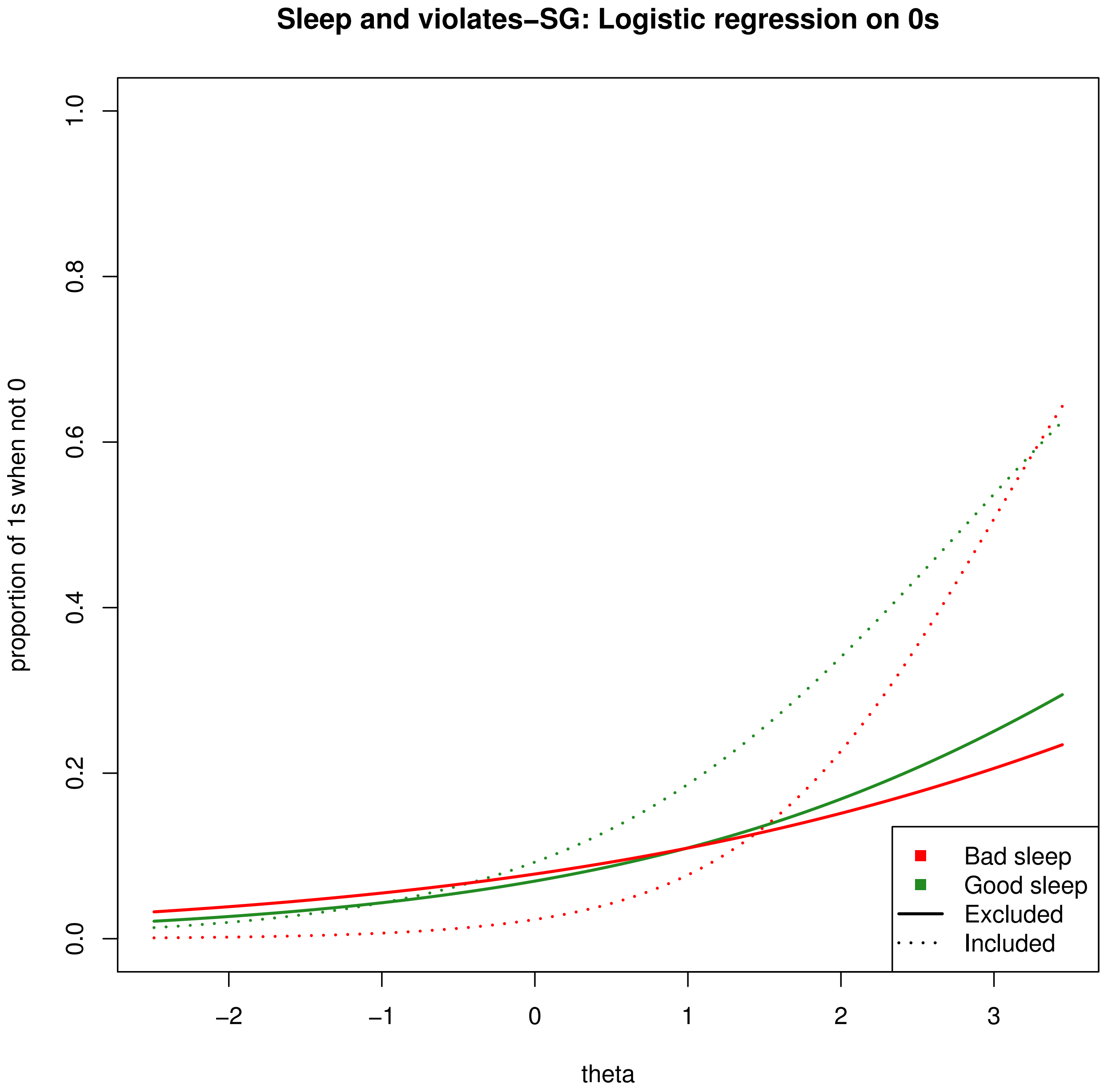


Figure 6.16: *Expected proportion of 0s from one logistic part of ZOIB model for sleep with violates-SG, incorporating participant sleep health.* Each curve shows the expected proportion of responses giving a utility of 0, for each of the four groups defined by exclusion and participant health.

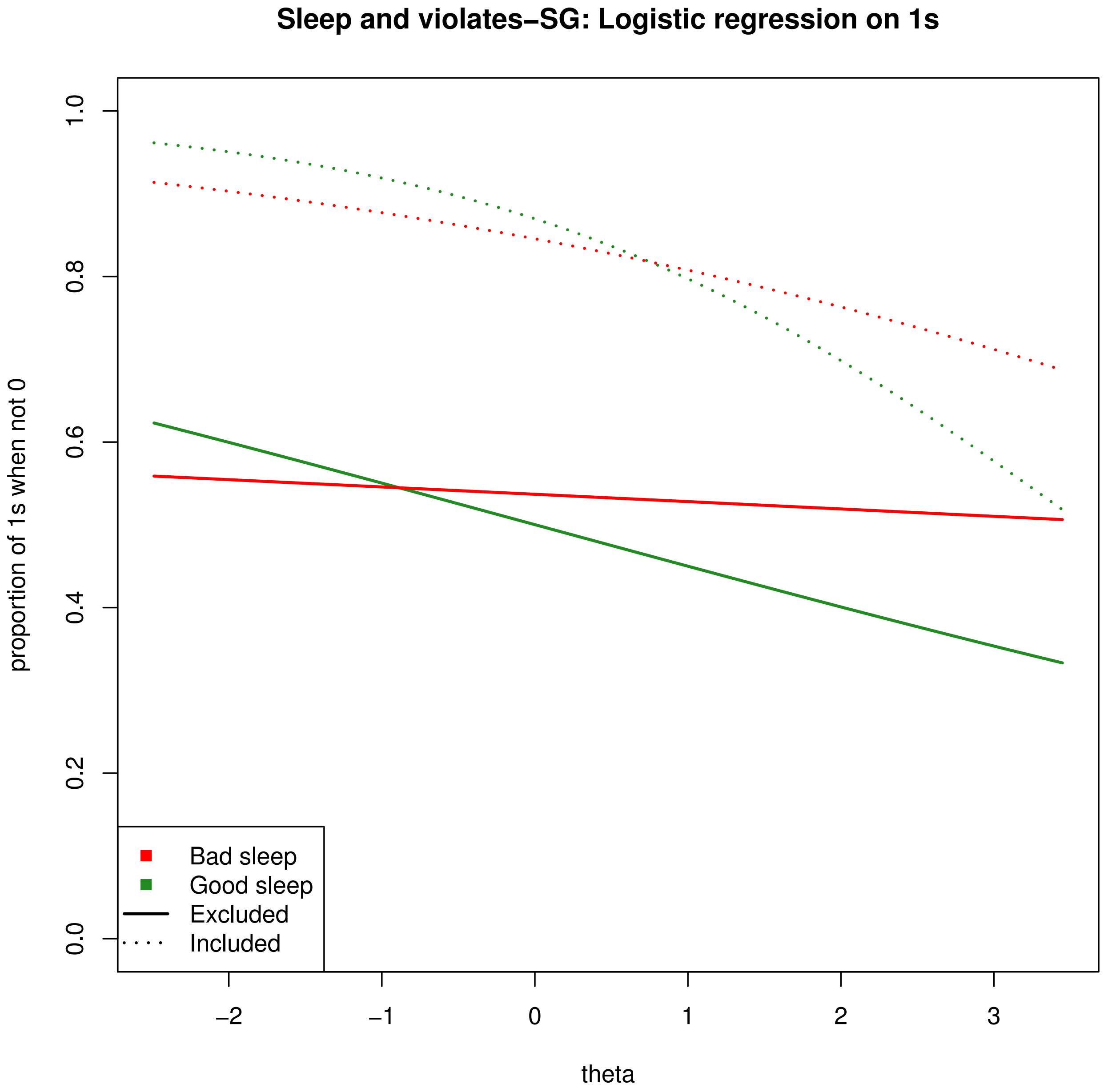


Figure 6.17: *Expected proportion of 1s from one logistic part of ZOIB model for sleep with violates-SG, incorporating participant sleep health*. Each curve shows the expected proportion of responses of giving a utility of 1, for each of the four groups defined by exclusion and participant health.

sleep states than are excluded ones, especially with respect to poor sleep states.[11]

Analogous ZOIB models can be estimated for every domain and criterion combination, although that task is left for future work. Figure 6.20 shows a second example, with a different domain (social functioning) and exclusion criterion (*dead-all-worst*). Of 1,169 responses, 170 are 0s, 563 are 1s, and 436 fall between 0 and 1. The beta regression on the non-0/1 data (first panel) shows the predicted utility curves of the excluded responses (solid lines) are in the opposite direction of the included ones (dotted lines). Among the included, those with poor health give higher utilities (dotted red) than those with good health (dotted green), with the opposite result among the excluded. For the expected proportion of 0s (second panel), all curves are in the expected direction, as in the previous example (i.e., fewer 0s as descriptions of health get better). There is less change in that expected proportion of 0s among the excluded responses, with an almost flat curve for those who are excluded with good health (solid green). For the expected proportion of 1s (third panel), the proportion of 1s is lower among the included, with the excluded in poor health (solid red) having the highest expected proportion – with that proportion *decreasing* as health states get *better*. For the other curves, the expected proportion increases as health states improve, as in the previous example (sleep, *violates-SG*).

Combined, these models produce the full ZOIB model for the social and *dead-all-worst* example (bottom panel). Overall, participants whose responses were excluded (solid lines) were insensitive to the value of theta that they were evaluating, with those reporting poorer social functioning assigning higher value to this domain, across the range. Participants who would be included, after applying this criterion, were sensitive to the health state that they were valuing, with those in poorer health assigning higher utility to poorer social health states.

In both examples (Figure 6.18 and Figure 6.21), participant health and the exclusion criteria interact, such that the differences between excluded and non-excluded groups change as a function of participant health. That means the exclusions have different implications when applied to participants with better and worse health. Moreover, the utility curves for those in poor health differ from those in good health, in particular among those who would be included. These have

[11] Figure 6.19 shows Figure 6.18 and its three constituent parts (Figure 6.15, Figure 6.16, and Figure 6.17) on one page, for ease of reference.

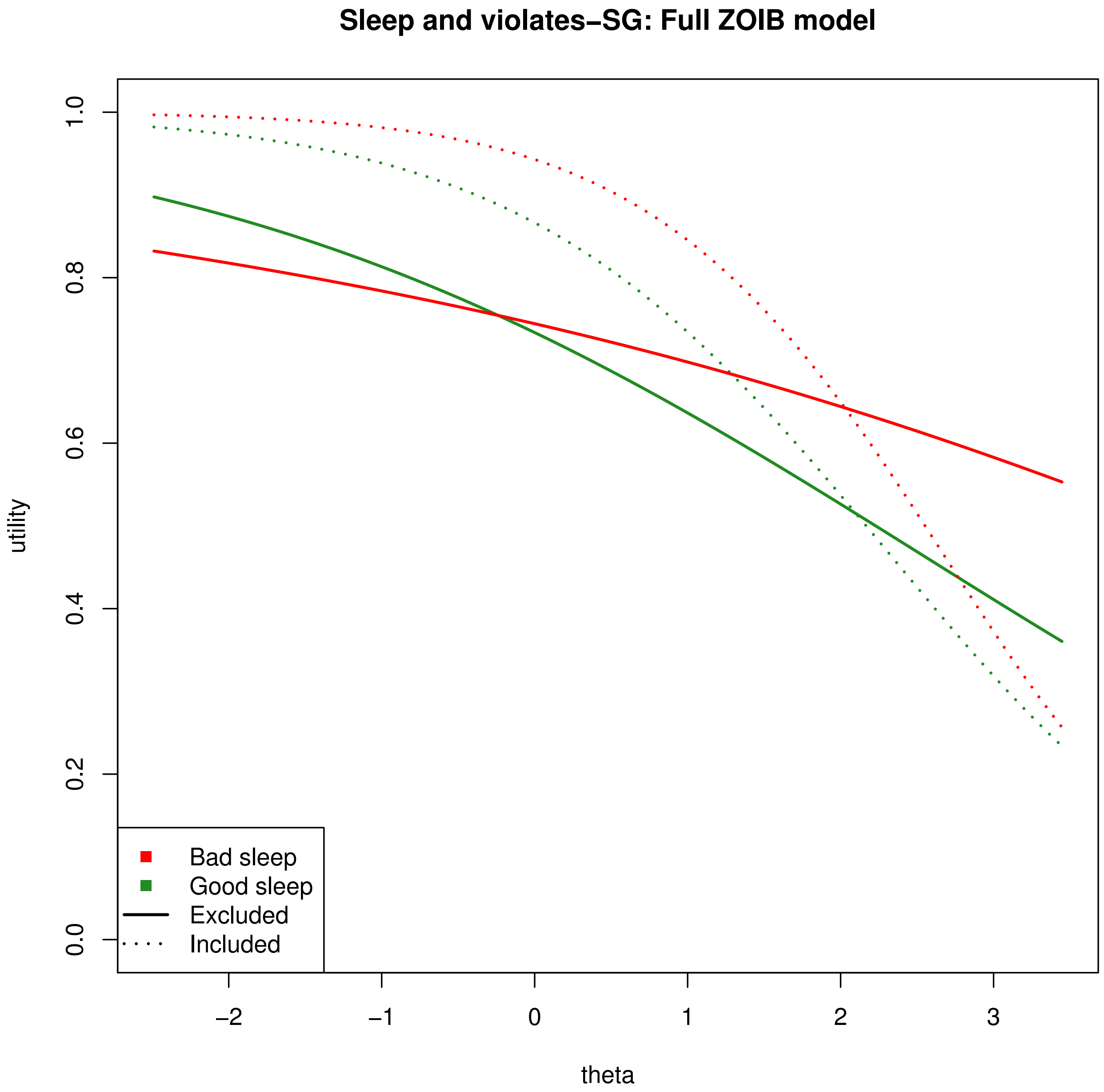


Figure 6.18: *Utility curves from ZOIB model for mean sleep utilities as a function of violates-SG, theta, and participant sleep health.* The figure shows the conditional means from the full ZOIB model for sleep utilities as a function of theta, *violates-SG*, and participant sleep health. It combines the models shown in Figure 6.15, Figure 6.16 and Figure 6.17, according to equation 6 in Liu & Kong (2015).

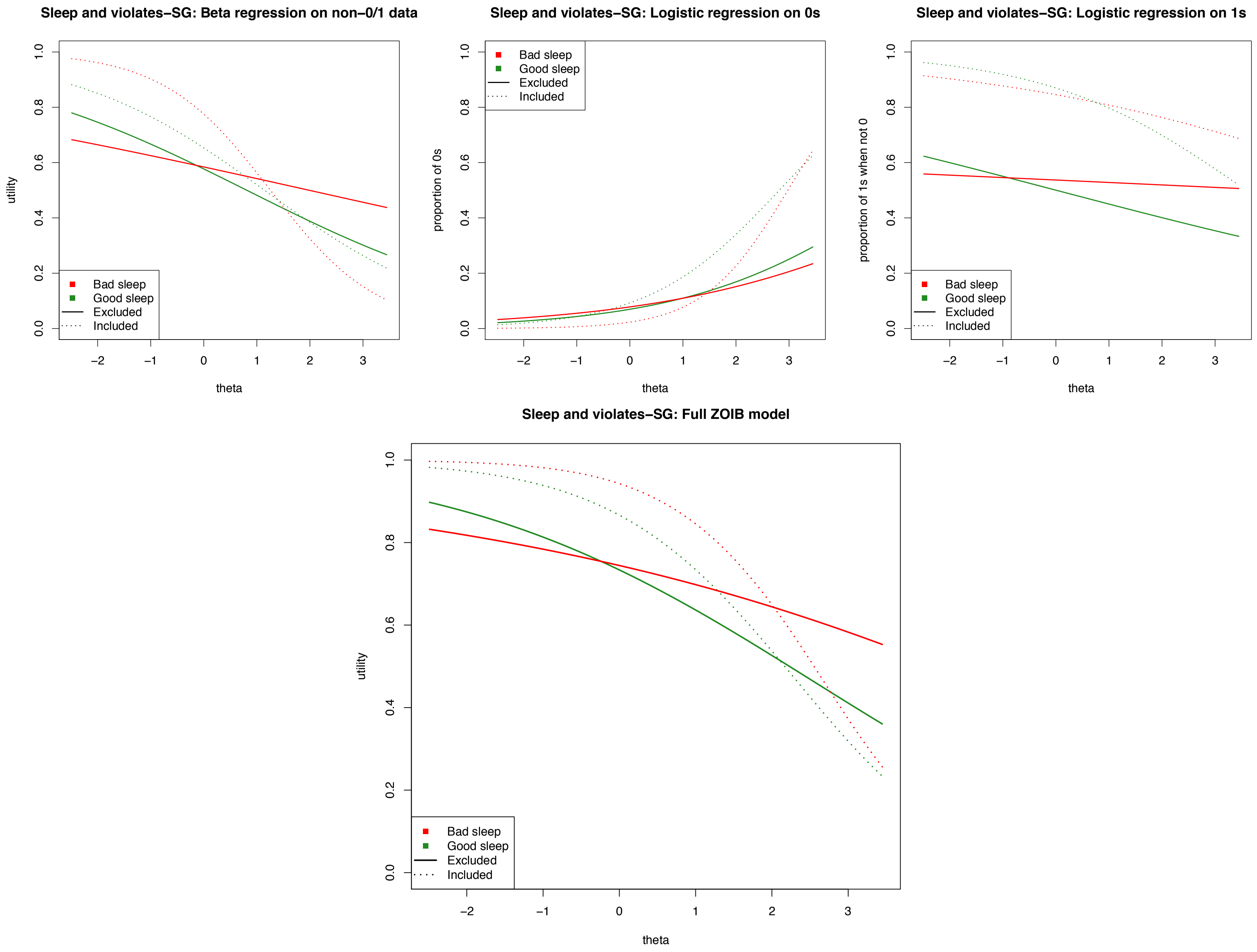


Figure 6.19: *All parts of the ZOIB model for sleep utilities as a function of violates-SG, theta, and participant sleep health.* This figure shows the constituent parts of the ZOIB model (Figure 6.15, Figure 6.16, and Figure 6.17), along with the final ZOIB model Figure 6.18.

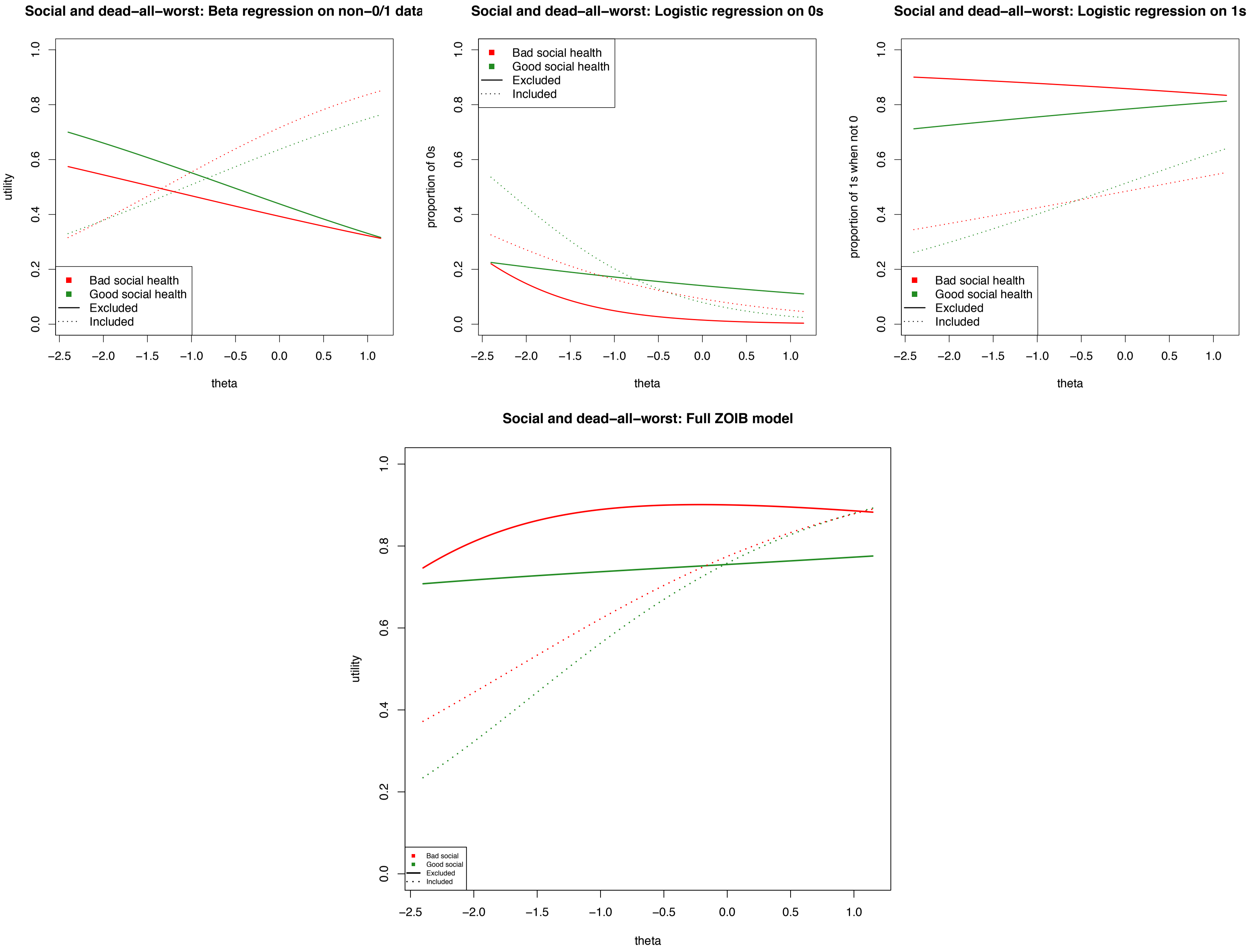


Figure 6.20: *All parts of the ZOIB model for social utilities as a function of dead-all-worst, theta, and participant social health.* This figure shows the constituent parts of the ZOIB model for mean social utilities as a function of the *dead-all-worst* criterion, theta, and participant social health. The top left shows the beta portion, estimated on the non-0/1 data; the upper centre shows the logistic regression on the 0s; the top right shows the logistic regression on the 1s; and, the largest figure shows the full ZOIB model, combining the beta part with the two logistic regressions.

implications for both the implementation of the criteria and their use in applied analyses, discussed below.

6.5 Beta regression sensitivity analysis and model-checking

To assess the quality of our models, we assess the functional form of equations (6.1) and (6.3), examine the residuals of the models, and use the zero-one inflated beta (ZOIB) models as an alternative (where they were not our main modeling strategy), examining whether our conclusions change using a more flexible model. We make heavy use of simulation in our assessment of the beta models (Davis, 2017a; Shalizi, n.d.). Details are included in the Appendix. In summary, we find that conclusions using the squeezed models are robust for estimating the effects of applying the exclusion criteria. The robustness of the resulting utility estimates requires additional analysis, as do the patterns revealed by more complicated ZOIB models.

Producing robust uncertainty estimates for our models requires completing two tasks. For the squeezed models, we need to perform simulations to ensure the standard errors produced (e.g., Table 6.1) are accurate despite the squeezing, and if not, estimate new values. For the ZOIB models, we need to either a) use the output of the **zoib** R package to produce uncertainty estimates (Liu & Kong, 2015), or b) develop our own estimates. The first strategy is computationally intensive (e.g., eight hours to estimate the sleep and *violates-SG* example), making it difficult to scale for all domain-criterion pairs, let alone all pairs while including the three-way interaction with participant health. The second strategy might then be the more appropriate, but if we cannot make analytical arguments, it could also require extensive simulation. In the interim, we use the size of the effects and the number of excluded (Table 5.5) as an indication of the uncertainty of the results.

6.6 Discussion of beta regression results

The purpose of this study was to see what practical difference exclusion criteria made on mean utility curves, in part, to see how important it was to resolve the ethical questions that the criteria raise. Furthermore, the responses are a reflection on the exclusion method as well as the participants, and thus also have implications for whether the criteria act as claimed.

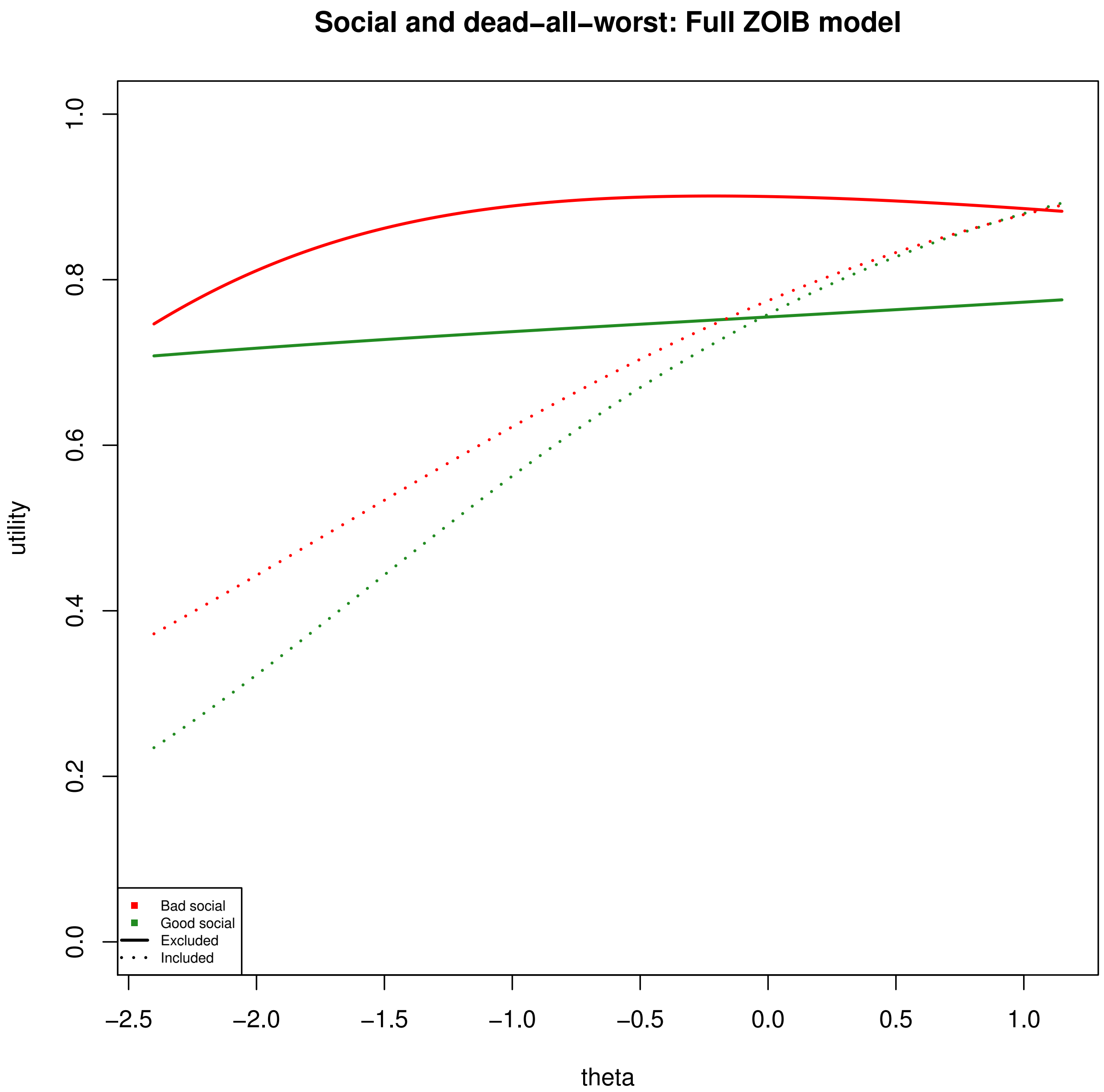


Figure 6.21: *Close-up of utility curves from ZOIB model for mean social utilities as a function of dead-all-worst, theta, and participant sleep health.* A close-up of the full ZOIB model from Figure 6.20.

6.6.1 Differences among utility curves

Utility curves can differ in two primary ways relevant to their use in health policy analyses: *elevation*, showing how much health states are valued, and their *sensitivity* (curvature), showing how much health states are differentiated. When those properties depend on participant health, that adds another dimension to their practical and ethical implications.

Differences in elevation

Changes in elevation matter if the severity of health states (i.e., functional capacities described by values of theta) matters. For example, some procedures (e.g., so-called "cost-value" analysis) (Nord, 2014) privilege treating states with lower utility, which are considered more severe. Higher utilities increase the cost-effectiveness of life-saving and life-extending interventions, by increasing the value of all (including poorer) health states, while decreasing the value of life-improving interventions, because the utility of poorer health states is already higher.

Differences in sensitivity

Changes in sensitivity matter because the steeper the slope (or curvature), the greater the change in utility for a fixed change in health status (theta). Thus, steeper slopes will increase the chances of treatments demonstrating their value.

Differences that are a function of participant health

The policy implications of changes in elevation and slope interact with participant health. Suppose those with poor health always gave higher utilities than those with good health, for all domains. Focusing on the preferences of those with poor health would then change the final multi-attribute utility function (Part II) in a predictable way: the relative utility of a multi-attribute disabled health state based on the utilities of those with poor health would be higher than the value calculated from the whole sample or from their healthy co-participants. That would *increase* the effectiveness of a life-saving or life-extending intervention for the disabled.

Differences in sensitivity when accounting for participant health could have similar implications. Suppose those in poor health show less sensitivity in their utility curve than their counterparts in good health. Then, treatments from which they could benefit become *less* effective from their perspective. That could be one argument for not relying on their preferences, if their adaptation to poor health leads to a lower probability of health systems buying treatments that could improve their quality of life. Conversely, these participants might be more motivated to answer attentively in the survey, given their experience with the survey topic.

Thus, how elevation and sensitivity change as a function of participant health affects whether interventions are more effective or less effective from the perspective of those who could immediately benefit from them. Deciding whether and how to weight the preferences of the healthy and the unhealthy is beyond the scope of our work, although it appears to be an instance of the meta-preference problems addressed in Part I (Menzel, Dolan, Richardson, & Olsen, 2002; Sanders et al., 2016; Versteegh & Brouwer, 2016).

6.6.2 Modeling utilities as a function of exclusion criteria

Figures 6.1-6.9 and Figure 6.13 show the results of estimating equation (6.1) for all nine[12] exclusion criteria for all seven domains. Figure 6.19 and Figure 6.20 show the results of incorporating participant health status for select domain-criterion combinations. We begin by discussing the worked examples from the previous sections, before turning to an overview of each criterion.

Sleep and social examples

In the example of sleep and *numeracy* (Table 6.1 and Figure 6.10), applying the exclusion criterion affected elevation, but not sensitivity. Those excluded (solid line) assign systematically higher utilities than those included (dotted line). One possible explanation is that the low numeracy criterion creates a pattern of exclusions similar to the *upper-tail* criterion. As explained in Box 1 of the MDS section, the implementation of the SG means that two out of the three options on the first screen of any SG lead to high utility values. This pattern would support the use of the *numeracy* criterion, as it excludes participants based on their (low) ability to express their preferences, not what preferences they have to express. Given the size of the exclusion effect and the number of

[12] Recall *no-variance* was not included in the modeling.

participants involved in the sleep domain (Table 5.5), the practical implication of applying *numeracy* turns out to be small (Figure 6.12), whatever one thinks about the principle.

In contrast, applying *violates-SG* in the sleep domain would exclude many responses (Table 5.5, Table 6.2 and Figure 6.11), have a larger effect on the utilities used to represent societal preferences, and make a strong normative statement, namely, that it excludes participants who are not providing their true preferences because those preferences do not conform to a certain structure (i.e., monotonicity in the correct direction). As seen in Figure 6.11, the utility curve for those who are excluded (solid line) is less sensitive than that for those who remain (dotted line). As a result, applying the criterion would favor interventions that improved sleep in more sensitive regions of the curve. As elsewhere, the legitimacy of that exclusion would depend on whether it removed participants who are not using the SG task properly (e.g., as a result of inattentiveness or carelessness) or participants who were thoughtfully expressing something other than preferences (e.g., not really caring, protesting).

Incorporating participant health in the modeling of sleep and *violates-SG* (Figure 6.18) revealed a larger difference between the excluded (solid line) and included (dotted line) when participants' sleep quality is poor (red) compared to when it is good (green). As an example of the implications of these results, consider a policy meant to meet the needs of those who sleep more poorly at times when they sleep the worst (high sleep theta). Excluding participants with violate-SG produces a steeper curve, thereby placing greater value on treatments that improve sleep in this range of theta. Focusing just on those who sleep more poorly, among those who remain (dotted red line), would have little effect on such decisions (given the similar shapes of those two curves). However, focusing on them would reduce the overall importance of sleep treatments, because their curve is higher for all values of theta. A substantive interpretation of the elevation effect is that it is caused by their adaptation to impaired sleep (Neumann et al., 2000; Peeters & Stiggelbout, 2010).

In the model of utilities for the social domain as a function of theta, the *dead-all-worst* criterion, and participant social health (Figure 6.21), those with poorer social health provide higher utilities than those with better social health, whether they are excluded or included (red above green, conditional on dotted or solid). The utilities from the excluded are also above those from the included for almost the entire range of theta, as well as being less sensitive to theta. That means removing them decreases the utility of social health states while increasing the sensitivity of the utility curve, making poor health states more severe and any change in social health more important than if the criterion were not implemented.

Decomposing the ZOIB model of this example (Figure 6.21) into its constituent parts (upper row of Figure 6.20), we see a stark result: in the beta regression on the non-0/1 responses (leftmost panel), those excluded by *dead-all-worst* are monotonic in the wrong direction, violating dominance. By comparison, those excluded by *violates-SG* are monotonic in the correct direction (leftmost panel of Figure 6.19). Furthermore, those excluded by *dead-all-worst* provide fewer 1s than their included counterparts (solid above dotted in the rightmost panel of Figure 6.20). In contrast, those excluded by *violates-SG* are providing *fewer* 1s than their included counterparts (solid below dotted in the rightmost panel of Figure 6.19). Assuming that high frequencies of 1s are indicative of non-preference SG responses (e.g., Box 1), that pattern could support the claim that *dead-all-worst* removes non-preference data, and that *violates-SG* might not (as argued in the MDS section). In addition, when flagging those who rate dead/all-worst as equal-to-or-better-than full health, *dead-all-worst* captures a group with *overall* preferences that violate dominance, whereas *violates-SG* does not. We need to repeat the analysis, implementing the criteria for every domain, to generalize these results.

Although we lose model parsimony through the ZOIB approach, these worked examples demonstrate the advantage of analyzing the 0s and 1s separately. Those two values appear to be of special significance among all the possible SG responses. In fact, 0 and 1 are the two most frequent responses among all the SGs (with 1 being the most frequent).

Next, we discuss the global results from all the criteria and domain combinations estimated using equation (6.1) and displayed in Figures 6.1-6.9.

6.6.3 Discussion of each criterion

The practical implications of applying exclusion criteria depend on how they affect the utility curves used to represent societal preferences. Those that reduce the elevation for a health state could increase the resources devoted to helping those at that state. Those that increase the steepness at a health state also increase the resources devoted to changes near that state. The size of those effects is always an empirical question. Even when the effects of applying an exclusion criterion are qualitatively predictable, how widely it will apply is not. For example, applying the *low-range* exclusion criterion (or the *no-variance* criterion) will necessarily leave a steeper utility curve. The size of that change will, however, depend on how many people use less than 10% of the range. The effect on elevation cannot be predicted at all.

Here, we consider the patterns that emerge with each of the criteria, across the seven domains. We begin with the non-preference based criteria.

Numeracy

The mean utility curve for those excluded by *numeracy* (Figure 6.1) is higher in all domains except for depression and physical functioning. That result is consistent with the MDS finding that *numeracy* and *upper-tail* are relatively close, indicating that numeracy tends to exclude high responses. The higher utilities of those excluded by numeracy supports the conjecture (Box 1) about how the first screen on an SG may direct confused participants to high values. There is little discernible pattern in the slopes. Given the small difference in the curve and the few participants affected (7.8%) (Table 5.4), applying the numeracy criterion would make little practical difference.

Time

The mean utility curve for those excluded by *time* (Figure 6.2) shows no consistent difference in elevation across the domains from the curves of the included. In contrast, there is a distinct pattern of extremely low sensitivity among the excluded, with most lines being close to horizontal. That could be interpreted as support for the criterion, if it indicates inattentiveness. The drastic insensitivity of the excluded across the entire range of theta for each domain, combined with the moderate number of participants (15.0%) excluded by *time* (Table 5.4), means removing those participants would increase the sensitivity of the societal utility curve and affect calculations involving any health state.

Understanding

Among those excluded by *understanding* (Figure 6.3), only fatigue shows a systematic difference in elevation of the utility curves, with those excluded giving higher utilities on average. There is diminished sensitivity to theta throughout; however, the effect is small for cognition, fatigue, and pain. The rationale for *understanding* is that it captures those who admit that they could not use the survey tool to provide their preferences (Table 4.1). However, our results suggest that this might not be the case: the curves of the excluded and included for cognition, depression, fatigue, and pain are close in shape. If *understanding* removed those who could not express themselves, it is difficult to understand how they would provide utilities that are so similar to those of their counterparts who rated their self-assessed understanding higher. Rather, one might expect something completely nonsensical (e.g., monotonic in the wrong direction), or perhaps even a constant average. Thus, it could be that this *criterion* is capturing thoughtful individuals, who, upon reflecting on an abstract and unusual task, underestimate the quality of their data. Although 14.3% of participants would be excluded by *understanding* (Table 5.4), the closeness of the excluded to the included curves mean the effect would be small, except for fatigue, where the societal utility curve would lower in elevation, making each fatigue state more severe.

Low-range and *no-variance*

For every domain, the mean utility curve for those excluded by *low-range* (Figure 6.4) is higher for most (and sometimes all) of the entire range of elicited theta values. That pattern – along with the necessarily insensitive curves of the excluded – could be caused by the ease of providing high responses, as discussed earlier in Box 1. The removal of 12.2% of the sample via *low-range* (Table 5.4) means that, if one were to apply the criterion, there would be a nontrivial increase in the sensitivity of the resulting utility curve and a decrease in the elevation of that curve, across the whole range of theta. The effects of implementing *low-range*'s nested criterion *no-variance* would be similar.

Dead-all-worst

Those excluded by *dead-all-worst* (Figure 6.5) have systematically higher utilities overall (higher elevation), and show diminished sensitivity to theta. That means those who rated dead or the all-worst state with a utility of 1 or greater were also prone to give high utilities for the other states they valued. *Consistently* providing high utilities could indicate an inability to use the SG to communicate one's preferences (e.g., Box 1), and necessarily leads to low sensitivity. It could be that making an extremely large violation of dominance, as in *dead-all-worst* – where dead or the all-worst state is set equal to or better than full health – is a signpost for non-preference responses. More than a quarter of the sample (28.0%) is flagged by *dead-all-worst*, meaning that implementing it would increase the sensitivity of the societal utilities and decrease their values, producing larger changes in utility for a given change in health status and increasing the severity (lowering the utility) of all states.

Violates-SG* and *violates-VAS

Violates-SG and *violates-VAS* (Figure 6.6 and Figure 6.7), show similar patterns of differences between their respective excluded and included, which are more pronounced for *violates-SG*. There is not a consistent elevation effect, and the included show much more curvature than the excluded, usually rating the good health states with much higher utilities. Of note is that the *mean* utility curves of the excluded do *not* violate dominance.[13] The large number of participants excluded by *violates-SG* (71.6%) and *violates-VAS* (84.7%) implies that the societal utilities would change drastically if either one were implemented. That effect would, for many domains, be less with *violates-VAS*, despite the larger number excluded by that criterion, because of the smaller differences between the excluded and included curves.

[13] The utility curves are monotonic by construction because they are linear in theta, although they could be monotonic in the wrong direction, which would violate dominance (as in the top-left panel of Figure 6.20). However, the factor models have no such restrictions. These show only a handful of violations, e.g., for depression among the excluded in *violates-SG*, or among the non-excluded for physical functioning in *violates-VAS*.

***Upper-tail* and *lower-tail* (collectively *10% trimming*)**

The mean utility curves for those excluded by *upper-tail* and *lower-tail* show the elevation effects that are expected. If we were to apply *upper-tail* or *lower-tail* by removing all the responses of those flagged – analogous to the implementation of the other criteria – it would produce a variety of other effects, as seen in Figure 6.8 and Figure 6.9. The elevation effects are predictable, but some domains show large sensitivity effects as well (e.g., much lower sensitivity among those excluded by *lower-tail* in the fatigue domain). However, that type of implementation would not be the usual practice. Rather, convention is to implement them together via the criterion of *10% trimming* (Table 4.1), in which more than 10% of the sample might have some response flagged by the criterion (as in Table 5.4), but only 10% of responses would be removed at every health state. The large number of 1s and 0s, discussed in the MDS chapter, means that *10% trimming* would hardly change the resulting utility curve, but would change the variance of its estimates.

6.7 Conclusion

The effects of exclusion criteria can be organized into two categories: increasing or decreasing the elevation of the societal utility curve and increasing the sensitivity (curvature) of the societal utility curve. Elevation and sensitivity effects are observed for two of the three non-preference-based criteria (*numeracy* and *time*). For *numeracy*, the higher elevation and lower sensitivity among the excluded could be caused by low numeracy leading to difficulty using the SG, which requires an understanding of probability (Box 1). For *time*, rushing through the survey could lead to producing the same pattern of choices for every state, with a propensity for higher values given the mechanics of the SG. In contrast, it appears that the preferences of those excluded by *understanding* are not much different than their included counterparts, possibly indicating that those excluded participants are providing thoughtful responses.

Among the preference-based criteria (Table 4.1 and Table 5.3), we saw diminished sensitivity and higher utilities among those excluded by *dead-all-worst*. That could indicate it is removing responses that are not true preferences, though, as elsewhere, that depends on whether insensitivity to theta is evidence of inattention or inability to use the SG. We saw less sensitive utility curves among those excluded by *violates-SG* and *violates-VAS*, although given the focus on mean utilities, it is unclear whether it is worth trusting these criteria if the excluded, on average, do *not* violate dominance, and both criteria exclude more than half the sample (Table 5.4). Incorporating health status showed an example of *dead-all-worst* (Figure 6.20) capturing responses that are difficult to explain as true preferences – the excluded violated dominance *as a group* and provided more 1s than the included. In contrast, among the responses captured by *violates-SG* in the sleep example (Figure 6.19), the excluded were monotonic in the correct direction and provided fewer 1s than the included. Fully unpacking those results requires extending our ZOIB analysis to other domains.

The changes in elevation and sensitivity observed in this section are important if they affect analyses that depend on the utilities, such as cost-effectiveness analyses. The interaction between criteria and the health of participants is important for similar reasons; differential effects of criteria on healthy and unhealthy samples will affect the legitimacy of analyses that (do not) focus on the latter group. All of the results of this section also impact the representativeness of the *societal* utility estimates, if there is the potential for wrongful exclusions – excluding responses that represent true preferences. Next, we focus on these topics.

7

Policy Implications

7.1 Summary of the MDS analysis and the modeling of utilities

Our analyses in the previous two sections have focused on two tasks: uncovering the similarities between exclusion criteria's classifications of survey participants, and determining how the preferences of the excluded and non-excluded differ.

The first used multidimensional scaling (MDS) to produce a map (Figure 5.1) of 10 exclusion criteria (Table 5.3). The proximity of exclusion criteria on the map is a proxy for the similarity of their exclusion decisions, which are a function of how many of the same participants they exclude (and do not exclude). The criteria are divided into *preference-based* and *non-preference-based* criteria (Table 5.2), depending on whether they are defined using the preference elicitation tasks from the PROPr survey. The two sets show distinct types of relationships. The three non-preference-based criteria – *numeracy*, *time*, and *understanding* – form a cluster, suggesting that they represent a common construct. Each reflects a somewhat different mechanism related to struggling with the task: low numeracy leading to poor use of the SG, rushing leading to inattentiveness, and admitted inability to use the SG.

In contrast, the seven preference-based criteria are dispersed in the space (Figure 5.1), suggesting that they represent different constructs. The most dramatic of those might be *upper-tail* and *lower-tail* (Table 5.3), which fall far apart, despite being routinely combined by analysts in the *10% trimming* criterion (Table 4.1). Their neighbors suggest features that might characterize the responses that each excludes. *Upper-tail*, *low-range* and *no-variance* are close to *numeracy*, which we conjecture reflects an artefact of the SG procedure for those not fully comfortable with it: As detailed in Box 1, the mechanics of that procedure mean that a likely mistake on the first screen of the SG will produce utilities leading to exclusion by possibly all three of these preference-based criteria – depending on which states the mistakes are made – by virtue of producing the highest possible responses. The axis with *no-variance* at one end and *violates-SG* at the other could indicate a "precision" dimension, as the easiest way *not* to violate dominance in the SG is to provide identical (i.e., no variance) responses. Thus, those who violated the SG were at least trying to distinguish among the health states.

The MDS revealed that not all criteria have equal empirical support for their use. Ten-percent trimming might be the archetype: split into its constituent parts, *upper-tail* and *lower-tail*, we saw that participants who get flagged by one are not the same as those who get flagged by the latter, despite a common rationale. An exclusion criterion lacking empirical support would, at best, exclude the responses of those who do not use the SG to communicate their preferences – but possibly for reasons that do not align with the criterion's stated rationale – and, at worst, exclude those whose responses are true preferences. Both would undermine the quality of conventional practice, and the second would be particularly pernicious, given that it would weaken the claim to the "societal perspective" taken by most analysts using the data left after exclusion. However, these concerns could be trivial if the preferences of the excluded and the included do not differ in ways that affect analyses.

The MDS results model the criteria in terms of how they flag participants for exclusion. In contrast, by modeling utilities as a function of the exclusion criteria, we can learn whether they act on preferences as claimed, and, when there is a difference between the excluded and included, determine if that difference has practical or ethical content for someone who would want to use the criterion. The beta regression models from the second section of this chapter (Figures 6.1-6.9) showed two patterns that were true to varying degrees with almost every exclusion criterion: those excluded were less sensitive to theta and they provide systematically higher or lower utilities than those included. Applying *upper-tail*, *numeracy*, *dead-all-worst*, *low-range*, and *time* generally results in higher utilities for the excluded compared to the included; applying *lower-tail*, *violates-SG*, and *violates-VAS* generally results in lower utilities. Those excluded by *low-range*, *numeracy*, *dead-all-worst*, *time*, and to a lesser extent *understanding*, *upper-tail*, *violates-SG* and *violates-VAS*, are less sensitive to theta than their included counterparts.

Extracting the implications of these results requires some interpretation. For example, those excluded by *time* provide nearly constant utility curves, a pattern that suggests inattentiveness, hence the validity of applying that exclusion criterion. In principle, that pattern could reflect participants who are truly indifferent to theta and quickly express those preferences. However, the proximity of *time* to the other non-preference based criteria in the MDS supports the former interpretation. As a second example, those excluded by *understanding* had very similar utility curves to those not excluded, suggesting that those judgments captured how they felt about their performance, rather than their actual ability.

An additional clue to interpreting exclusion criteria is provided by participants' health on a domain. In the focal example of sleep and *violates-SG* (Figure 6.11), those excluded are less sensitive to theta than those included. We then incorporated the participants' own sleep quality into that model (Figure 6.18). We focused on sleep because the distribution of participant sleep PROMIS scores is the most symmetric, whereas the other distributions are skewed towards good health. We chose *violates-SG* because it has the largest effect on sleep out of any criterion when applied to the sleep domain (Figure 6.12). Its isolation in the MDS plot (Figure 5.1) also made it a good candidate for further investigation. Our analysis showed that the difference between the excluded and included was larger for those who sleep poorly compared to those who sleep well, while those excluded who sleep poorly are less sensitive to theta than their counterparts who sleep well, and those who are included and sleep poorly show similar sensitivity to their counterparts who sleep well, but have a higher elevation. Thus, implementing the exclusion criterion would lead to utilities for sleep that varied more over the range of sleep states than if the whole sample were used, while doing so and over-weighting those who sleep poorly would increase the utility of every sleep state.

A second example for the social domain and *dead-all-worst* revealed similar results and implications (Figure 6.21). There were two notable exceptions. One was that those with poor social health assigned utilities above their healthier counterparts, whether included – as in sleep and *violates-SG* – or excluded. The second was that those included with good health had more sensitivity to theta than their counterparts with poor health for low values of theta (states of poor social health). Thus, implementing the exclusion criterion would lead to utilities for social health that varied more over the range of social health states than if the whole sample were used, while doing so and over-weighting those with poor social health would increase the utility of every sleep state, but slightly decrease the sensitivity of the utility curve.

Above, and throughout the last two sections of Part III, we have touched on some of the policy implications of our results. In this section, we focus on them.

7.2 Analyses requiring HRQL estimates

In cost-effectiveness analyses (CEA), population health studies, and decision analyses, societal preference-based (i.e., utility-based) measures of HRQL are often used when comparing or tracking the health of different groups.

As discussed earlier and in Part II, the preference data analyzed here are the source for building the PROPr scoring system, a new societal preference-based HRQL tool. Part II presented sensitivity analyses, assessing the effects on the multi-attribute scoring (utility) function of applying some of the exclusion criteria studied here. Those analyses were constrained by multi-attribute utility theory, which requires estimating models for utility curves with properties that can conceal the patterns in the data: fixed endpoints and monotonicity. In contrast, our use of beta regression allowed estimation of the best-fit curve with the only constraints coming from the data and the assumptions of the beta model – which are less restrictive than those from multi-attribute utility theory.

In that vein, this section explores the practical implications of our findings from the previous sections. We illustrate our approach with single-domain examples, although it could be extended to multi-attribute ones.

A primary concern of CEA is the incremental cost-effectiveness ratio (ICER), the difference between the cost of two alternatives divided by the difference in their effectiveness. One way to define the latter is the change in HRQL. That can be calculated by measuring the change in health status (e.g., PROMIS score), using an instrument such as the PROPr scoring system. Then, one can use that value along with PROPr's utility functions to estimate the change in quality-adjusted life-years (QALYs) associated with the difference in treatment. The ICER is given in units of dollars/QALYs gained (Neumann et al., 2016).

Many organizations, such as the UK's National Health Service (NHS), the World Health Organization (WHO), and the Bill and Melinda Gates Foundation have cost-effectiveness thresholds: an ICER below which an intervention is considered sufficiently cost-effective, hence a candidate for reimbursement, and above which the intervention might not be offered. For example, the NHS has historically used an ICER of around £20,000/QALY (McCabe, Claxton, & Culyer, 2008).

To explore the possible effects of the exclusions on an analysis using societal preference-based HRQL, we consider a hypothetical CEA for Lunesta (*eszopiclone*), a sleep medication for sufferers of insomnia. CEAs for Lunesta have been completed (Botteman, 2009; Botteman et al., 2007; Snedecor et al., 2009), and take a multi-attribute approach to HRQL – although, to the best of our knowledge, none have used HRQL measures that assess sleep disturbance with the precision of PROMIS's sleep disturbance scale. PROMIS's scale is, indeed, particularly suited for measuring sleep symptoms in the context of insomnia (Yu et al., 2011).

Consider a hypothetical CEA of Lunesta that uses PROMIS's sleep disturbance scale, which corresponds to the sleep domain in PROPr, to measure sleep-related outcomes. Exclusion criteria change two features of the utility curve: its elevation and sensitivity.

Ceteris paribus, the elevation of a utility curve will not affect the calculation of the ICER, or anything else that relies only on a *change* in utility. However, elevation would matter if, for example, a health system only paid to treat more severe disease states, or used it to choose among interventions with the same ICERs (Dolan et al., 2005; Nord, 2014).

If a criterion affects the sensitivity (slope or curvature) of the utility curve, then it will change how difficult it is to meet a given cost-effectiveness threshold. The steeper the slope, the greater the change in utility for a given change in theta. For example, the curvature of the utility curves for sleep under no exclusion and under *violates-SG* are roughly the same for lower values of theta (< 0), which correspond to better sleep (Figure 6.11). For higher values, indicating poorer sleep, the utility curve when implementing the *violates-SG* criterion decreases more quickly. Thus, using *violates-SG* to remove participants will increase the value of any improvement in sleep quality when the initial sleep quality of those treated is low (i.e., a positive sleep theta value).

Figure 6.18 shows how these issues manifest when we also consider participants' PROMIS sleep scores. The difference between the included and excluded is larger for those who sleep poorly than those who sleep well. Among those excluded, the curve for those who sleep poorly is shallower than the curve for those who sleep well, and both are shallower than the curves for those included. Among the included, those who sleep poorly value every state more highly than their included counterparts. The curvature of the curve for those included who sleep poorly is nearly the same as that of their counterparts who sleep well. Thus, a treatment for poor sleep would be more cost-effective if the exclusion criterion is applied, in which case the severity of any initial state of sleep would be greater if extra weight were given to the preferences of those who *already* sleep well.

Note that a similar pattern with similar policy implications emerged in our second example, involving the three-way interaction model for social health utilities as a function of theta, participant PROMIS scores on the social domain, and the *dead-all-worst* criterion (Figure 6.21). One exception is that, although the utility curves for the included are similar whether social health is good or poor, the differences in curvature that do exist are small but more pronounced than in the sleep and *violates-SG* example (Figure 6.18). Thus, treatments that make improvements for those with dysfunctional social health would be slightly more cost-effective using the curve from those included with good social health than from the perspective of those included with poor social health. That result, and the final one from the previous paragraph, are variations on a theme: using the utilities of those who could benefit immediately from a treatment can make that treatment appear less attractive than using the utilities assigned by their healthy counterparts. That is one reason some argue against relying on the utilities of the unhealthy, despite their increased experience with illness.

As mentioned in an earlier section, the higher elevation of the utility curves of those with poor health could be a result of adaptation to poor health (Versteegh & Brouwer, 2016). Regardless of the mechanism, the implications of this elevation are well-known (Menzel et al., 2002). As described above, if two treatments were being considered and the severity of the initial state was used as an input to the decision, using the preferences of those with poor health – the very people who would stand to benefit from the treatment – might result in that treatment not being purchased. In more concrete terms, those with poor sleep quality would give a *higher* utility to a description of the symptoms of insomnia, which could lead to a decision *not* to purchase a treatment such as Lunesta, if that utility was considered too high.

In contrast, the utility of extending the lives of those who sleep poorly would be *higher* if using the utility curves of those included with poor health, because of their greater elevation. Thus, the effects of using the utility curves of the unhealthy compared to their healthy co-participants are not uniform; i.e., the systematically higher utilities of the unhealthy will not always disadvantage them (Versteegh & Brouwer, 2016). Rather, the policy effect depends on the type of intervention (e.g., quality-of-life improving, life-saving, life-extending) and what is deemed important to the analysis (e.g., the ICER, the severity of the initial health state).

7.3 Implementing exclusion criteria: Social choice theory

We have described some of the effects exclusion criteria can have on analyses that depend on the utilities those criteria are used to define. One question that we cannot completely resolve with our data is whether the diminished sensitivity to theta exhibited by many of the excluded should be interpreted as evidence of non-preference responses, or if it is plausible that the criteria are capturing representatives of a subpopulation that truly discriminates less between health states.

If the near-constant mean utilities of those excluded is evidence of their inattention or inability to use the SG, then that supports their removal. If we doubt that interpretation, it can still be legitimate to remove those participants. We describe how in this section: one could legitimize their exclusion by appealing to the normative underpinnings of the utility function, using social choice theory, following the approach of Part I.

Social choice theory addresses insensitive utility curves with an axiom called the *elimination of the influence of indifferent individuals* (or *separability*) (Roberts, 1980). Our discussion in Part I did not invoke this axiom, but one could choose to do so: it says to ignore those individuals who have constant utility across all the alternatives (Deschamps & Gevers, 1978).

An important note is that the definition of separability only applies in its usual description in the context of the *no-variance* exclusion criterion, as, by definition, that criterion captures *exactly* those who have constant utility across all the alternatives. In contrast, *low-range* is defined to capture the same participants as *no-variance* as well as those who have *near* constant utility. Finally, we have seen that other criteria – such as *time* – appear to capture participants who, on average, have nearly constant utility. The social choice framework can still be useful for criteria other than *no-variance* – especially given the emphasis on the person-mean approach in the HRQL literature (see Part II) – if we view there being two types of participants: those with a utility function described by the mean curve of the included, and those defined by the mean curve of the excluded participants. Then, the question becomes how to combine these two groups.

Suppose that we did *not* think that separability should hold. Then, the final ranking of the health states could depend on the constant utility of the group of indifferent individuals. Our analysis from Part I would not change, and we would have the full set of aggregation procedures (societal utility functions) described there at our disposable (Roberts, 1980). Thus, we could use the mean, or a weighted average of the mean and the standard deviation, or a weighted average of the mean and the minimum utility, or of the mean and the maximum utility, etc. That is, we would be free to incorporate the many ethical principles discussed in Part I.

Suppose that we *did* think separability should hold. It turns out that this severely constrains the options for societal utility functions in our context: our only options are utilitarian procedures, such as the mean, and the lexicographic extension of the maximin and maximax rules (Deschamps & Gevers, 1978). Thus, the conventional (mean) form of the utility function would be permissible, but it would be impossible to incorporate relevant social values – for example, operationalizing equity via the standard deviation or the Gini coefficient – into the utility function in a way that was normatively grounded, unlike in the base case explored in Part I.

The conclusion, then, is that we could appeal to social choice theory to implement exclusion criteria that capture groups of constant mean utility, a characteristic we have discovered is common among the criteria. But, we would lose the ability to apply most ethical principles into our utility calculations. Thus, in the absence of a thorough understanding of whether the criteria are capturing participants whose responses are not truthful representations of their preferences, one can appeal to social choice theory to validate applying some of the criteria: but, in doing so, one loses the ability to undertake a large class of sensitivity analyses regarding the ethical implications of the functional form of the societal utility function. Said differently, most of the design options presented in Part I would no longer be possible. We would be forced to accept conventional preference aggregation as a consequence of resigning ourselves to conventional preference measurement and its approach to exclusion.

7.4 Implementing exclusion criteria: Altering convention

Current convention likely produces many false alarms – wrongly excluding those whose responses are true preferences – because it trusts that every exclusion criterion can make a correct exclusion decision, but only all of them together can make a correct *inclusion* decision. Even without the data necessary to calculate the accuracy of the (non-)exclusion classifications, we can provide some preliminary advice for the implementation of criteria that considers some of the ethical implications of our results and of current practice.

Excluding a response when a single criterion excludes, and including only if all criteria include, implies that including a response that is not a true utility is much more dangerous than wrongfully excluding one that is. Given the emphasis on the *societal* perspective in HRQL and CEA (Sanders et al., 2016), this seems counter to the prevailing goal of producing representative utility estimates.

One way to alter the use of criteria in order to have fewer false alarms is to change their implementation: if a set of criteria is imposed, make it more difficult for a participant to be excluded, by choosing "exclude" only if a participant is excluded by several criteria. That forces the researcher to consider whether any of their criteria are independent (e.g., *violates-SG* and *no-variance*), which preclude common exclusions, and to consider estimating the rate of correct exclusions of each criterion. Requiring several criteria to agree on the exclusion is tantamount to applying the conceptual framework of our MDS analysis to the design of the survey: one should be more confident excluding a participant flagged by multiple criteria that do not depend on the exact same characteristics than an exclusion of a participant flagged by only one criterion. This is exactly what we saw in our MDS results: the three non-preference based criteria excluded similar people, despite being defined by different parts of the survey, providing a kind of validity in their categorizations.

Another approach is to define *inclusion* explicitly. Although, formally, exclusion criteria perform this role – the complement of each exclusion criterion defines an inclusion criterion – there is reason to think that, in practice, producing a list of exclusion criteria and one of inclusion criteria would result in different final samples. To the best of our knowledge, there is no convention about whether exclusion criteria are chosen *ex ante* or *post hoc* (Engel et al., 2016). There is evidence that some choices of exclusion criteria probably occur after data collection: in a survey of NIH-funded researchers, Martinson et al. (2005) found that 15.3% dropped observations based on a "gut" feeling of them having been produced in error. The literature on evaluating evidence *post hoc* and thinking up ways that the data might look *ex ante* suggests that the two processes are not the same (Davis & Fischhoff, 2014). For example, Davis & Fischhoff (2014) find that an interpretation based on error (i.e., that the data are not representative of the phenomenon being measured) is common when evaluating surprising results (*post hoc*), but less common when evaluating expected results (i.e., considering how those results could have been produced by error). In our context, that could mean a researcher is more likely to think of a way that an unexpected set of utilities represents a participant who needs to be excluded than the researcher is to think of the ways a "normal" set of utilities could have been produced by an inattentive participant. Without a deeper understanding of the behavioral aspects of the SG, that could result in excluding many who should not be – as well as including many who are not providing their preferences, but are conforming to expectations.

Generating hypotheses about the characteristics of improperly completed surveys (i.e., exclusion criteria), and generating hypotheses about the characteristics of properly completed surveys (i.e., inclusion criteria) – all before data collection takes place – could reduce the hindsight and foresight bias that affects the interplay of evidence evaluation and researcher design choices during the completion of a scientific project (Davis & Fischhoff, 2014). Although these criteria-as-hypotheses would not be formally tested by the data – unless we could devise some measure of ground-truth of the participant's ability to communicate their preferences – they could force the researcher to consider the elicitation techniques in use more carefully than they would otherwise, and to more judiciously apply the exclusion/inclusion criteria they define. That could decrease what is likely a high false alarm rate in current practice. At the very least, it could generate ideas for new studies that would improve data collection (and exclusion).

7.5 Implementing exclusion criteria: Accounting for participant health

Finally, a researcher wanting to implement exclusion criteria might want to account for the differences between those with good and poor health status, and the interaction of health status with the exclusion criteria. As we saw in the sleep example, the effect of *violates-SG* is more extreme if we focus on participants who sleep poorly. Incorporating the health status of the sample defining the utilities is normatively permissible under social choice theory – these are called *non-neutral* social orderings. Our modeling approach could easily be adapted to determine whether participants' PROMIS scores affect the participants' valuations of all the domains (and with a little more effort, include the interaction with one or with many exclusion criteria). The literature investigating the differences between community preferences and those of patients or other experienced individuals is large (Dolders, Zeegers, Groot, & Ament, 2006; Gerhards, Evers, Sabel, & Huibers, 2011; Krabbe, Tromp, Ruers, & van Riel, 2011; Menzel et al., 2002; Mulhern et al., 2014; Rowen et al., 2015; Schwalm, Feng, Moock, & Kohlmann, 2015; Versteegh & Brouwer, 2016), with inconsistent findings (Dolders et al., 2006; Peeters & Stiggelbout, 2010; Versteegh & Brouwer, 2016).

We will not attempt a full description of the competing ethical perspectives advanced for emphasizing preferences from one group over another, which has been done elsewhere (Versteegh & Brouwer, 2016). Following Vertseegh & Brouwer (2016), community preferences and those of patients or experienced individuals all have a role to play, especially when moving from one group to the other changes whether an ICER falls on the cost-effective side of the cost-effectiveness threshold. However, to the best of our knowledge, our analysis is the first to investigate the interaction with exclusion criteria. If the construction of the utility function is sensitive to the interaction of the exclusion criteria and the participants' health, that could be relevant to the decision to apply the criteria. For example, we found that participants with poor sleep quality who are excluded by *violates-SG* are less sensitive to theta than participants who are excluded but sleep well, but those who remain and sleep poorly are equally sensitive as those who remain and sleep well. The larger effect of the exclusion on the poor sleep group could indicate a bifurcation in their preferences, unless *violates-SG* truly captures those not using the SG to express their preferences. If there is any doubt, then we are potentially ignoring preference heterogeneity among those who could benefit from treatment (e.g., by Lunesta).

That heterogeneity, including its practical effect, could be better understood by extending our analysis into the other domains and the disutility corner states (see Part II) that define the weighting of the health domains in the final multi-attribute function. For example, it could be that one group of poor sleepers shows little sensitivity among hypothetical sleep states because they are resigned to it (i.e., a form of adaptation) (Versteegh & Brouwer, 2016), and they are flagged by *violates-SG* because the noisiness of the SG measurement tool makes expressing low sensitivity difficult without violating dominance. But, this same group might show *increased* sensitivity to the ability to perform their social roles, a sequela of poor sleep (Snedecor et al., 2009). That might describe, for example, the preferences of new parents, who see perpetual sleep-deprivation as an unavoidable consequence of their decision to have children, but believe there is a way to recover some aspect of their previous social lives despite their lower-quality sleep. Conventional exclusion practices cannot detect these nuances, because they do not look for them.

8

Conclusion

We have attempted a systematic study of the exclusion criteria used in health state valuation studies to remove preference data, criteria that have the stated goal of enhancing the data's quality (Engel et al., 2016). We cannot provide unqualified suggestions to researchers about the exclusion criteria they should use, because we did not have a source of ground-truth of whether participants were using the standard gamble elicitations to communicate their preferences – the construct we believe these criteria are attempting to capture.

However, even in the absence of that data, we can say something. The non-preference based criteria posit explicit causal mechanisms: *numeracy* conjectures that a minimal amount of numeracy is required to understand the SG; *time* conjectures that it is impossible to use the SG if one is moving too quickly through the survey; and, *understanding* says that those who state that they were not sure if they understood the survey are admitting that the SG does not represent their preferences faithfully.

The proximity of these criteria in the MDS plot (Figure 5.1) provide them with some validity – though *a priori* they are unrelated, they end up agreeing on some of the exclusions. Of the three, we recommend *time* and *numeracy* over *understanding*, because participants removed by *understanding* have similar preferences to those who are not (Figure 6.3), indicating the criterion might mainly be capturing those who admit that an unusual task is exactly that.

The preference-based criteria are more difficult to evaluate. They did not cluster as well in the MDS as the non-preference based criteria. Of the preference-based criteria, our opinion is that *dead-all-worst* is the best among the group. The results from our worked example of the social domain (Figure 6.20) are particularly compelling, as we saw patterns of responses, such as group-level violation of dominance among the excluded, which are difficult to explain as true preferences. The rating of dead or the all-worst state as equal to or better than full health could be such an egregious ordering error that the criterion can successfully identify a participant providing non-preference data. The proximity of *dead-all-worst* to *numeracy* in the MDS plot provides one possible causal mechanism for that low-quality data.

Upper-tail and *lower-tail* – which together define *10% trimming* – exclude two different groups of people. Without more investigation of what differentiates those groups, we do not recommend these criteria. Furthermore, although *10% trimming* will increase the variance of the utility estimates, it is unlikely to change the point estimates, due to the symmetry of the criteria and the number of 0s and 1s in our data. Similarly, we cannot endorse *low-range* or its nested cousin, *no-variance*, until we understand whether those who use little of the utility scale are answering that way for a relevant reason (e.g., their health status, or as protest responses (Wittenberg & Prosser, 2011)).

Violates-SG cannot be recommended either. Even if one accepts the normative theory on which it is based, accounting for the noisiness of the SG – the MDS results suggest that those flagged by *violates-SG* could be those *trying* to provide the most precise SG responses – *violates-SG* must have a high false alarm rate, given the sheer number of exclusions (71.6% of the sample). (However, excluding that many is not unheard of in the literature (Engel et al., 2016).) Furthermore, the participants excluded by *violates-SG* are, as a group, just as well-behaved as those who are not excluded (Figure 6.6), which means that excluding them for modeling purposes is unnecessary – a rationale often provided for exclusions. Given the focus on the *person-mean* (Dewitt et al., 2017; Feeny et al., 2002; Furlong et al., 1998; Hanmer & Dewitt, 2017), it is difficult to motivate the criterion if the (person-)mean of those excluded by *violates-SG* does not itself violate dominance. As we have argued, the individual-level violations could be caused by measurement error, rather than inattention or some other mechanism connecting the violations to responses that are not true preferences (e.g., protest responses) (Wittenberg & Prosser, 2011).

In fact, one of the chief risks of current exclusion conventions is the missed opportunity to improve elicitation techniques. Although comparisons of the properties of these techniques is a topic of interest in the literature (Badia, Monserrat, Roset, & Herdman, 1999; Bleichrodt, 2002; Bleichrodt & Johannesson, 1997; Torrance et al., 2001), to the best of our knowledge there is little work that aims to determine if exclusion criteria set reasonable thresholds on the ability to use the SG properly. Therefore, the continuing implementation of conventional practices could mean we are missing the opportunity to understand a subpopulation that is trying, but failing, to express themselves in our surveys. Operationally, that missed opportunity might not matter if these individuals are properly represented by others included in the sample. But, given the importance in health state valuation studies of maintaining the "societal perspective," we should be confident that this representation-by-proxy is true, rather than continuing to follow convention and exclude participants. The challenge to representativeness could be severe if it turns out that the demographic variables that are meant to ensure a final representative sample are not the most important causes of preferences – as has been argued in several studies (Badia, Fernandez, & Segura, 1995; Devlin et al., 2003; Kind & Dolan, 1995). Adding demographic variables to our model is an important next-step in our work.

To apply our recommended criteria judiciously, we also recommend that researchers implement them in a way that reduces false alarms. For example, one could require being flagged by two (or more) criteria before a participant would be excluded, assuming the agreement of *a priori* independent criteria increases the confidence of the categorization. Considering explicit *inclusion* criteria would also help ensure researchers understand the assumptions they are making about participant preferences, and help avoid *post hoc* exclusions that reduce the variance of the data, but might be *introducing* bias.

The main practical danger of applying exclusion criteria incorrectly – even the ones we provisionally recommend – is that utility curves could change the results of analyses based on them. Our results suggest that we risk producing utility curves that are more sensitive to theta than they should be. If cost-effectiveness thresholds are fixed, then this will make an intervention's ICER appear smaller and over-estimate the intervention's cost-effectiveness, potentially causing misinformed resource allocations if the ICER estimate is close to a cost-effectiveness threshold.

As we saw, because near-constant utility curves are a common result among the excluded, one theory-driven way to avoid the danger of wrongful exclusion is to take the normative position that we do not care about groups of constant (mean) utility. Social choice theory allows one to take that stance – at the expense of being unable to incorporate many ethical principles of interest (e.g., instantiations of equity) into the calculation of utilities. Whether to go this route appears to us to be an instance of the meta-preference problem defined in Part I, and resolving it would likely require the analytical-deliberative framework described there. In our opinion, it is not worth sacrificing the freedom to incorporate social values into utility calculations, and we should instead strive to improve the accuracy of our exclusion criteria and their implementation through better preference measurement.

To that end, we plan to: analyze more criteria defined via the VAS, with a view to improving preference elicitation techniques; alter our models to account for participant health status beyond our examples of sleep and *violates-SG* and social roles and *dead-all-worst*, to see how participant health affects the implementation of each criterion; and, extend our analyses to jointly model commonly-used subsets of exclusion criteria, such as those in the sensitivity analysis section of Part II.

Ultimately, the goal is to provide systematic guidance about how to use lay preference data of all kinds so that data “cleaning” is not outside of the core analyses, nor are its effects focused on the final utility calculations and relegated to sensitivity analyses in a supplementary information section. Given the importance of such data in policy analysis, we should understand the measurement properties and policy implications of how we choose the final sample of data, especially if we wish to claim that the final product gives a societal perspective.

9

Postscript

Societal preference-based health-related quality of life is an example of synthesizing a multi-faceted concept – health – into a metric that can be used uniformly across (health-related) policy analyses (Fischhoff, 2015). Those analyses (e.g., cost-effectiveness analyses) are done according to exacting standards (Neumann et al., 2016), providing them with a common global structure. That common structure allows for a kind of commensurability among analytical conclusions. Confidence in that commensurability leads to confidence in the policies informed by the analyses, policies that ultimately allocate resources to address the diverse health needs of a population.

In this dissertation, we have attempted to illustrate the consequences of the design choices required to produce health-related quality of life utility values. Those choices have practical and ethical implications; for example, changing how one values the lives of the disabled (i.e., the utility of poor states of health). Central to the design of tools meant to represent society – a process that includes summarizing data and deciding whose data count – is the fact that these tools are created, and are not estimations of some true underlying construct. With that knowledge, a researcher, advisor, or policy-maker can better appreciate the influence of their analytical choices and strive to make them and their political-ethical consequences transparent. That transparency could improve the communication between scientists, analysts, policy-makers, and the public, leading to a better understanding of the limitations of policy analysis, and thus better-informed policy decisions.

A

Appendix

This appendix provides more background on societal preference-based measures of HRQL, focusing on the Health Utilities Index Mark 2 (HUI:2) and Mark 3 (HUI:3). As in the main text, we use "HUI" to refer to both the Mark 2 and Mark 3 versions when distinguishing between them is not required. It also provides a list of every component of the PROPr survey, as well as a detailed description and discussion of MDS and beta regression.

A.1 Standardized societal preference-based measures of HRQL

The HUI system, along with related generic societal preference-based measures of HRQL (e.g., EuroQol-5D, SF-6D, Quality of Well-Being Scale), defines a *state space* that describes possible states of health. The developers of the HUI then used an elicitation procedure to assign a number to each state in that space, representing the quality of that health state. These numbers are treated as *utilities*, compactly representing respondents' preferences (Keeney & Raiffa, 2003). The HUI:2 and HUI:3 scoring functions are explicitly estimated using *multi-attribute utility theory* (MAUT) (Feeny et al., 2002; Keeney & Raiffa, 2003; Furlong et al., 1998; Torrance et al., 1996). Other measures use other methods (e.g., regression analysis) to produce their scoring functions.

As an example, a state space might have two attributes: mobility and vision. Health states within it would be represented by a vector $v = (x_{mobility}, x_{vision})$, where $x_{mobility}$ is a level of mobility and x_{vision} is a level of vision. A societal utility (scoring) function, U, for the space would assign a real number, $U(v)$, to each possible vector (i.e., health state) v. One health state is preferred to another if it has a higher $U(v)$. See Figure A.1.

Table A.1 shows examples of the attributes (e.g., mobility, vision, cognition) used in HUI:2 or HUI:3. HUI:2 has seven attributes, each having three to five levels; HUI:3 has eight attributes, each having five to six levels. Level 1 is best for any attribute. The health state described by the vector of all 1s is the full health state (i.e., the most-able state). Taking into account all of the attributes and the levels of each system, HUI:2 involves a seven-dimensional state space with 24,000 unique health states; HUI:3 involves an eight-dimensional state space with 972,000 unique health states.[1] Valuation study participants were asked to compare a tiny fraction of the possible HUI states. The analysis then relied on assumptions about the coherence of their preferences and the interactions between the attributes to generate a utility function for the entire state space.

HUI:2 and HUI:3 use two different methods to aggregate individual utility functions into a societal utility function:

1. *Mean (overall) utility function.* During the creation of the HUI:2 system, multi-attribute utility functions are estimated for each individual in the sample of participants. These are then averaged to produce an overall societal multi-attribute utility function over the entire state space.

2. *Person-mean (attribute-based) utility function.* This approach first produces a utility function for each attribute (e.g., mobility, pain), using the mean of the elicited preferences for each level of the attribute. Each of these single-attribute functions is conceptualized as a single-attribute function of a hypothetical individual, the *person-mean,* whose preferences equal the mean of individual preferences within each attribute. These single-attribute functions are then combined (using MAUT methods) to form the overall societal function (Keeney & Raiffa, 2003). HUI:3 uses this method exclusively. HUI:2 uses it in comparison with the mean (overall) utility function method, in order to determine the extent of their disagreement, advocating for the person-mean approach, given its simpler elicitation procedure.

For a simplified example, consider the two-attribute state space, (*mobility, vision*), and a sample

[1] The state spaces for HUI:2 and HUI:3 each include an additional state that is not described as a vector of attributes: dead. That is, the state is described by the worst level of each attribute is *not* assumed to be the same as dead. The former is called the *all-worst state*, or the *most-disabled state*, also sometimes called the “pits.”

of n people. Assume that each attribute has three levels, creating nine health states. The mean (overall) utility function approach would first elicit each individual's utilities for the three levels of mobility and vision separately, using these values to produce a single-attribute utility function for mobility and for vision for each individual. For each person i, their two single-attribute utility functions would be combined using assumptions from MAUT to produce an overall (multi-attribute) utility function $u_i(x_{mobility}, x_{vision})$. The societal (multi-attribute) utility function is the average of these individual (multi-attribute) utility functions, i.e.,

$$U = \frac{1}{n} \sum_i u_i.$$

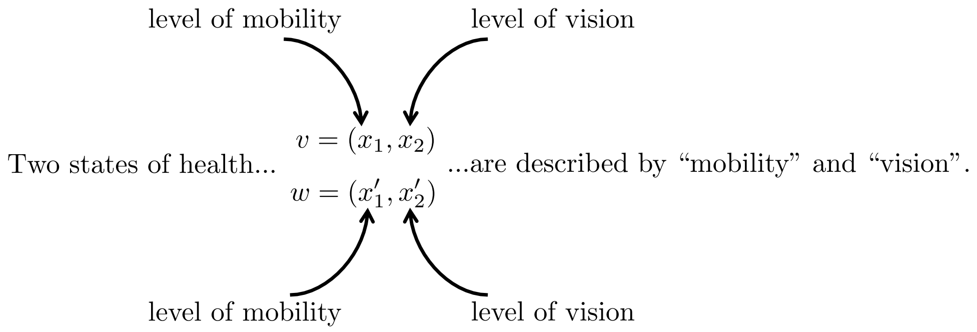
The person-mean (attribute-based) approach would elicit single-attribute utility functions for mobility and for vision from individuals in the sample, but not necessarily both from *every* (or any) individual. The single-attribute functions would then be averaged to produce mean *single*-attribute utility functions for mobility and for vision, $U^{mobility}$ and U^{vision}. The societal multi-attribute utility function combines these two societal single-attribute functions using MAUT modeling assumptions, i.e., $U = f(U^{mobility}, U^{vision})$. See Figure A.2.

Thus, although the methods differ, both *average* individual preferences in order to produce a societal aggregate, with one averaging over the multi-attribute utility functions and the other averaging over single-attribute functions.

For more background on societal preference-based HRQL measurement, including the HUI systems and others, see (Neumann et al., 2016, Chapter 7).

Table A.1: Example attributes and descriptions of their levels from HUI:2 and HUI:3 (Feeny et al., 2002; Torrance et al., 1996).

Attribute	Level	Description
Sensation (HUI:2)	1	Able to see, hear, and speak normally for age.
	2	Requires equipment to see or hear or speak.
	3	Sees, hears, or speaks with limitations even with equipment.
	4	Blind, deaf, or mute.
Mobility (HUI:2)	1	Able to walk, bend, lift, jump, and run normally for age.
	2	Walks, bends, lifts, jumps, or runs with some limitations but does not require help.
	3	Requires mechanical equipment (such as canes, crutches, braces, or wheelchair) to walk or get around independently.
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well.
	5	Unable to control or use arms and legs.
Speech (HUI:3)	1	Able to be understood completely when speaking with strangers or friends.
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well.
	3	Able to be understood partially when speaking with strangers or people who know me well.
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well.
	5	Unable to be understood when speaking to other people (or unable to speak at all).
Cognition (HUI:3)	1	Able to remember most things, think clearly and solve day to day problems.
	2	Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.
	3	Somewhat forgetful, but able to think clearly and solve day to day problems.
	4	Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.
	5	Very forgetful, and have great difficulty when trying to think or solve day to day problems.
	6	Unable to remember anything at all, and unable to think or solve day to day problems.



A societal utility function U assigns a number to every health state:

$U(v) > U(w) \Rightarrow$ the health state represented by v has better quality of life than that represented by w.
$U(v) < U(w) \Rightarrow$ the health state represented by v has worse quality of life than that represented by w.
$U(v) = U(w) \Rightarrow$ the health state represented by v has equal quality of life to that represented by w.

U allows **comparisons** between all health states than can be described as a vector $(x_{mobility}, x_{vision})$.

Figure A.1: An example of a simple health-related quality of life measurement system.

A.2 The PROPr survey

The survey had the following components:

1. Consent to participate.
2. Demographic information.
3. Participant's overall self-rated health: excellent, very good, good, fair, or poor (Hays, Spritzer, Thompson, & Cella, 2015).
4. One of 4 existing patient-reported outcome measures, chosen at random:
    - The PROMIS Global Health Items (Hays, Bjorner, Revicki, Spritzer, & Cella, 2009).
    - The EQ-5D-5L with visual analogue scale VAS (Herdman et al., 2011).
    - The Health Utilities Index Mark 2 and 3 (Feeny et al., 2002; Torrance et al., 1996).
    - Chronic Health Conditions List (12 conditions) (CDC, 2016).
5. The PROMIS-29 questionnaire (Gershon et al., 2010), plus 4 questions from the Cognition short form (PROMIS, 2015).
6. The participant's self-assessed additional life expectancy.
7. Valuation of 1 of the 7 health domains, assigned at random.
8. Task engagement questions.
9. Additional questionnaires presented in randomized order:

There are 9 health states $(x_{mobility}, x_{vision})$ describable in the system: $\begin{array}{ccc} (1,1) & (1,2) & (1,3) \\ (2,1) & (2,2) & (2,3) \\ (3,1) & (3,2) & (3,3) \end{array}$

Goal: Build a societal utility function U that assigns a number to each state.

The mean (overall) utility function approach

Elicit single-attribute functions for both attributes from each of the n individuals in the sample:

$$\{u_1^m(x_{mobility}), u_2^m(x_{mobility}), \dots, u_n^m(x_{mobility})\},$$
$$\{u_1^v(x_{vision}), u_2^v(x_{vision}), \dots, u_n^v(x_{vision})\}.$$

Use these to produce a multi-attribute utility function for each individual:

$$u_i(x_{mobility}, x_{vision}) = f_i(u_i^m(x_{mobility}), u_i^v(x_{vision})),$$

where f_i is determined by MAUT [4]. The societal utility function U is their average:

$$U = \frac{1}{n}\sum_{i=1}^{n} u_i.$$

The person-mean (attribute-based) function approach

Elicit k $(k \leq n)$ single-attribute functions for mobility.

Elicit j $(j \leq n)$ single-attribute functions for vision.

Create societal **single**-attribute utility functions for both mobility and vision:

$$U^{mobility} = \frac{1}{k}\sum_{i=1}^{k} u_i^m$$
$$U^{vision} = \frac{1}{j}\sum_{i=1}^{j} u_i^v.$$

Define U to be a function of the two societal single-attribute functions:

$$U = f(U^{mobility}, U^{vision}),$$

where f is determined by MAUT [4].

Figure A.2: A simplified example explaining the two preference aggregation approaches used in the HUI systems.

- The 3 questionnaires from (4) not yet administered.
- The 3-question short form of the Subjective numeracy Scale (Fagerlin et al., 2007; McNaughton et al., 2015).
- Experience with disability.
- Distributional preferences.

These are described in more detail in the technical report (Hanmer & Dewitt, 2017), available at `http://janelhanmer.pitt.edu/PROPr.html`.

A.3 Multidimensional scaling

A.3.1 Proximity indices

The choice of index depends on the form of "proximity" relevant to a problem. In a more familiar context, we might define the distance between two buildings in a city "as the crow flies" – or, we might measure it in terms of the length of the road we need to travel to complete the journey (the so-called *city-block metric*). These can be the same, if there is a perfectly straight road between every building, but otherwise, they will differ. For example, two buildings that are back-to-back might be almost maximally proximal in the first distance; but, if you must walk around the block to get from the front door of one to the front door of the other, they would be much further apart in the second. The first would be appropriate if we were delivering packages by drone; the second, if we were providing walking directions. Similarly, the choice of proximity index for exclusion criteria must be sure to capture the type of "closeness" we desire.

We chose phi (ϕ) because it captures two important features in the exclusion criteria space, which we call: 1) *large complementarity* and 2) *small simultaneity*. *Large complementarity* refers to a large difference between two criteria if they each flag many participants, but flag *different* participants; they act in a complementary way. *Small simultaneity* refers to a small difference between two criteria if they each flag few participants, and flag (almost) the same participants; they are both selecting for a rare set of attributes that often co-occur. Phi has these properties. For example, large complementarity says that if we have a sample of 1,000 people and (*i*) criterion A_1 excludes 500 and A_2 excludes 500, with none of those people in common, we should be more certain in the large "distance" between A_1 and A_2 than if (*ii*) each excluded 10 of 1,000, with none in common. The phi value for (i) is -1, while it is -0.01 for (*ii*). Large complementarity reflects that, if A_1 and A_2 are in fact proximal with respect to their exclusions of individuals – they exclude the same people – it is much more likely to observe (*ii*) than (*i*). Conversely, small simultaneity says that if each of A_1 and A_2 excluded only 10 out of 1,000 people and excluded *the same* people, we should define them as more similar than if they excluded 500 out of 1,000, with a large overlap. The latter is more likely to arise by chance than the former. Accordingly, the phi value of the former is 1 and is 0.2 for the latter (with an overlap of 300 of 500 exclusions).

A.3.2 MDS implementations and goodness-of-fit

MDS takes a proximity matrix p – where each entry p_{ij} is the proximity between two objects – as input. An MDS algorithm then searches for a configuration X in m-dimensional space, such that the Euclidean distance in X between object i – an exclusion criterion, in our case – and object j, denoted by d^X_{ij}, is related to their proximity p_{ij} by some function f. This can be viewed as the problem of finding the configuration X, and the function f, such that the residual $e_{ij} = f(p_{ij}) - d^X_{ij}$ is minimized over all objects. More specifically, we seek to minimize the sum of squared e_{ij}s, which, when normed to account for the scale of X – recall that X is a geometric space, like a map, with arbitrary units of distance – is called *stress* (Borg & Groenen, 2005).

The functional form of f defines the type of MDS, and depends on the scale type of the input data. For example, *interval* MDS – also known as *classical* MDS or *metric* MDS – assumes that the data is on at least an interval scale. That means the ratios of differences between proximities are meaningful, as they would be if the proximity index itself was Euclidean distance, as in our example of using MDS to recover the distances between cities. Interval MDS takes f to be a linear function of the proximities, i.e., $f(p_{ij}) = a + bp_{ij}$. In interval MDS, one minimizes stress by choosing different configurations and different parameters for f. Thus, finding an MDS solution is akin to performing linear regression to estimate f, while simultaneously adjusting the configuration of X, all in an attempt to find the lowest stress combination possible (Borg & Groenen, 2005).

However, there are problems with interval MDS. Many proximity indices, including phi, are not on an interval scale (Gower & Legendre, 1986). Furthermore, although it is possible to transform indices to have interval-scale-like properties, doing so increases the number of assumptions one must make about the data: It requires that we trust the size of the differences between proximities just as much as we trust the proximities themselves. Allowing f to take on other functional forms provides more flexibility, requires making fewer assumptions about the data, and allows one to find better – i.e., lower stress – configurations.

One such alternative type of MDS is called *ordinal* (non-metric) MDS. Ordinal MDS uses ordinal regression, rather than linear regression, in the process of finding an optimal configuration X. Ordinal regression only tries to preserve the *rank order* of the proximity data. Thus, it only assumes that the arrangement of the objects (exclusion criteria) from least proximal to most proximal – however defined via a proximity index – is relatively stable. We use ordinal MDS because of its fewer assumptions.

To see the difference between the types of MDS, Figure A.3 shows a *Shepard plot* for each. The Shepard plot displays the regression line of f. For technical reasons, the x-axis is not the proximity scale data, but a transformation of the proximities into *dissimilarities*, i.e., where 0 is the least dissimilar (most proximal) and positive numbers are increasingly dissimilar (less proximal).[2] The y-axis shows the distances (the d_{ij}^Xs) in the associated configuration. The points on the plot show how the dissimilarities are mapped to distances in the configuration. A good solution has small residuals. Note that a solution where every fitted value $f(p_{ij})$ coincided with its d_{ij}^X is not necessarily ideal, even though it obviously minimizes stress, because it is likely overfitting the data, just as in a regression with too much curvature. Rather, we look for a spread of the distances around the fitted values, such that we are not systematically violating the structure of the proximity data (e.g., the order of the distances roughly match the order of the distances), and the residuals do not show a systematic difficulty of scaling a particular size of dissimilarity.

Another choice to make in the implementation of MDS is the dimensionality of the configuration space. The larger the number of dimensions, the better the fit of the scaling (in terms of stress), because there is more freedom to arrange the objects in the space. However, dimensions beyond three are difficult to use; even a 3-dimensional scaling of a few number of objects can be difficult to interpret. There is no definitive rule for choosing the number of dimensions. By

[2] In order to turn the ϕ index into an index of dissimilarity, there are two possible transformations: $1 - \phi$ or $\sqrt{1 - \phi}$ (Gower & Legendre, 1986). They have the same ordinal properties, but can change the results of the other MDS algorithms.

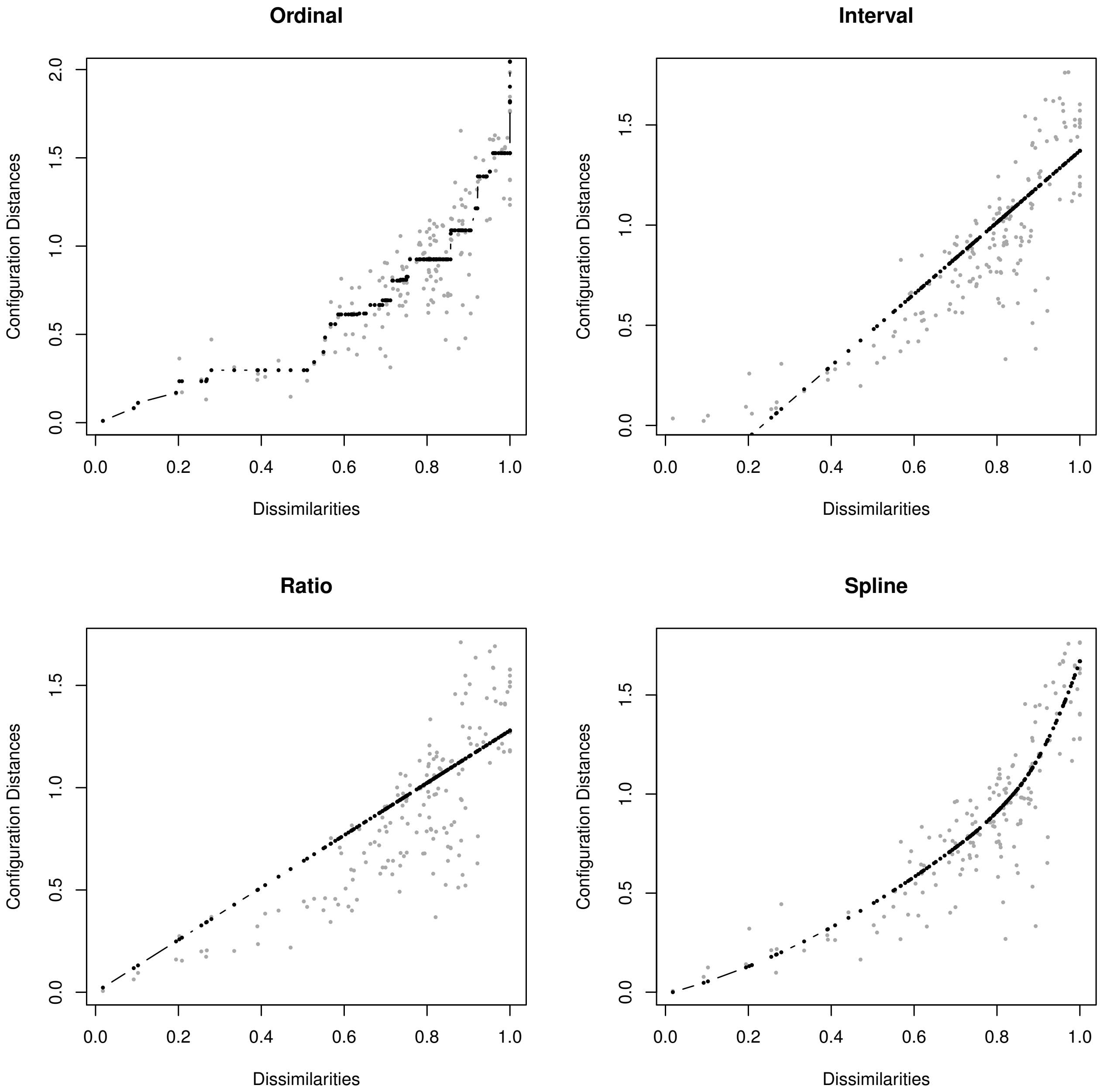


Figure A.3: *Example Shepard plots.* Example Shepard plots from scalings using different MDS algorithms. The x-axis shows the *dissimilarities* – a transformation of the proximities – and the y-axis shows the distances between objects in the associated configuration. Each point is the distance between two objects, so there are $\frac{n(n-1)}{2}$ points for n objects. The type of MDS algorithm – ordinal, interval, ratio, spline – determines the type of scale-type assumed for the dissimilarity and proximity data, and the type of regression used to minimize the residuals (the *stress*).

convention, researchers rely on the change in stress combined with content knowledge of the application. Examining the *Scree plot*, which has the number of dimensions on the x-axis and the stress of a scaling of given dimensionality on the y-axis, can help with this process. The "elbow" of the plot, the step at which the change in stress diminishes visibly from the previous step, is taken as suggesting that one might be starting to scale the noise in the data. Figure A.4 shows an example.

A complementary approach starts with a 1-dimensional scaling, and adds dimensions, examining the resulting analyses for new structural relationships. When those cease, nothing is gained scientifically by increasing the dimensionality of the solution, even if, formally, the stress decreases. As the entire purpose of MDS is to gain insight on the structure of the objects by simultaneously representing all $\frac{n(n-1)}{2}$ proximity relationships in a low-dimensional geometric configuration, then the lowest-dimensional useful representation can be "optimal," in the sense that it avoids potentially scaling noise (Borg et al., 2012).

A.3.3 Results of the MDS sensitivity analysis and model-checking

Figure A.5 shows the Shepard plot for Figure 5.1. It shows a good fit, with no systematic relationship to the residuals. There are other plots that provide useful diagnostics for evaluating the fit of an MDS solution, which are derived from the Shepard plot. Figure A.6 shows the *stress-per-point*. Recall that every point on the Shepard plot shows the relationship between one pair of criteria in terms of their proximity (as measured by phi) and its distance in the configuration. Each object in a scaling of n objects is associated with $n-1$ points in the Shepard plot. One can apportion the stress in the Shepard plot to each of the objects, by taking the sum of all the squared residuals involving that object and dividing by the total sum of squared residuals. Combining the stress-per-point plot for our core criteria (Figure A.6) with the configuration (Figure 5.1), one can produce a configuration where each object is represented by a point whose size indicates the amount of stress associated with the scaling (Figure A.7). This is called a *bubble plot*. The larger the bubble, the more difficult it is for the MDS solution to scale that criterion. Figure A.6 and Figure A.7 show that *low-range*, *no-variance*, *violates-SG*, *violates-VAS*, and *lower-tail* were relatively easy to configure in 2-dimensional space, whereas *upper-tail*, *time*, *understanding*, *dead-all-worst*, and

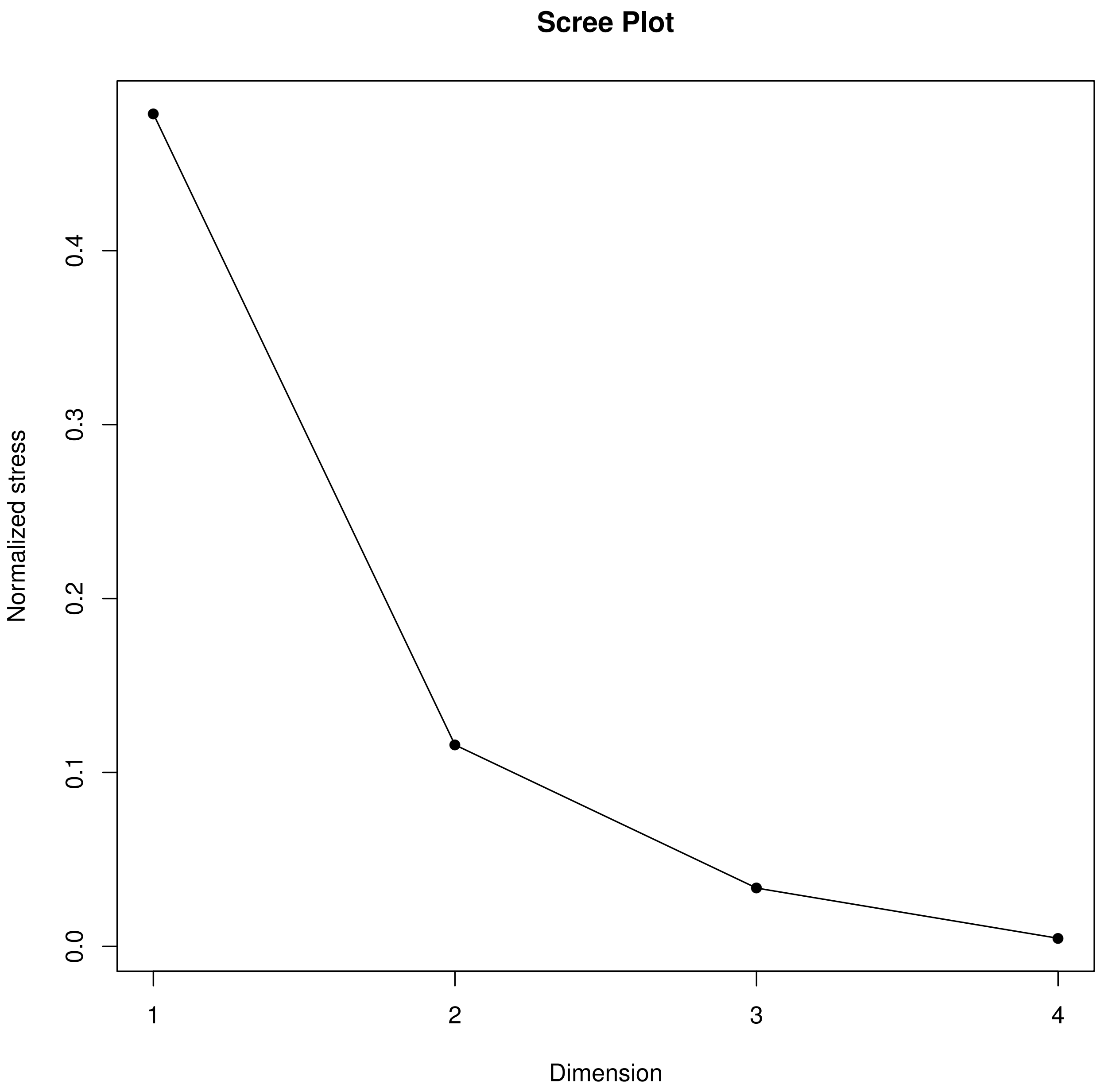


Figure A.4: *Example Scree plot*. The x-axis shows the dimension of the MDS solution, and the y-axis shows the stress of that solution. By definition, stress decreases as the dimensionality increases, because there are more ways to scale the objects and maintain the relationships of the proximities. The "elbow" of the plot is the point at which the marginal decrease in stress changes abruptly, which is usually assumed to point to the best dimensionality for the MDS configuration. Here, that elbow is at 2-dimensions.

numeracy were the most difficult to scale.[3]

Figure A.8 shows the Scree plot for ordinal MDS for the core criteria, with one, two, three, and four dimensions. The elbow appears at two dimensions. Thus, our primary scaling for the core criteria is a 2-dimensional MDS solution.

The results of other tests of our MDS model include the following:

a) *Dimensionality*: Figure A.9 shows the 3-dimensional solution. It necessarily has a better fit than the 2-D solution, but at the risk of scaling noise. The relationships are similar to the 2-dimensional configuration, but the five difficult-to-scale criteria are separated from the others along the third dimension – lowering the stress of the configuration. The pairwise distances of the 3-dimensional configuration and the 2-dimensional configuration have a linear correlation of 0.90 (p-value < 0.01).

b) *MDS jackknife*: Figure A.10 shows an MDS jackknife plot, demonstrating how each criterion moves when one of the other criteria are removed from the MDS and the configuration is re-calculated. Most criteria are stable, except for *time* and *numeracy*, which both move when *upper-tail* is not included in the scaling. We can use those changes to calculate a dispersion statistic, which measures the average difference between the "leave-one-out" solutions and the original solution. It has a maximum of 2 – meaning that the original solution is highly sensitive to the inclusion of each object – and a minimum of 0 (low dispersion). The dispersion value for the core solution is 0.05.

c) *MDS algorithm*: Moving to other MDS models – interval MDS, and spline MDS – produces similar configurations (Figure A.11), but with one notable difference. The arrangement of *numeracy*, *understanding*, and *dead-all-worst* differ from the ordinal scaling. They are shown alongside a different transformation of the phi proximity index into a dissimilarity index.

d) *Clustering*: In a k-means clustering analysis, we choose k in order to minimize within-cluster variance, and use a Scree plot to find the k at which adding another cluster is *not* uncovering more group structure in the data, by looking for a decrease in the improvement of within-cluster variance. (After all, when k equals the number of objects, within-cluster variance is 0 – but we learn nothing about the potential group structure.) Figure A.12 shows the configuration with a clustering with $k = 3$, the optimal k for this configuration. It places *violation-SG, violation-VAS,* and *lower-tail* in one group; *upper-tail* and *numeracy* in a second group; and, *time*, *understanding*, *low-range*, *no-variance*, and *dead-all-worst* in a third.

A.3.4 Discussion of the MDS sensitivity analysis

a) *Dimensionality*: We cannot see any new structure in the 3-dimensional configuration that is not already present in the 2-dimensional configuration. The third dimension does not seem interpretable in-and-of itself. Applying the jackknife procedure to the 3-dimensional scaling, we get a worse dispersion score, of 0.10 (i.e., it is more sensitive to the inclusion of individual criteria). That, combined with the results of the Scree plot (Figure A.8), give us confidence that two dimensions provides the optimal configuration.

[3] Note that these figures, as well as some others in the Appendix, use the variable names for the exclusion criteria in our data set, rather than the shorthands from Table 5.3.

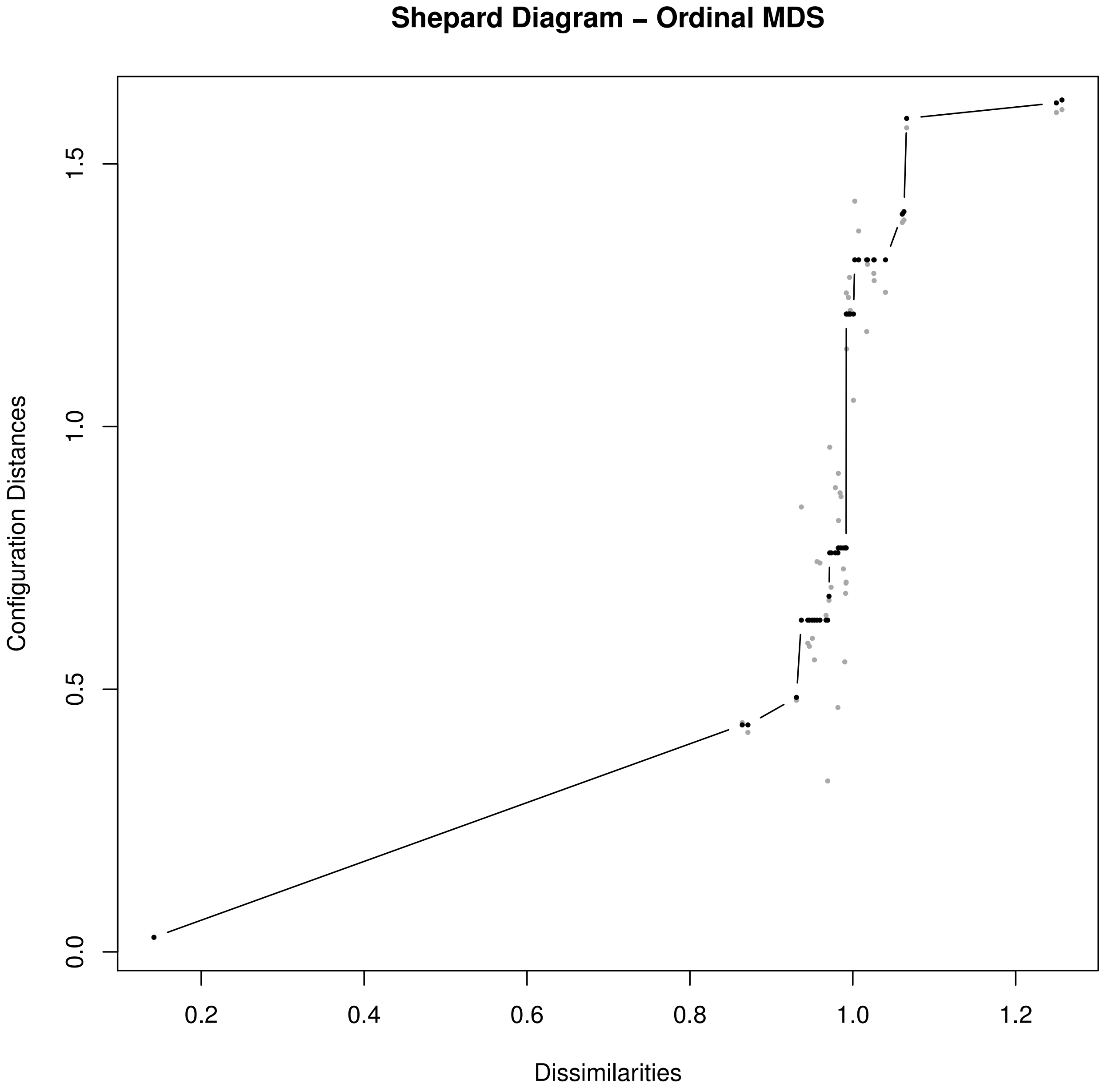


Figure A.5: *Shepard plot for core MDS configuration*. The Shepard plot for the core MDS solution, using the ordinal MDS algorithm.

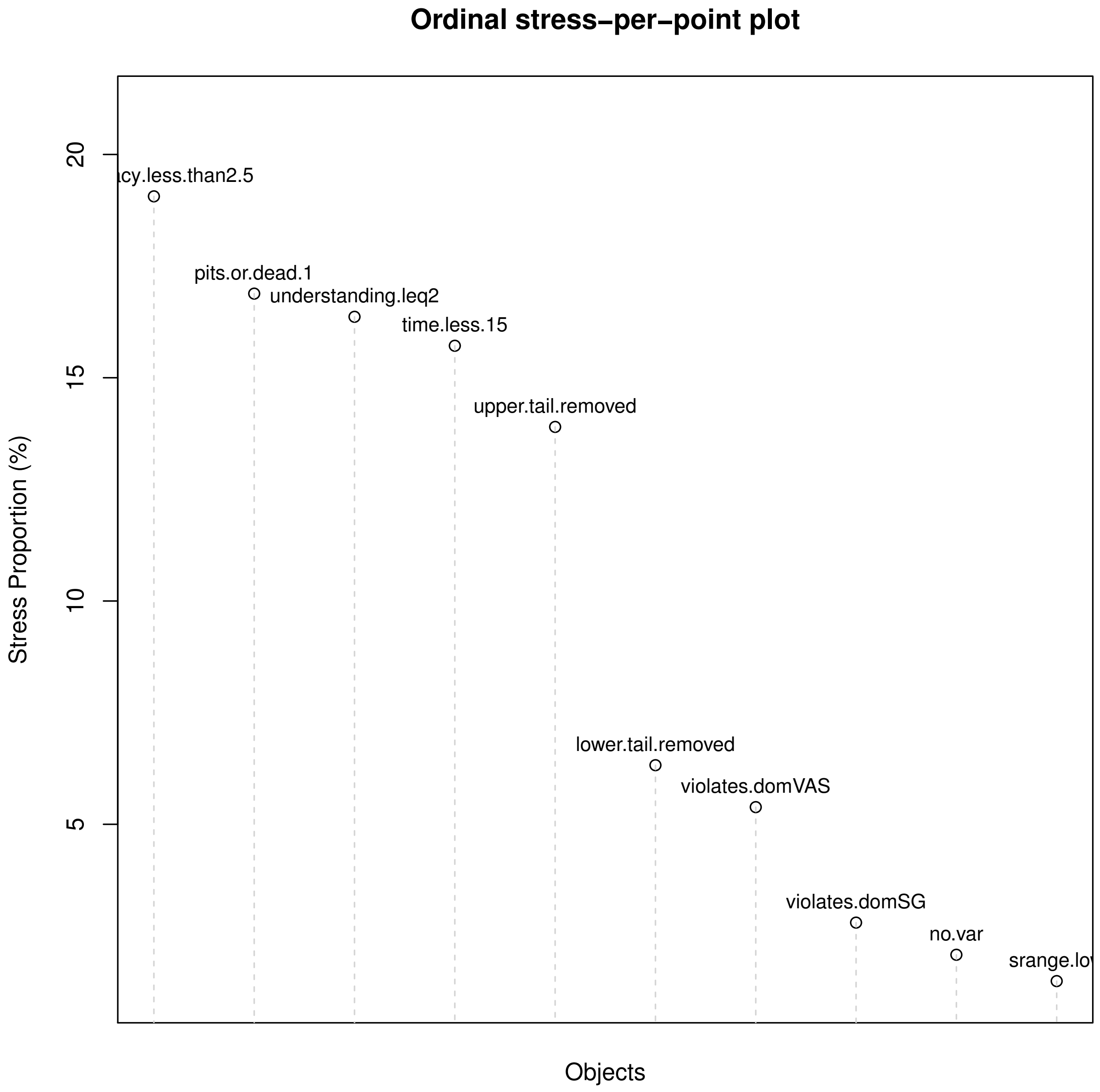


Figure A.6: *Stress-per-point plot for core configuration.* The stress-per-point plot calculates all of the squared residuals from the Shepard plot (Figure A.5), and apportions those residuals (the stress) among the criteria. Those with larger proportions of stress were more difficult to scale; i.e., it is more difficult to maintain all of the proximity relationships of that criterion when producing the configuration.

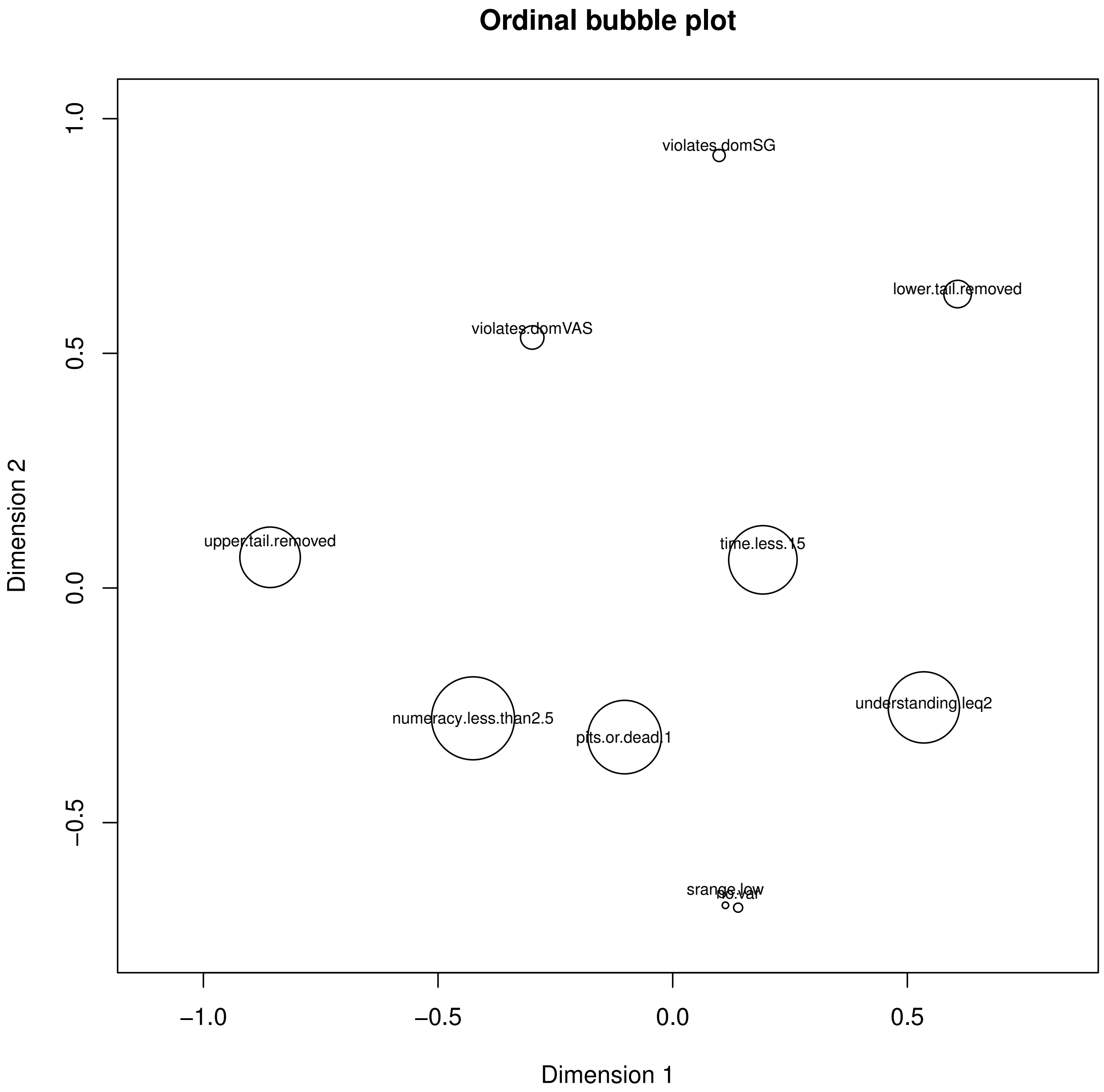


Figure A.7: *Bubble-plot for core configuration.* The bubble-plot of the core MDS solution combines Figure 5.1 and Figure A.6, showing the configuration with each object represented by a bubble demonstrating the stress apportioned to that criterion. The larger the bubble, the more difficult it was to scale that criterion.

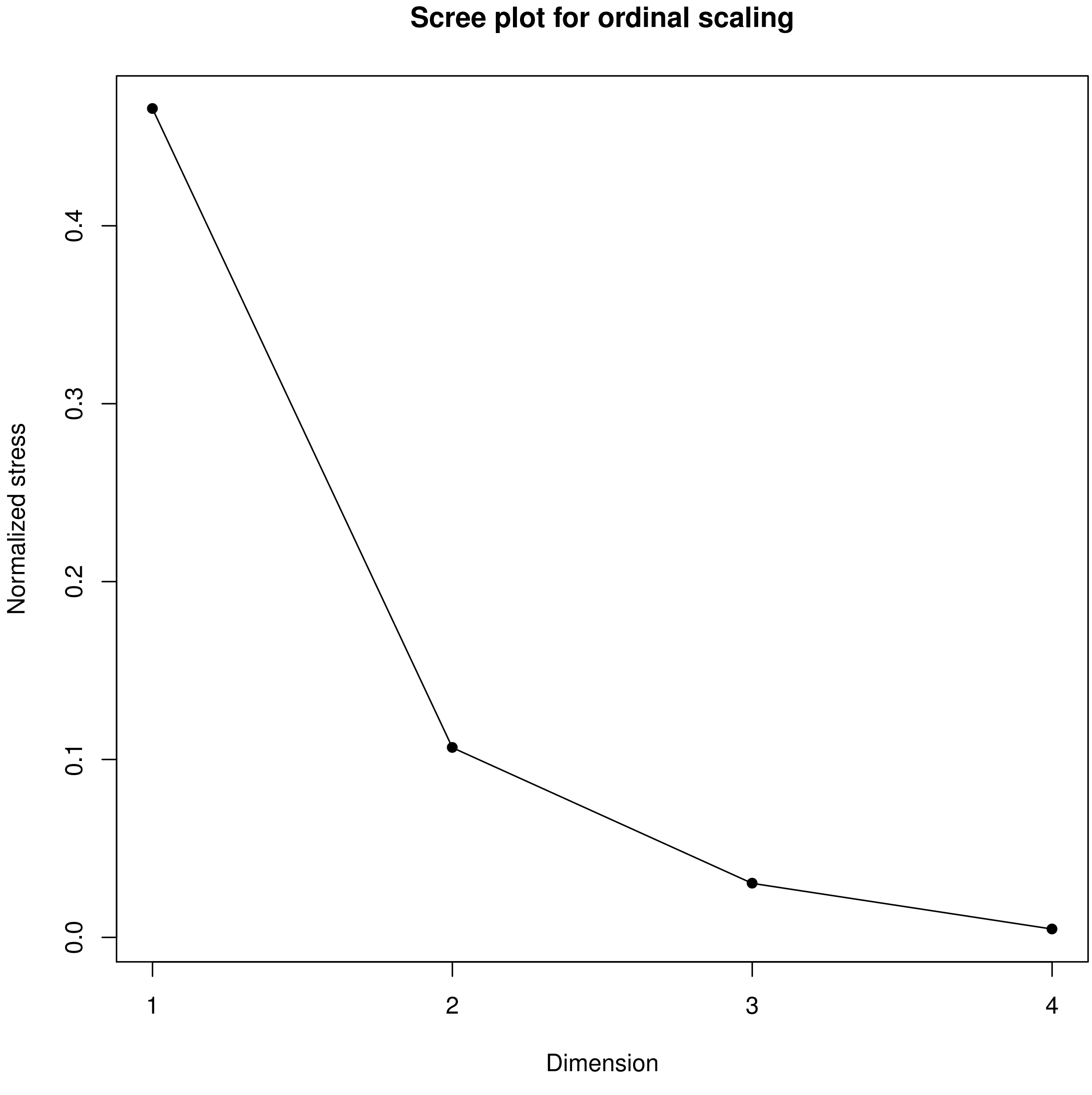


Figure A.8: *Scree plot for core criteria.* The scree plot for ordinal MDS solutions of the core criteria, with the elbow at 2-dimensions.

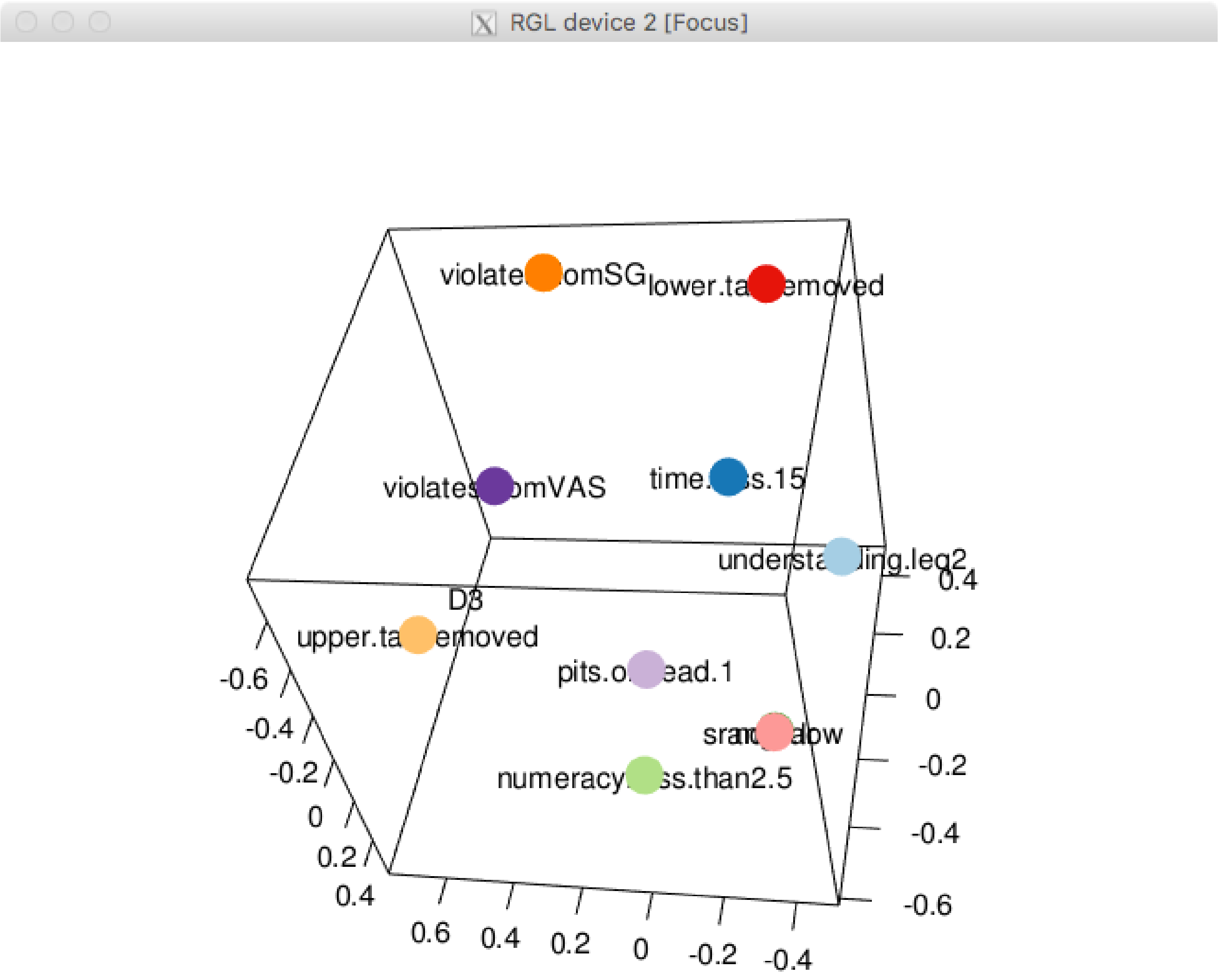


Figure A.9: *3-dimensional MDS configuration for core criteria.* A view of the 3-dimensional configuration corresponding to 3-dimensional ordinal MDS using the core criteria. Notice the similarity of this view to Figure 5.1.

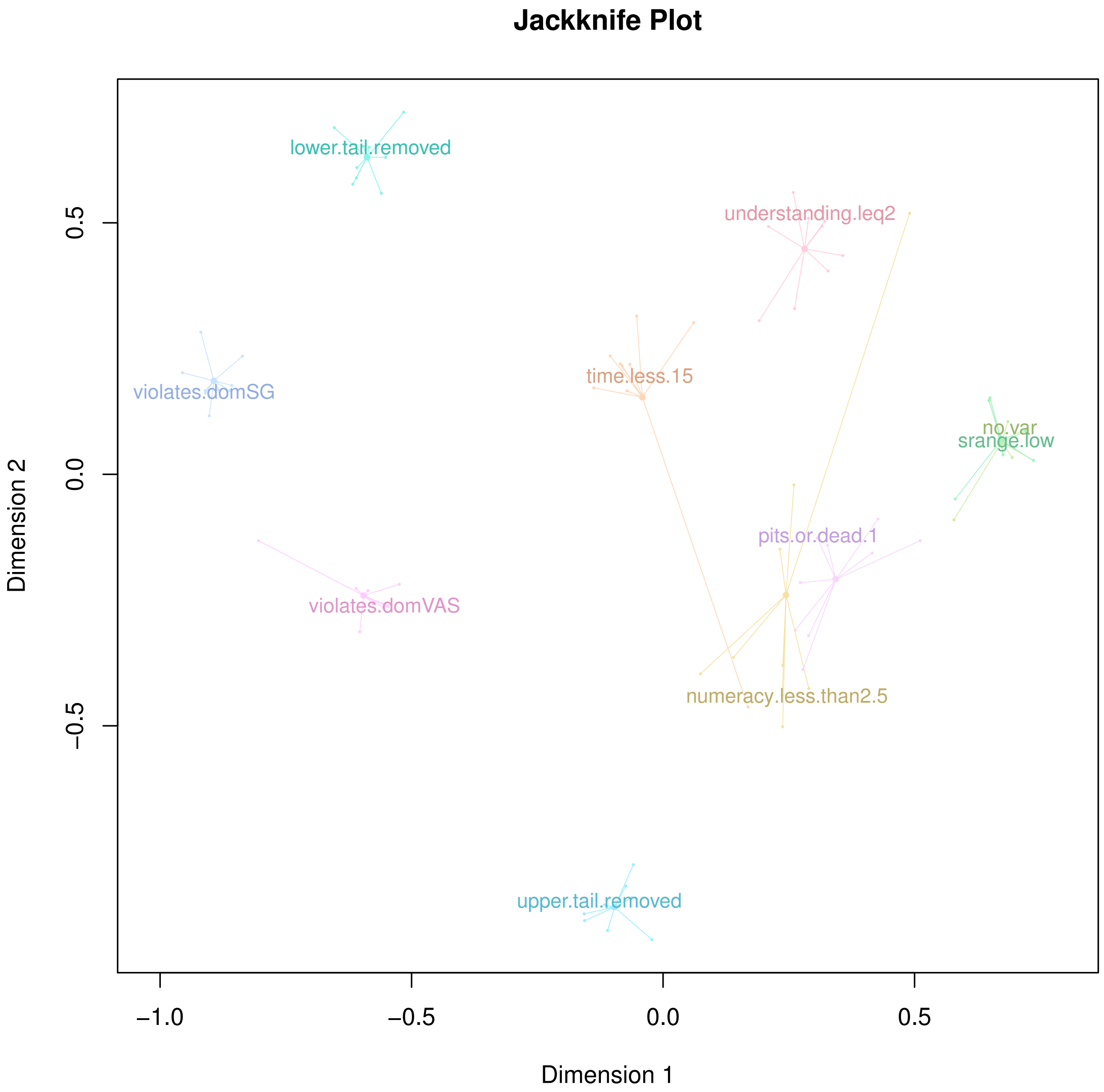


Figure A.10: *Jackknife plot for core configuration.* A jackknife plot corresponding to Figure 5.1, showing how each criterion moves when one of the others are removed and the MDS is re-computed. The labels show the original configuration, and the centre points the centroid of all of the leave-one-out MDS solutions.

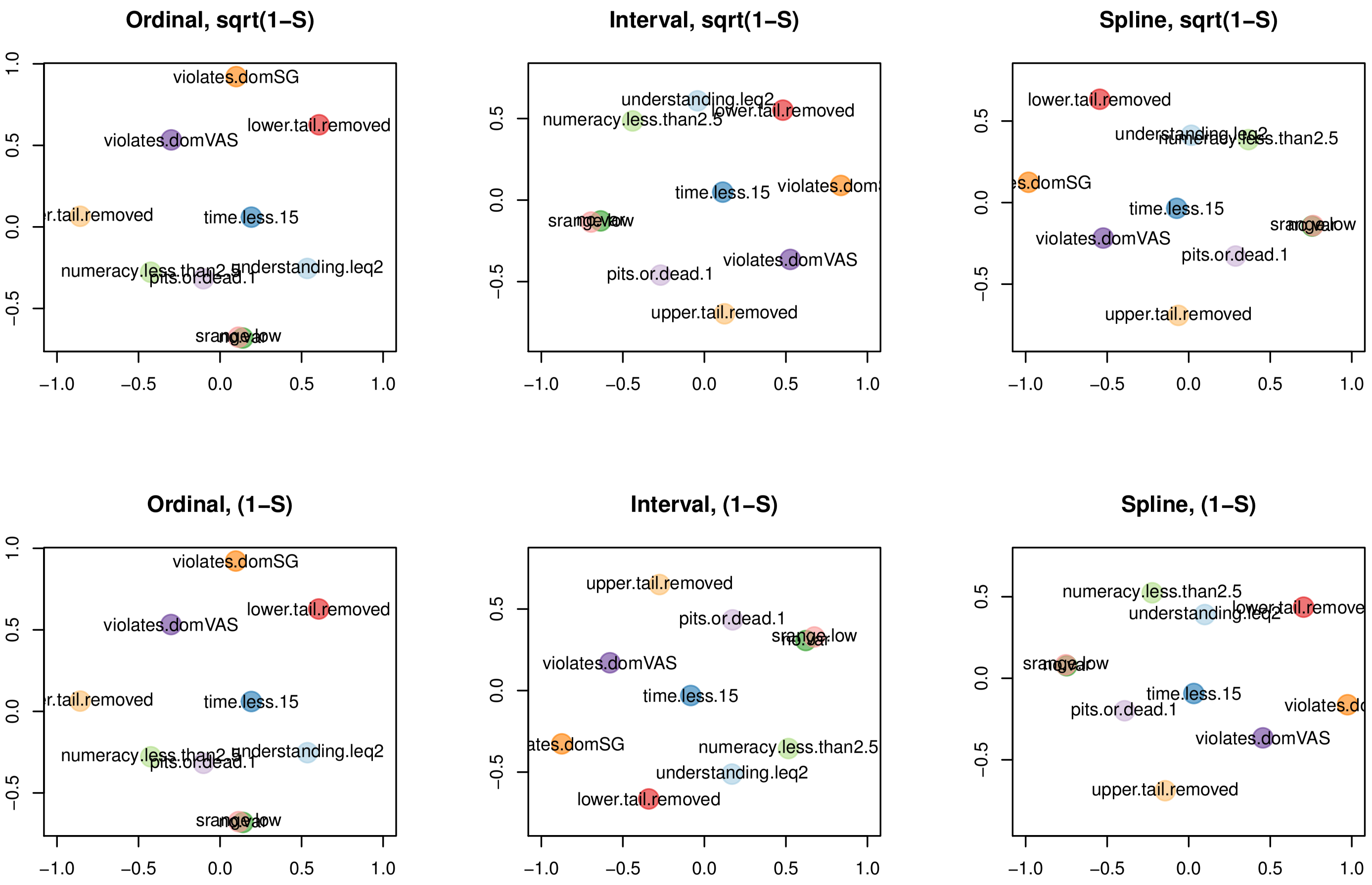


Figure A.11: *MDS using different algorithms and transformations of ϕ.* Each panel of this figure shows the MDS configuration using one of ordinal, interval, or spline MDS, and either the $1-\phi$ or $\sqrt{1-\phi}$ transformation of *phi* into an index of dissimilarity. The leftmost column shows two identical plots (by definition), equivalent to the core MDS configuration (Figure 5.1).

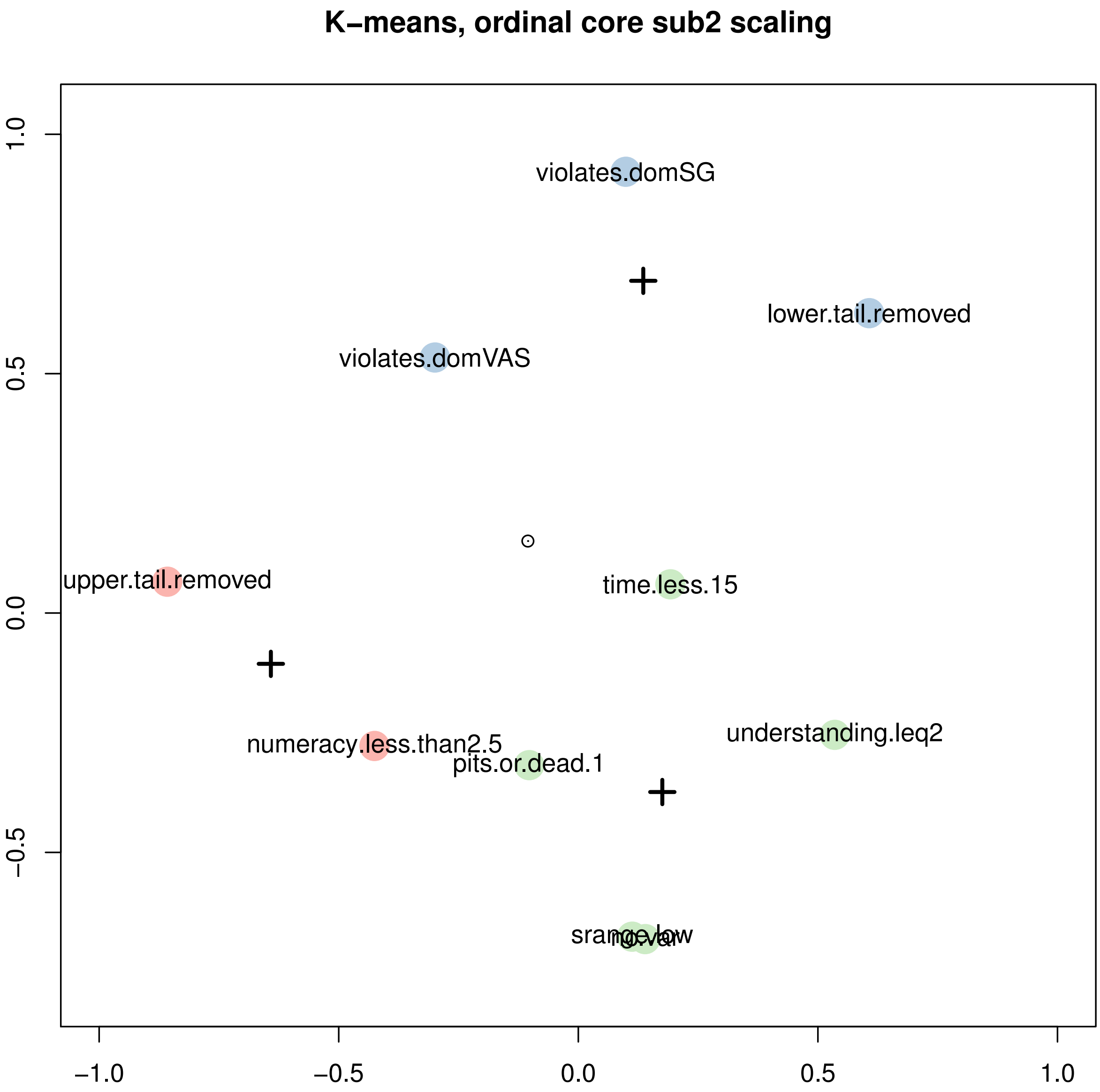


Figure A.12: *k-means on core MDS configuration.* Figure 5.1, with a colouring of the criteria demonstrating the solution to k-means clustering with $k = 3$. The crosses indicate the centroids of the clusters, and the empty dot the centroid of the entire configuration. K-means is a computational search for structure in the configuration, analogous to the researcher using his or her own eyes to see patterns in the configuration.

b) *MDS jackknife*: The jackknife procedure and the associated dispersion statistic suggest that the core solution is robust. The large change in *numeracy* and *time* when *upper-tail* is removed is notable, and could indicate that *upper-tail* fills a distinctive part of the space. Furthermore, the stress of that MDS solution compared to the core MDS solution with *upper-tail* deleted differ by a small amount (0.0892 versus 0.0811), indicating that the improvement in fit by moving *numeracy* and *time* is not substantial. (For perspective, the interval MDS solution of the core criteria has a stress of 0.270 versus the ordinal MDS solution, which has a stress of 0.110.)

c) *MDS algorithm*: There is a change in the arrangement of *numeracy*, *understanding*, and *dead-all-worst* among the other MDS models (interval and spline). There is also more proximity between *lower-tail*, *understanding* and *numeracy* than in our core configuration. However, the rest of the structure is similar, and we are hesitant to interpret these changes given that the interval and spline MDS models require *more* assumptions than the ordinal scaling. It is not clear that our data meet these assumptions (Gower & Legendre, 1986).

d) *Clustering*: The *k*-means clustering does not provide us with any additional information: the groupings do not lead to new conclusions that are not readily available from the configuration without the clustering. Furthermore, the categorization of *numeracy* and *dead-all-worst* in different clusters, despite their proximity – they are the second closest pair in the entire configuration – is counter-intuitive.

A.4 Beta regression

A.4.1 Details of the beta regression sensitivity analysis and model-checking

a) *Functional form*: Comparing the continuous and factor models provide one assessment of the appropriateness of writing the linear predictor as a linear function of *theta* in equation (6.1). The factor model is the extreme of non-linearity, as it calculates the utility value for a given theta using only information from that theta value, and none of its neighbors. We leave the same type of analysis for the three-way interaction models (equation (6.3)) for future work, given its computational challenges. In the interim, we compare the sample means to the three-way interaction model that is linear in theta, as those means are unbiased estimates of the population means, providing one assessment of the model specification.

b) *Residual analysis*: Residuals for beta regression are constrained by the bounded scale of the beta regression. To examine the fit of the beta model, we simulate a beta distribution data-generating process and fit a *correctly specified* model. We do this for a case not requiring squeezing, and a case requiring squeezing where beta regression still nearly recovers the correct parameters. We then compare the residual versus fitted values plots with those from our data.

c) *Zero-one inflated beta regression*: There is an alternative to the squeezing procedure, that still allows one to take advantage of the benefits of beta regression (e.g., explicit modeling of the variance), called a *zero-one inflated beta* (ZOIB) model (Liu & Kong, 2015). Essentially, this is a mixture model: two binomial models, estimated via logistic regression, describe the 0 and 1 responses, while a third model describes all the data between 0 and 1 through beta regression.

One disadvantage of the ZOIB approach is that it assumes that there are two processes producing responses, which might not be appropriate depending on the setting: one described by two binomial distributions (producing 1s and 0s) and the other described by a beta distribution. A second is that it is more computationally intensive. A third is that, if we assume each part of the model has the same linear predictor (e.g., $\eta = \beta_0 + \beta_1 theta + \beta_2 criterion + \beta_3 theta : criterion$), a ZOIB model is described by twice the number of coefficients than the equivalent squeezed beta regression model, because we are estimating not only the μ and ϕ parameters of the beta but also the means (proportions) of the two logistic regressions.

An advantage of the ZOIB model is that it can better reproduce the data, because of its additional flexibility. In our context, the parsimony of the squeezed beta regression models is useful, because it allows us to easily compare the effects of the different criteria. However, to explore the fit of the squeezed model, we compare it using some of our data to three approaches to ZOIB, which differ in their estimation procedures: a Bayesian approach using Markov chain Monte Carlo (MCMC) sampling, from the R package **zoib** (Liu & Kong, 2015); separate logistic regressions on the $0/1$ data using R's default function for estimating general linear models, which estimates parameters via iteratively reweighted least squares, combined with a beta regression using the **betareg** package on the *un-squeezed* $(0, 1)$ data;[4] and, a method that uses simulated annealing (a type of numerical optimization) to find the maximum of the joint likelihood of the two logistic regression and the beta regression (Davis, 2017b).

A.4.2 Summary results of the beta regression sensitivity analysis and model-checking

a) *Functional form*: As described earlier, comparing the models in equation (6.1) and (6.3), which are linear in theta, to their associated factor models, where theta is treated as a categorical variable, allows us to test the linearity assumption. For example, the model of those not excluded by *numeracy* (the dotted curve in Figure 6.10) tracks its associated factor model (open dots) well, while the model of those excluded (the solid curve) has some large differences with its factor model (solid dots). This demonstrates that there could be a model that is non-linear in theta and has a better fit for the excluded. Overall, however, most of the continuous models across the domains and the exclusion criteria track their factor counterparts well. As mentioned earlier, performing the same procedure for the three-way interaction models (equation (6.3)) requires further work, as the computational tools to perform this task need to be built. In the interim, we examine the functional form of our three-way interaction models (Figure 6.19 and Figure 6.20) by plotting the best-fit curves against the sample means (Figure A.13, Figure A.14, Figure A.15), as the sample means are unbiased estimates of the mean. These figures showed that the interaction models appear to be well-specified (i.e., they mostly recover the sample means). Further detail is provided below.

[4] We will refer to this method, which uses the `glm()` function in R for the two logistic models and `betareg()` in the **betareg** package for estimating the beta model, as the "double-logistic plus beta" version of ZOIB, because it estimates coefficients by running those models separately. Of course, formally, the other methods are also a mixture of two logistic regressions and a beta regression, but their implementation in R differs from simply combining `glm()` and `betareg()`.

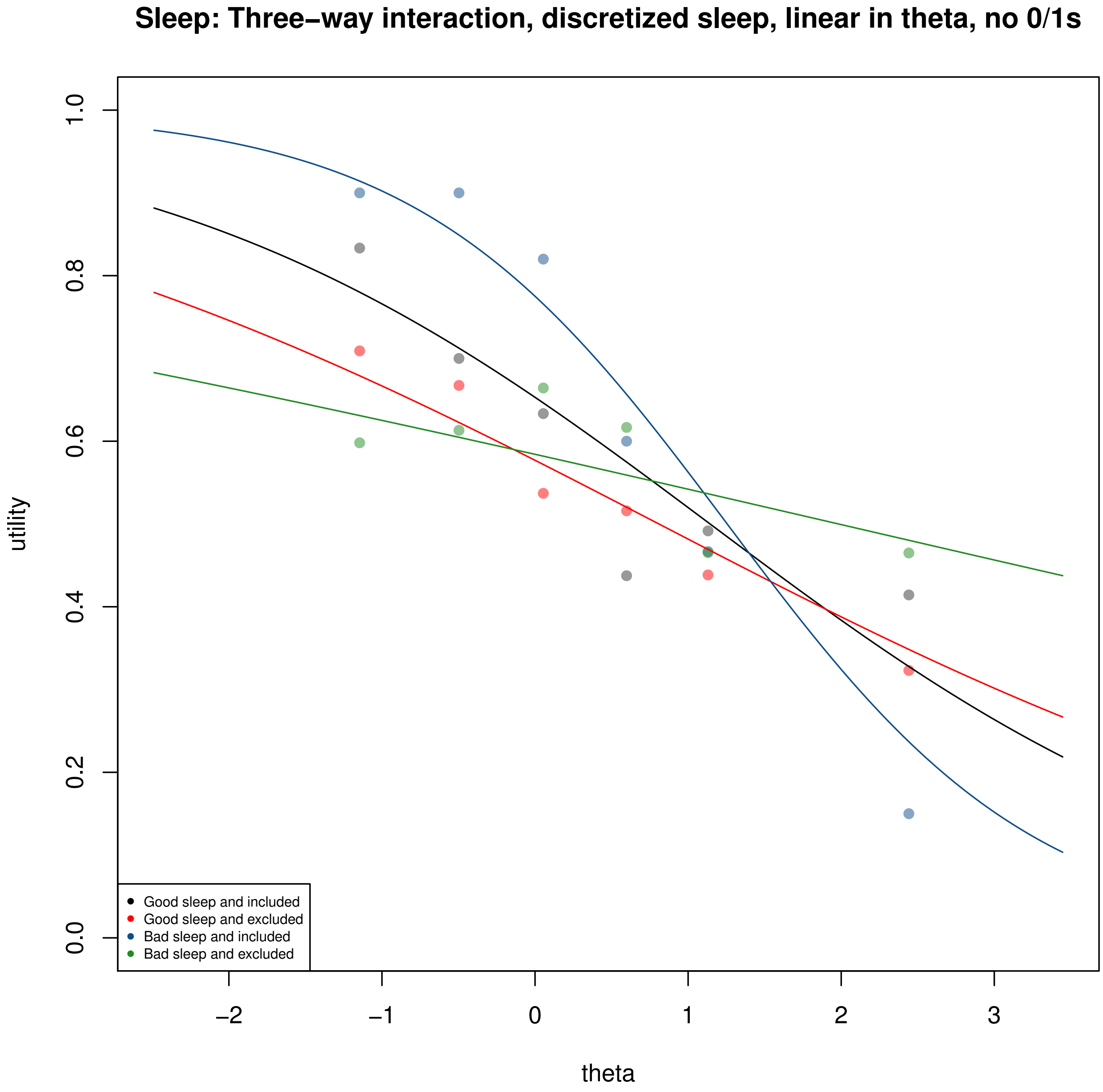


Figure A.13: *Utility curves and sample means for a version of the model for sleep utilities as a function of theta, violates-SG, and discretized participant sleep quality, with no 0s or 1s in the data.* Sample mean utilities and the conditional mean curves estimated for four groups: those who sleep well and are included by *violates-SG* (black), those who sleep well and are excluded (red), those who sleep poorly and are included (blue), and those who sleep poorly and are excluded (green). The conditional mean curves come from the beta regression model in equation (6.3), where PROMIS scores on sleep are discretized as in the main text, and utilities of 0 and 1 are removed. That is, the figure shows the beta regression portion of the ZOIB model of these data. This figure is formally equivalent to Figure 6.15 in the main text. Finally, notice the improved fit between the lines and points over the squeezed model in Figure A.23, shown below.

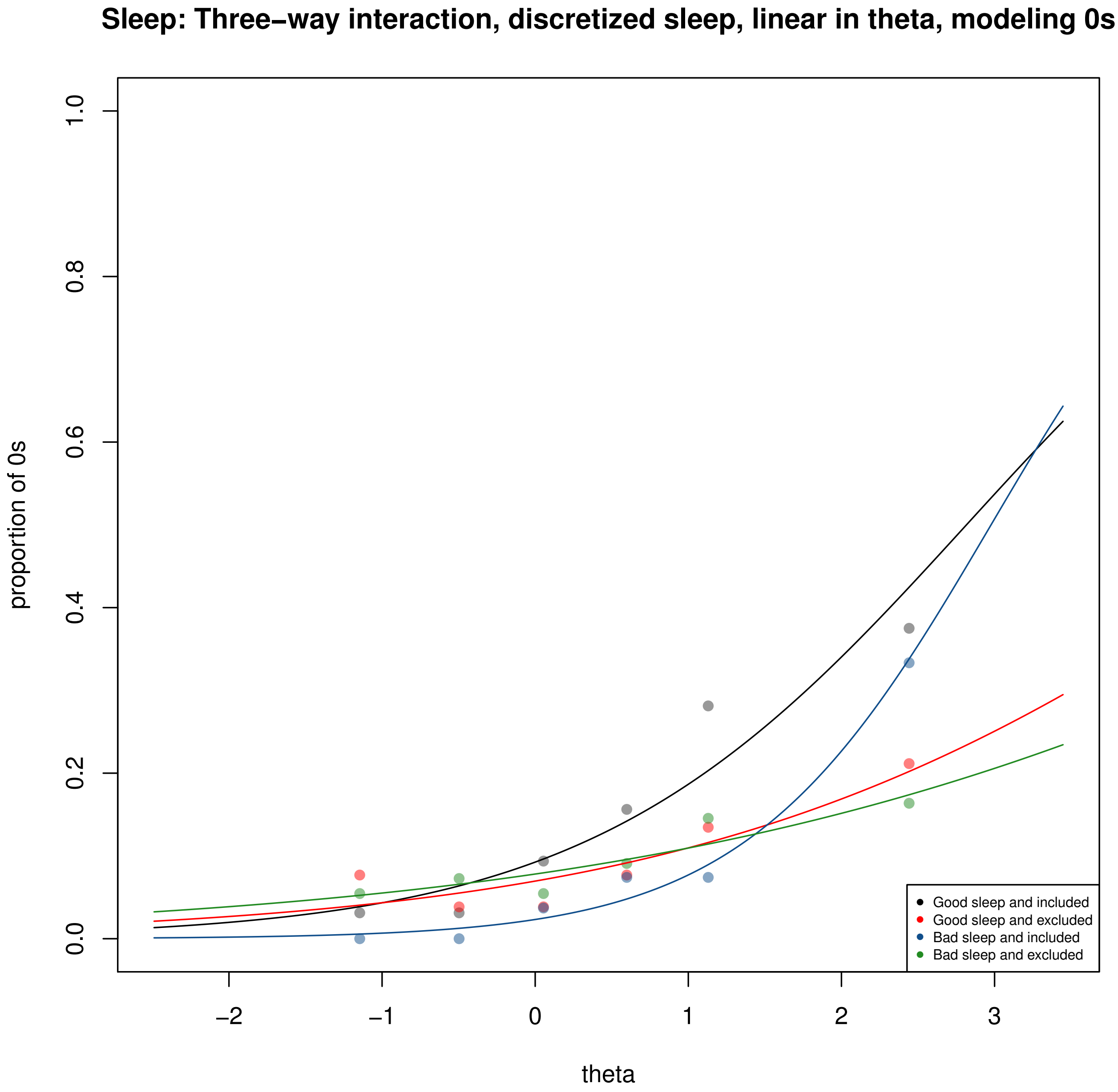


Figure A.14: *Expected proportion of 0s from logistic part of ZOIB model for sleep utilities as a function of violates-SG, theta, and participant sleep health, as well as sample proportions of 0s.* As in Figure A.13, this figure is equivalent to Figure 6.16, but includes the sample mean proportions.

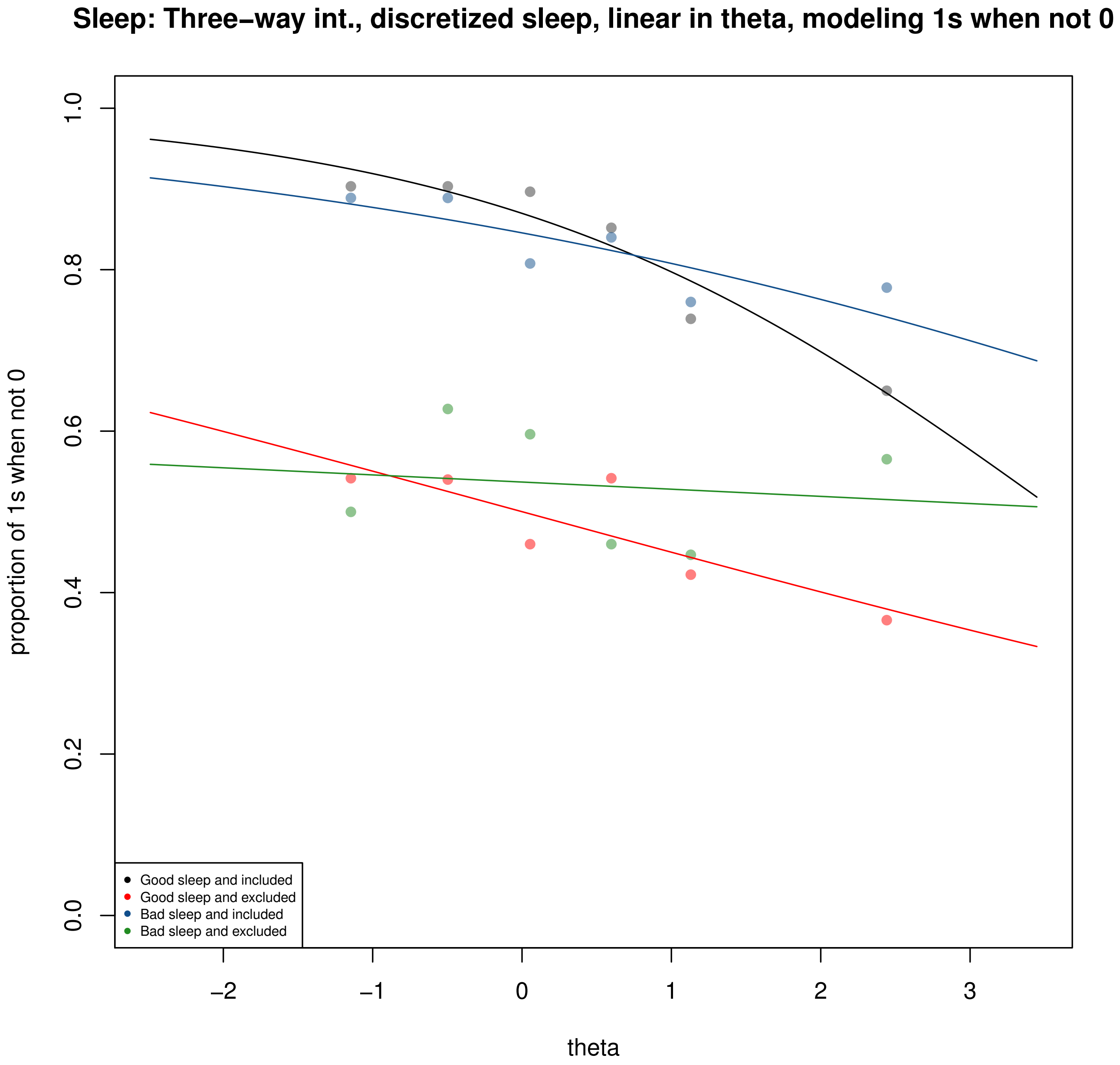


Figure A.15: *Expected proportion of 1s from logistic part of ZOIB model for sleep utilities as a function of violates-SG, theta, and participant sleep health, as well as sample proportions of 1s.* As in Figure A.13, this figure is equivalent to Figure 6.17, but includes the sample mean proportions.

b) *Residual analysis* and c) *Zero-one inflated (ZOIB) regression*: Examining the residuals of our main models from equation (6.1) (Figures 6.1-6.9) shows that they systematically vary from those expected from a conditionally beta-distributed random variable because of the number of 0s and 1s in the data at every value of theta. Moving to the ZOIB models necessarily produces a better fit. However, differences between the included and excluded groups remain, and are, in fact, more pronounced in the ZOIB models. Thus, we believe the squeezed models provide good within-sample comparisons for the main models (equation (6.1)), which are the focus of our study.

In order to move to the ZOIB models for all of the models in Figures 6.1-6.9, and in order to extend the three-way interaction analysis to all criterion and domain combinations, we need to advance the currently available computational tools for estimating beta regressions. That task adds a pure methodological component to our future work, and its end-product will have applications beyond the current study (Davis, 2017b; Smithson & Verkuilen, 2006).

A.4.3 Detailed discussion of the beta regression sensitivity analyses

Residual analysis

The correctly specified beta model with no squeezing shows how residuals in a beta regression are affected by the boundedness of the beta distribution (Figure A.16). Large residuals are only possible near the endpoints. For example, when the fitted values approach 1, the beta distribution – which always places some probability density over the whole open interval – will sometimes produce small values, producing large negative residuals. In the figure, the residuals “step down” as one moves to the right because the beta is skewed unless the two shape parameters are equal, which necessitates a (fitted) mean of 0.5. Otherwise, a mean value less than 0.5 necessitates a right skew, and one greater than 0.5 necessitates a left skew, producing many positive and negative residuals, respectively.

The residuals versus fitted values plot for the correctly specified beta model requiring squeezing, in which beta regression nearly recovers the correct coefficients, shows nearly the same pattern (Figure A.17). That provides us with confidence that squeezing can work well.

Figure A.18 shows residuals versus fitted values for the model of cognition utilities as a

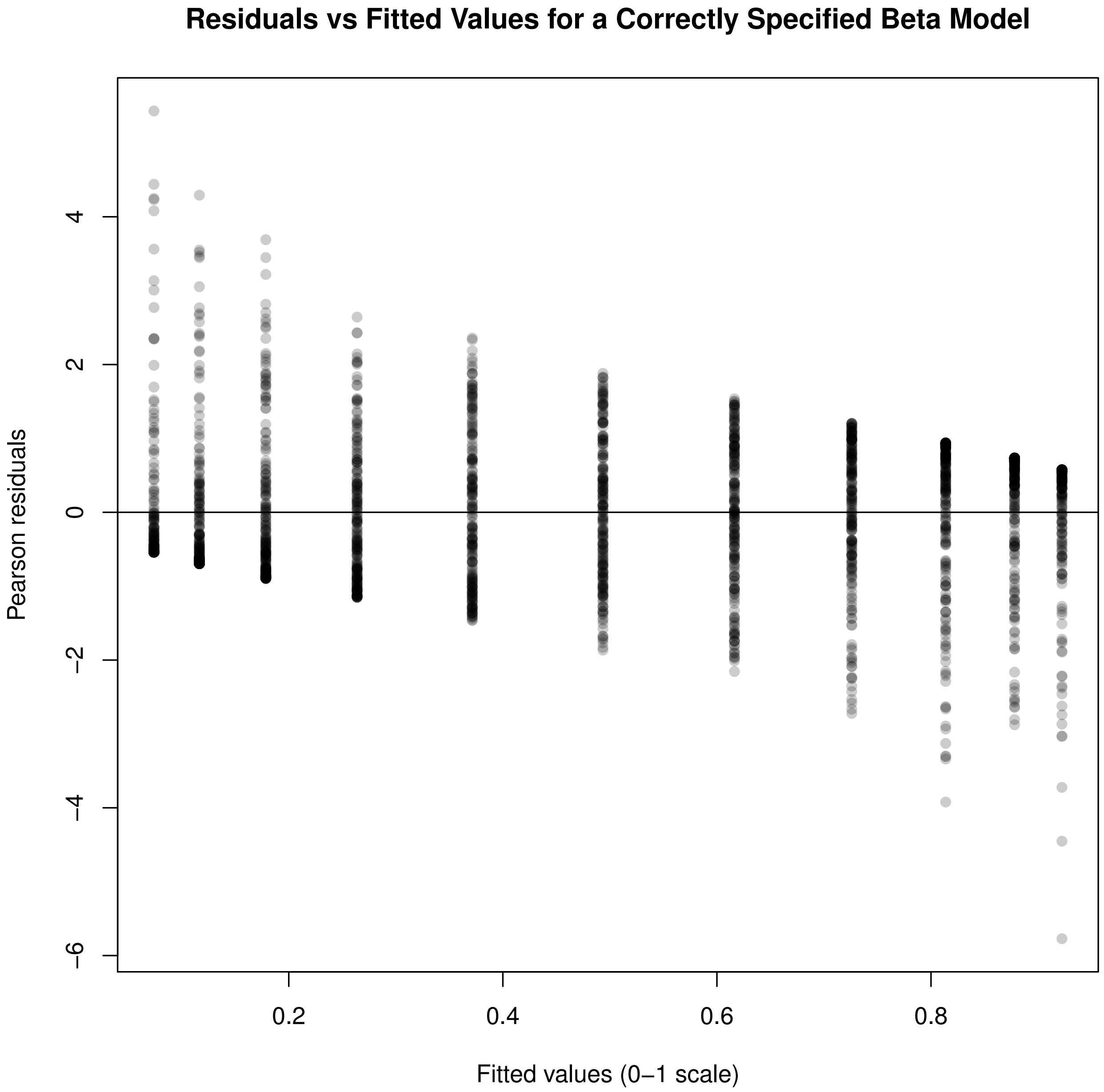


Figure A.16: *Residuals versus fitted values for a simulated, correctly specified beta model.* Plot of residuals versus fitted values for a simulated, correctly specified beta model, where beta regression recovers the correct parameters.

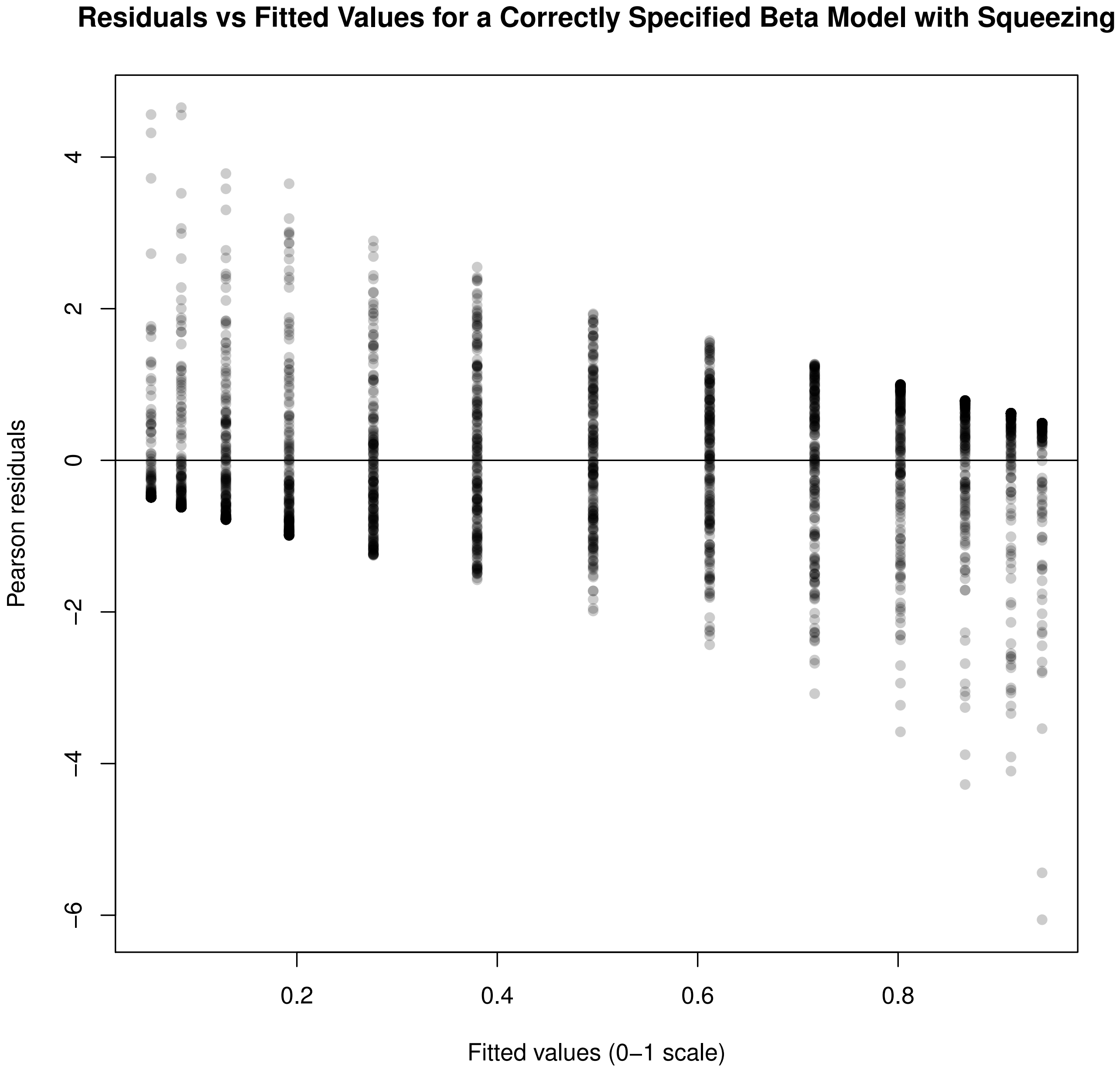


Figure A.17: *Residuals versus fitted values for a simulated, correctly specified beta model requiring squeezing*. Plot of residuals versus fitted values for a simulated, correctly specified beta model.

function of *theta* and the *time* criterion, i.e., whose mean is modeled by:

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 theta_{cognition} + \beta_2 time + \beta_3 theta_{cognition} : time.$$

It shows a similar pattern to Figure A.16 and Figure A.17. However, the columns of residuals at each fitted value have much more consistent length. That is due to the fact that every observed conditional distribution has some 0s and 1s in the data. The spread of the residuals for a fixed fitted value is discrete and often evenly spaced because of the SG elicitation method: utilities could only be given for every 0.05 of the 0-1 utility scale. Notice, too, that the fitted values *start* just above 0.5, in contrast to our simulations. That is, in our data we are never modeling distributions whose means are close to 0.

As a comparison, Figure A.19 shows the residuals versus fitted values for the same model using the cognition data and *excluding* the 0s and 1s from the dataset – and thus not requiring squeezing. The pattern is very similar to the squeezed model, although there is a more pronounced lowering of the columns of residuals, more closely matching the residual plot of the correctly specified model (Figure A.16). Note, too that the range of fitted values is shifted compared to Figure A.18, and that we do not get close to either endpoint of the scale. There are significantly more 1s than 0s – 574 versus 150 in the cognition domain – and so removing all the 0s and 1s shifts the (fitted) means downwards.

Given the patterns of residuals, we believe beta regression provides a useful method for summarizing our data. Although it does not appear that the conditional distributions are exactly beta distributed conditional on our covariates, the general shape of the residual plots reflects some of the key structural features of correctly-specified beta regressions.

We also believe the squeezing procedure is theoretically legitimate for our data. By assumption, the utilities are on an interval scale; more importantly, they are *not* on a ratio scale. Thus, the 0 is not an absolute, as it would be in a scale for mass or length. In fact, throughout the estimation of the PROPr scoring system (Part II), we often translate between different utility scales where 0 corresponds to the utility of various states (e.g., dead, the all-worst state). Therefore, moving the 0s in the data to $0.5/n$ via a linear transformation is conceptually equivalent to the other translations

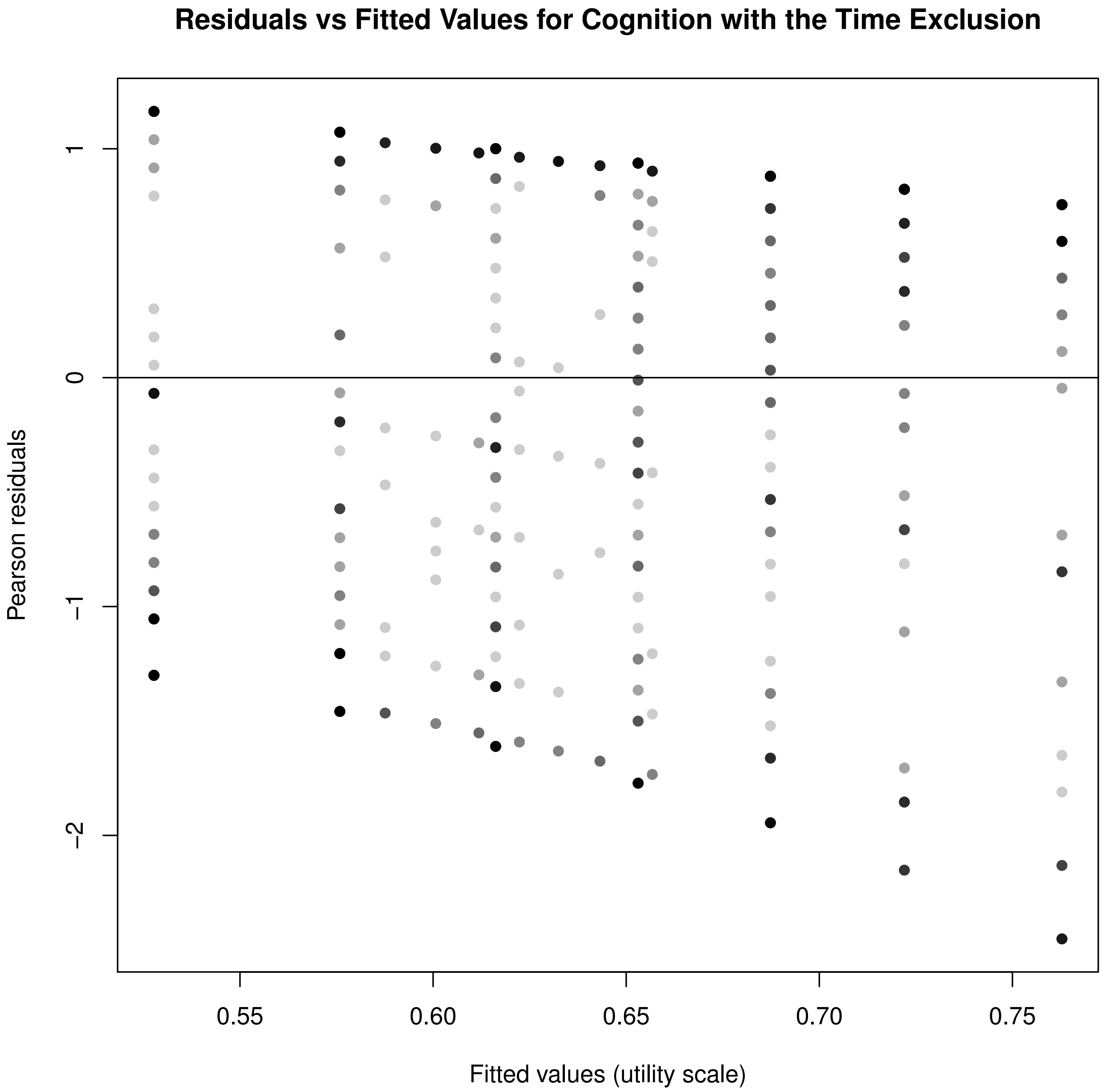


Figure A.18: *Residuals versus fitted models for cognition as a function of theta and time.* Plot of residuals versus fitted values for the squeezed model $\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 theta_{cognition} + \beta_2 time + \beta_3 theta_{cognition} + \beta_3 theta_{cognition} time$.

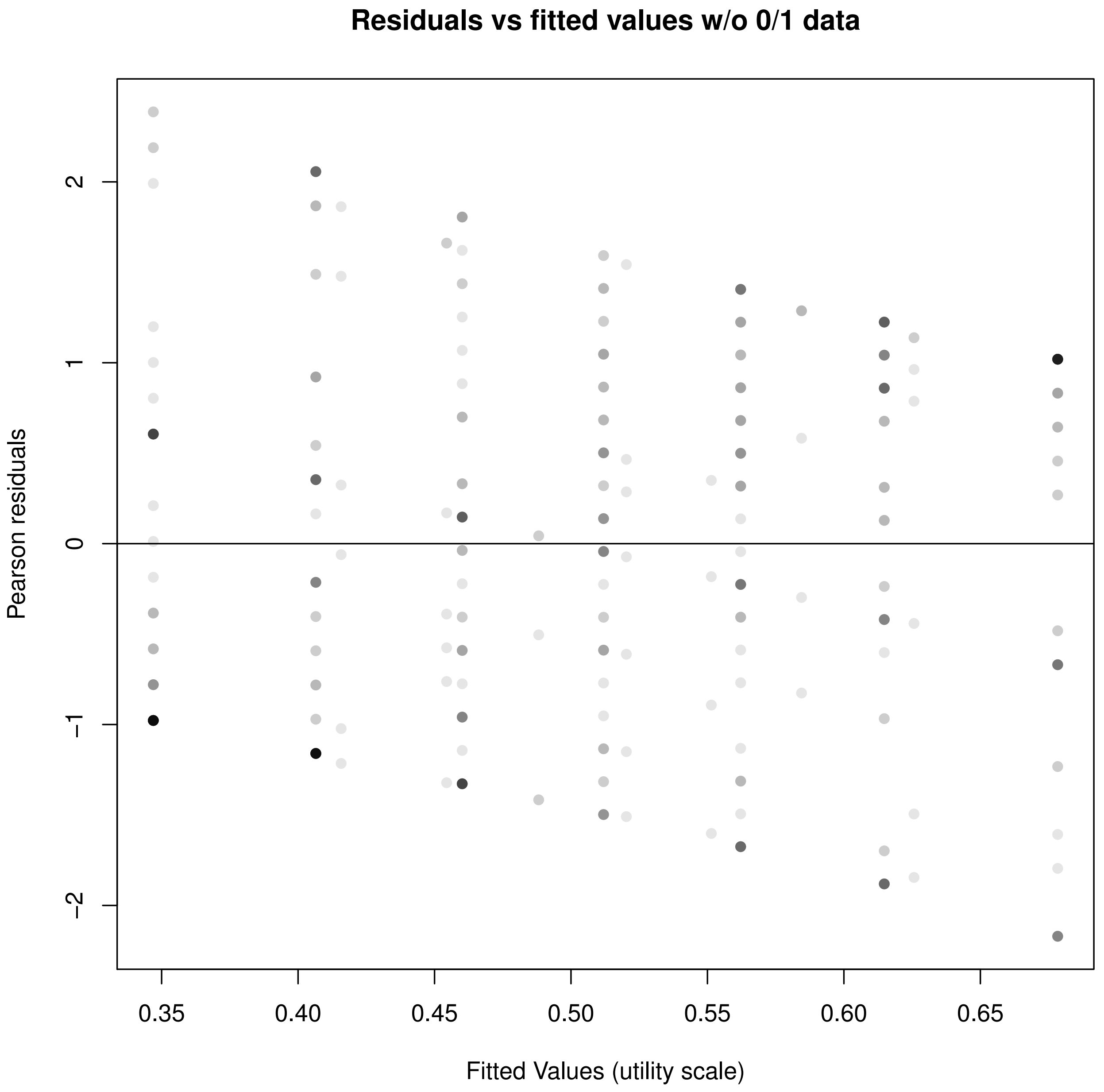


Figure A.19: *Residuals and fitted values for cognition utilities as a function of theta and time, no 0s or 1s in the data.* Plot of residuals versus fitted values for the model $\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 theta_{cognition} + \beta_2 time + \beta_3 theta_{cognition} : time$, using the cognition data without the 0 and 1 responses. (Note this would be the beta portion of the associated ZOIB model.)

of 0s that occur throughout the PROPr scoring system: the interval-scale properties of the data are retained, but the data is transformed for modeling purposes.

By construction, the single-domain utility data is built such that a state with more extreme functioning than the endpoints – that is, with a theta value outside of the values in Table 3.1 – is given a utility equal to 0 or 1 (depending on which side of the scale it lies). The societal valuation of a state – usually taken to be a mean value – is always between the endpoints, and in practice never equals those extremes, unless it is one of the pre-defined endpoints (i.e., full health and the all-worst state or dead). Thus, making it impossible for an intermediate state to have predicted mean utility of 0 or 1 is not, within the context of health state valuation studies, an unusual assumption.

Zero-one inflated beta (ZOIB) models

To further investigate the squeezing method used in our modeling, we used the data from the cognition domain and compared its results with those of the three ZOIB alternative models. (Recall that, if squeezing were unnecessary, the ZOIB models would reduce to normal beta regression.)

We estimated conditional mean curves for the four models where the linear predictor is only *theta* (i.e., where there are only two coefficients, the intercept and the slope on theta). Figure A.20 shows the squeezed model (solid black), the double-logistic plus beta version of the ZOIB model (solid red), the Bayesian ZOIB model (dashed brown), the ZOIB model estimated via simulated annealing (dashed blue), and the sample mean utilities of the un-squeezed data (dots). The deviations between the black curve and the sample means suggest some bias in the squeezed model. The Bayesian ZOIB and ZOIB estimated via simulated annealing are almost collinear, and closely reproduce the sample means. The double-logistic plus beta regression have a shape like the two other ZOIB models, but a different intercept.

We also estimated conditional mean curves for the four models where the linear predictor included an exclusion criteria (*time*), as in equation (6.1). These curves, two for each model – one for those participants excluded by *time* and one for those left included in the sample – are plotted for each method in Figure A.21. The double-logistic plus beta regression (red) and Bayesian ZOIB (blue) are almost collinear, while the included curve of the ZOIB using simulated annealing (dotted brown) is close to the included curve of the other two ZOIB models, but its excluded curve (solid

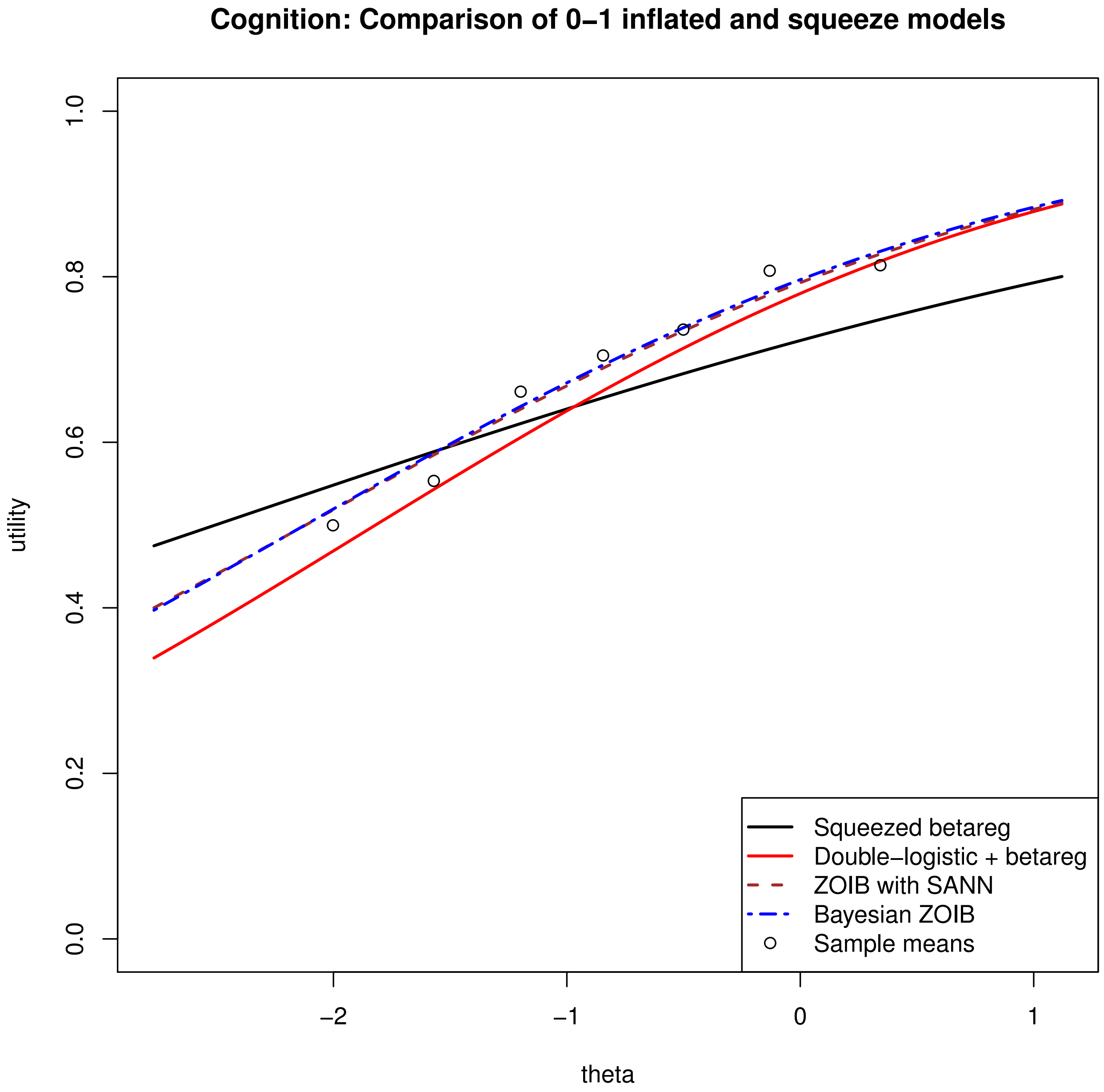


Figure A.20: *Comparison of ZOIB estimation techniques using the cognition data.* A comparison of beta regression on squeezed data (black) and three methods to estimate zero-one inflated beta (ZOIB) models: two logistic regressions on the 0-1 data plus a beta regression on the (0, 1) data (solid red), ZOIB estimated using simulated annealing for maximum likelihood estimation (dashed brown), and a Bayesian ZOIB approach that uses Markov chain Monte Carlo sampling to produce coefficient estimates (dashed blue). The sample means from the data are plotted as points. Every component of every model has the same linear predictor, $\eta = \beta_0 + \beta_1 theta$.

brown) is further from their excluded curves. The two parts of the squeezed results (black) have a similar shape to the others, but show less extreme slopes in both curves than the ZOIB models, and also show the bias in the excluded (solid) and included (dotted) curves that is seen in Figure A.20.

We implemented the ZOIB models to see the effect of modeling the untransformed data, and compare those models with our main, squeezed models. As seen in Figure A.20, the ZOIB models more closely recover the sample means, which are unbiased estimates of the conditional mean utilities. As there are many 0s and 1s in the data, squeezing them into data in the open interval $(0, 1)$ then requires the optimization algorithm to search for beta parameters that will produce many high and low values, and recover the mid-range of values as well. Unless the shape of the empirical distribution of the data $(0, 1)$ has peaks at the extremes, this procedure will necessarily do a worse job at recovering the shape of the $(0, 1)$ data, as a beta distribution cannot take on every shape imaginable (e.g., it can have at most two modes). The ZOIB models fit a beta regression on only the data in the open interval $(0, 1)$, allowing its beta regression to have fewer constraints. Thus, the differences could be caused by squeezing, or the different model assumptions (i.e., the multiple data-generating processes in ZOIB models) or even over-fitting from the more flexible ZOIB models, or by a combination of factors.

In Figure A.21, where the model also includes the interaction, we see that the differences between the included and excluded curves are more pronounced than in the squeezed beta regression. This could be because squeezing brings all the data closer together, thus diminishing the differences between otherwise untransformed utilities. It is promising that the general shapes of the curves are the same in the squeezed and the ZOIB models, and that any relationships are only more pronounced in the more complicated models. Thus, we believe it is more likely we have missed an effect of the exclusion criteria than produced one that would disappear with a ZOIB analysis. Again, two of the ZOIB estimation procedures (the Bayesian and double-logistic plus beta regression) closely recover the sample means.

Based on the results of our comparisons, we believe our squeezed models are best interpreted as parsimonious descriptions of the relationships between utility, health domains (i.e., theta), and exclusion criteria, allowing within-sample comparisons of those relationships, rather than as descriptions of mean population utility estimates. Future work is needed to determine the role of

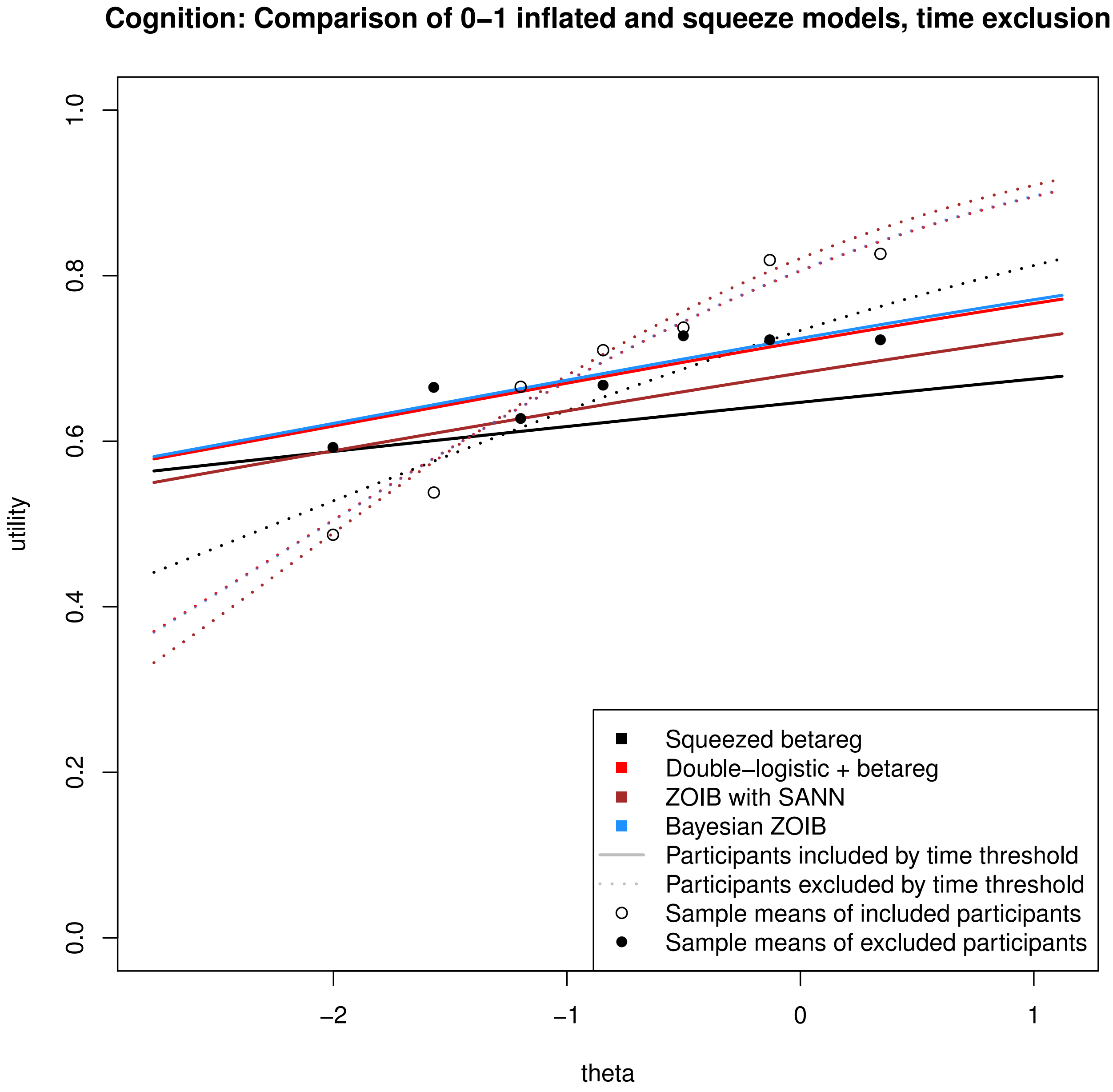


Figure A.21: *Comparison of ZOIB estimation techniques using the cognition data and time.* A comparison of beta regression on squeezed data (black) and the same model estimated using three different methods for zero-one inflated beta (ZOIB) regression on the un-squeezed data: two logistic regressions on the 0-1 data plus a beta regression on the (0, 1) data (red), ZOIB estimated using simulated annealing for maximum likelihood estimation (brown), and a Bayesian ZOIB approach that uses Markov chain Monte Carlo sampling to produce coefficient estimates (blue). The solid curves show the estimated conditional mean values for those participants excluded by the time criterion; the dashed curves show the estimated conditional mean values for those participants left in the sample. Every component of every model has the same linear predictor, $\eta = \beta_0 + \beta_1 theta + \beta_2 criterion + \beta_3 theta : criterion$.

the potential causes of the differences seen between the ZOIB and squeezed models, which include the squeezing, the (in)appropriateness of the beta distribution for describing the data conditional on our chosen covariates, and the extra parameters estimated in the ZOIB models. Moving to the ZOIB models in all aspects of the work could improve the modeling, but would require additional insight to interpret these complicated models for those who wish to use them to inform their survey design – in contrast to those who want to run their own ZOIB models, who might appreciate the thorough methodological discussion. It is unclear whether there is a large (or any) intersection of these two groups, so care must be taken when proceeding with an even more complex modeling strategy.

Details of sensitivity analyses for incorporating participant health status

In our analysis of including participants' PROMIS sleep scores as a predictor in our model for mean sleep utilities, we first ran the model from equation (6.3) with all the data – a necessarily squeezed model – keeping the PROMIS scores as a continuous variable. Figure A.22 shows that model's residuals vs fitted values. The figure displays the residuals separated by the four groups described earlier, even though the PROMIS sleep scores were not discretized as covariates in this model: those who sleep well and who were not excluded by *violates-SG* are shown in black; those who sleep well and were excluded are shown in red; those who sleep poorly and were not excluded are shown in blue; and, those who sleep poorly and who were excluded are shown in green. The residuals do not suggest that the beta is a good fit for the conditional distributions. In particular, the residuals for those included with good or poor sleep quality deviate from the typical beta pattern seen earlier (Figure A.16 and Figure A.17). We also ran the squeezed model with the discretized sleep scores. The sample means and conditional mean curves from that model are displayed in Figure A.23. The figure shows that the model is biased, and does a poor job recovering the sample means for the groups. For example, the model for the poor sleep and included group (blue curve) is systematically lower than the sample means.

Given Figure A.22 and Figure A.23, it does not appear that the squeezed data is conditionally beta distributed. Thus, we moved to the ZOIB models, focusing on the $(0, 1)$ data, as reported in the main text.

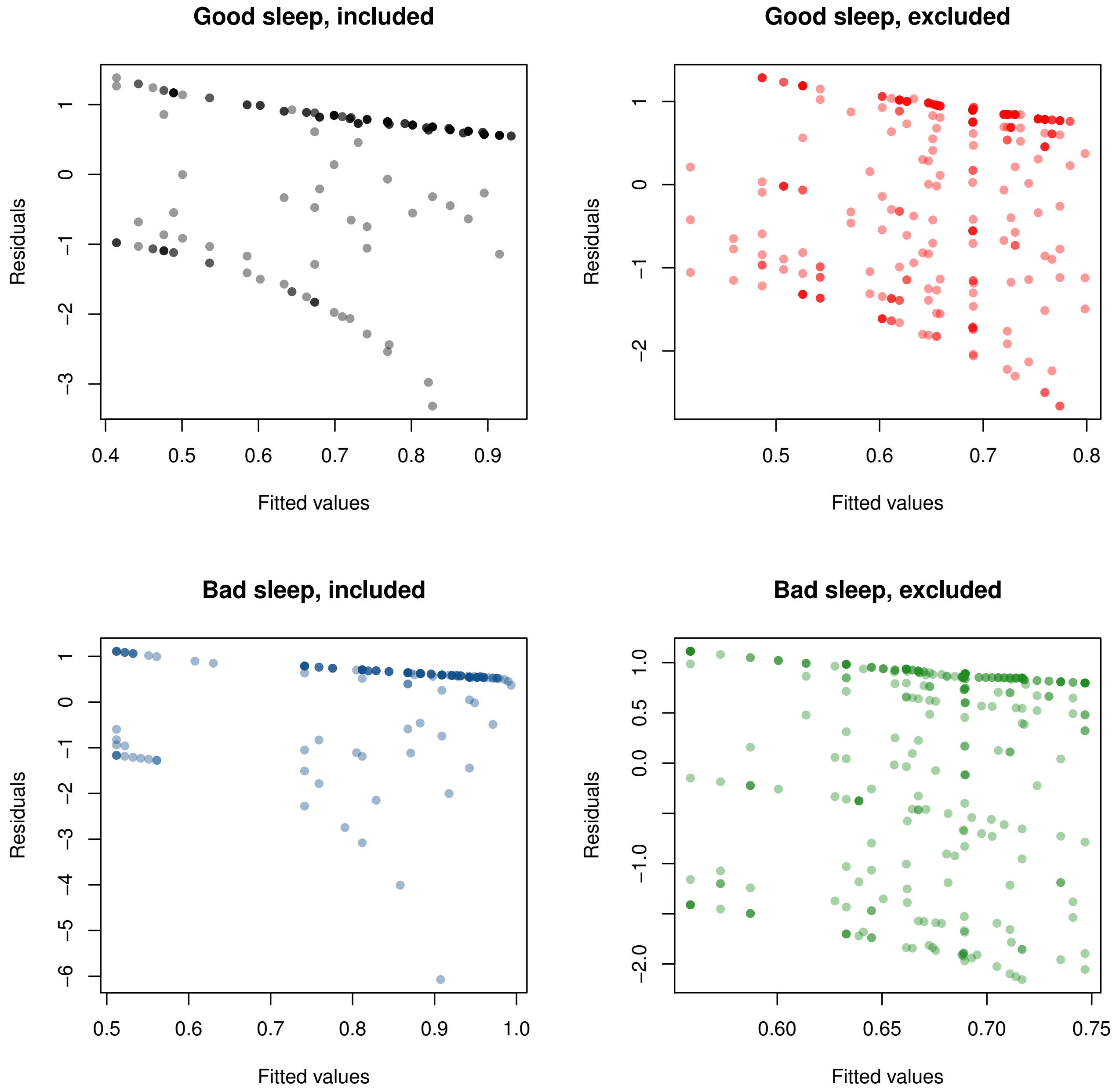


Figure A.22: *Residuals versus fitted values for a squeezed version of the model for sleep utilities as a function of theta, violates-SG, and participant sleep health (treated as continuous).* Residuals from a three-way interaction model (equation (6.3)) of the mean utility for sleep as a function of theta, the *violates-SG* exclusion criterion, and the participants' PROMIS scores on the sleep domain. The residuals are separated into four groups, based on exclusion/inclusion by violates-SG, and their sleep quality, but the participants' sleep quality in the model is treated as a continuous variable (i.e., not discrete).

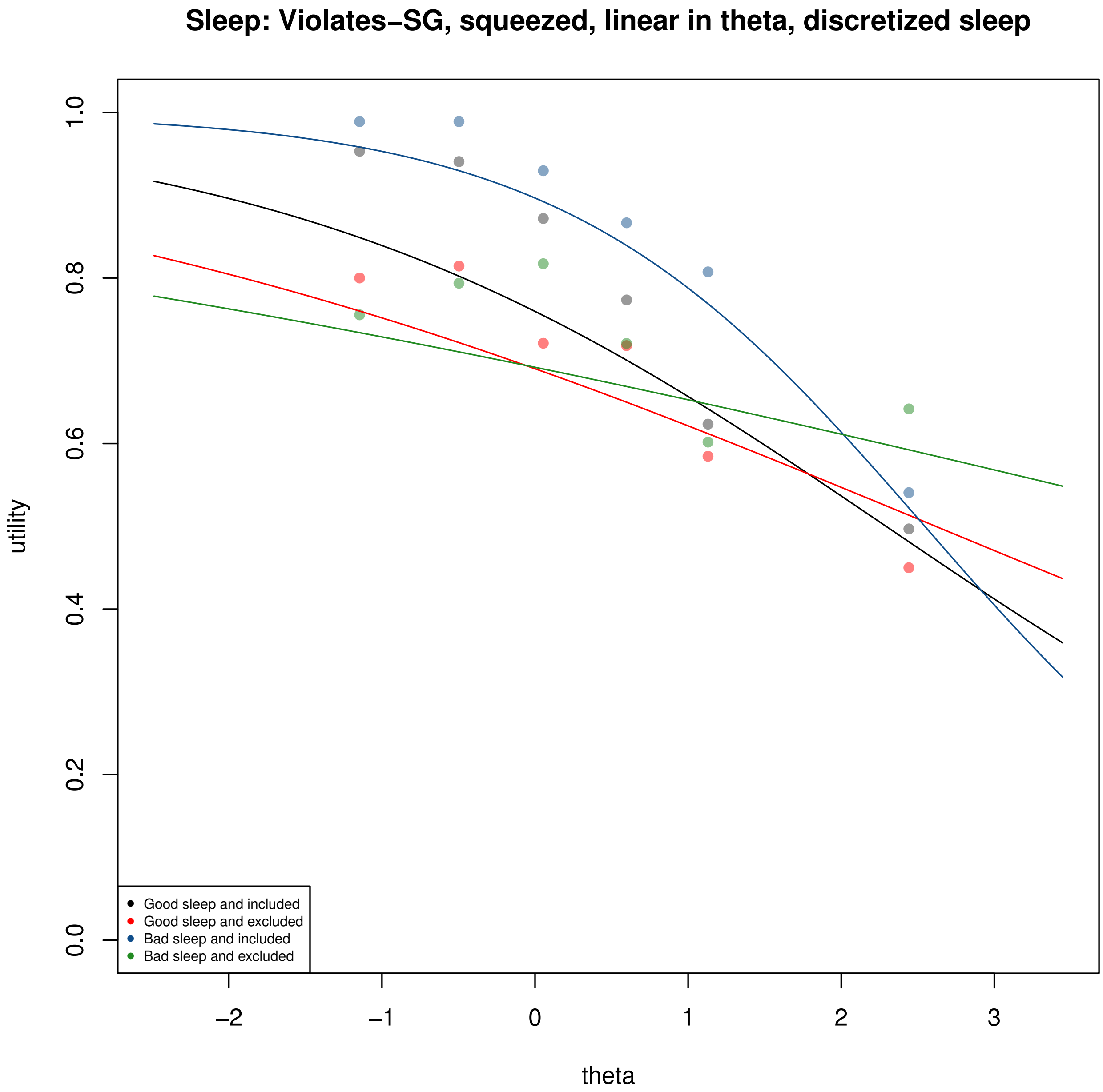


Figure A.23: *Utility curves and sample means for a squeezed version of the model for sleep utilities as a function of theta, violates-SG, and participant sleep health, discretizing participant sleep health into "sleep poorly" and "sleep well".* Sample mean utilities and the conditional mean estimated for four groups: those who sleep well and are included by *violates-SG* (black), those who sleep well and are excluded (red), those who sleep poorly and are included (blue), and those who sleep poorly and are excluded (green). The conditional means come from the beta regression model in equation (6.3), where PROMIS scores on sleep are discretized as described in the main text, and the utility data is squeezed. Notice the poor fit of the sleep model (e.g., the blue line consistently under the blue dots).

First, we estimated the original model, treating PROMIS scores as continuous, removing the 0s and 1s from the data. The residuals are plotted in Figure A.24. These are much closer to exhibiting the typical shape for the beta, although there are too few values in $(0, 1)$ among the included groups to say much about their conditional distributions. As before, we then estimated the three-way interaction model, treating PROMIS scores as discretized – the main model described in the results – and compared the sample means for each of the four groups with the conditional mean curves implied by the model (Figure A.13). This does a much better job of fitting the sample means. For example, unlike in Figure A.23, the curve for the included group who sleep poorly (blue) is not systematically under-predicting the conditional mean utility. However, the non-monotonic sample means in the excluded group who sleep poorly remains, just as when the 0s and 1s were included.

All of the models tried so far have been linear in theta (and in the PROMIS scores, when those have been used as a continuous variable). To see the effect of a non-linear model, we added a quadratic term to the previous model, producing Figure A.25 (note the sample means are unchanged from Figure A.13). By inspection, this does not look like it produces a much better fit, with the predictions for very positive theta values (where there is no data) for the good sleep and included group (black curve) suggesting a non-monotonic relationship between theta and utility. The fit for the bad sleep and included group (green curve) does look better than the straight line in the previous model. The only quadratic term with a significant coefficient is the good sleep and included group (black curve). Furthermore, a likelihood-ratio test of this model and the previous model does not reject the model linear in theta for the model that is quadratic in theta. Hence, we settled on the model from Figure A.13, as described in the main text.

Although we did some testing of the three-way interaction model specification, we need to improve the procedure so that we can scale that analysis to include all exclusion criteria and all domains. One procedure for doing this could proceed as follows: We could more finely discretize participants' PROMIS scores, by defining it as a factor variable with one level of the factor for every possible score they could achieve on the domain they valued, given the precision of the PROMIS-29 and Cognition 4-item short form. Then, by interacting this factor with the factor version of the six or seven health states valued in a given domain, as well as the binary indicator variable for a given exclusion criterion, we could write out the full factor version of the model in equation (6.3). This

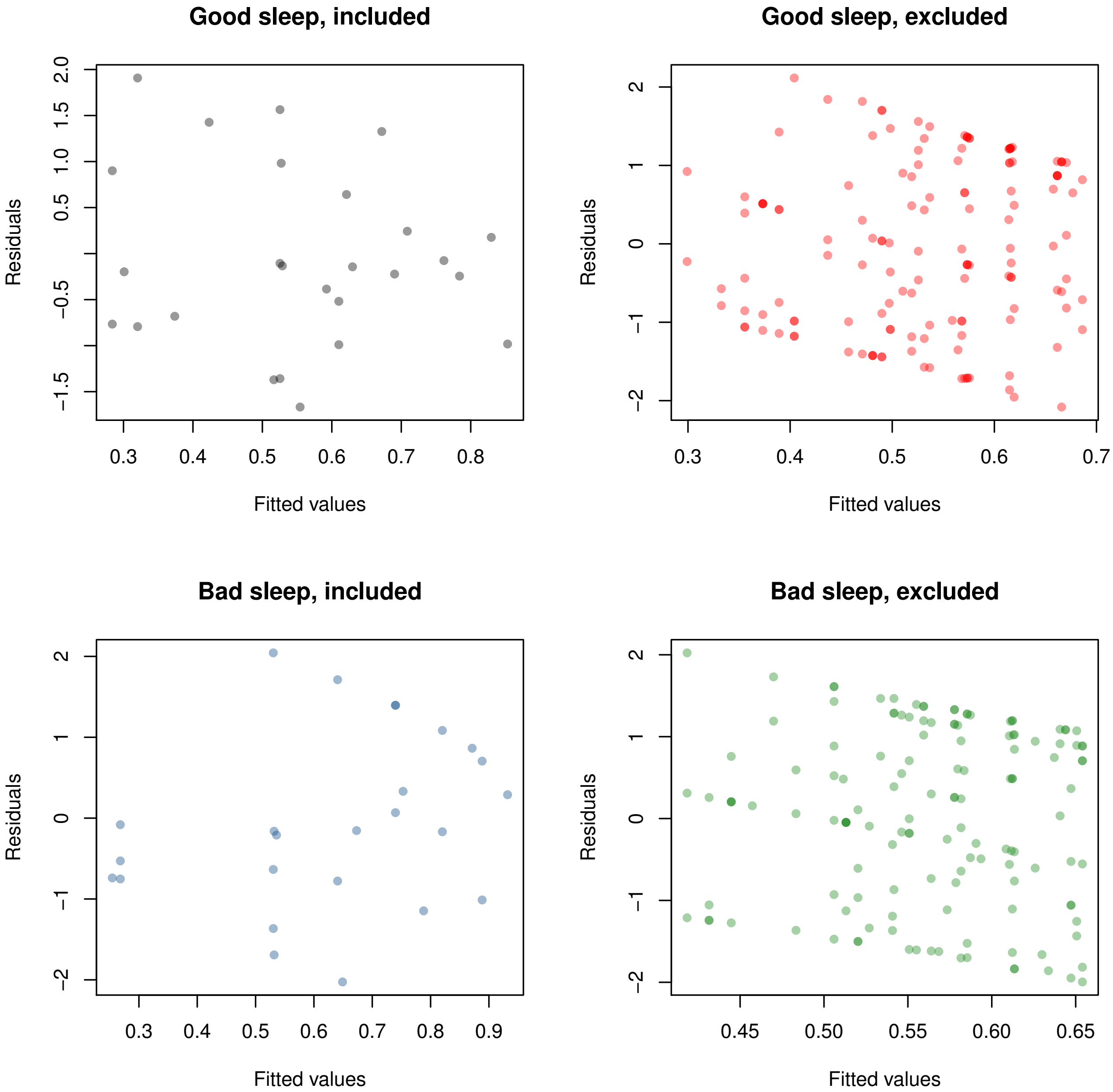


Figure A.24: *Residuals versus fitted values for the beta portion of a ZOIB model for sleep utilities as a function of theta, violates-SG, and discretized participant sleep quality.* Residuals from a three-way interaction model (equation (6.3)) of the mean utility for sleep as a function of theta, the *violates-SG* exclusion criterion, and the participants' PROMIS scores on the sleep domain (continuous), excluding all utilities equal to 0 and 1. The residuals are separated into four groups, based on exclusion/inclusion by *violates-SG*, and participant sleep quality.

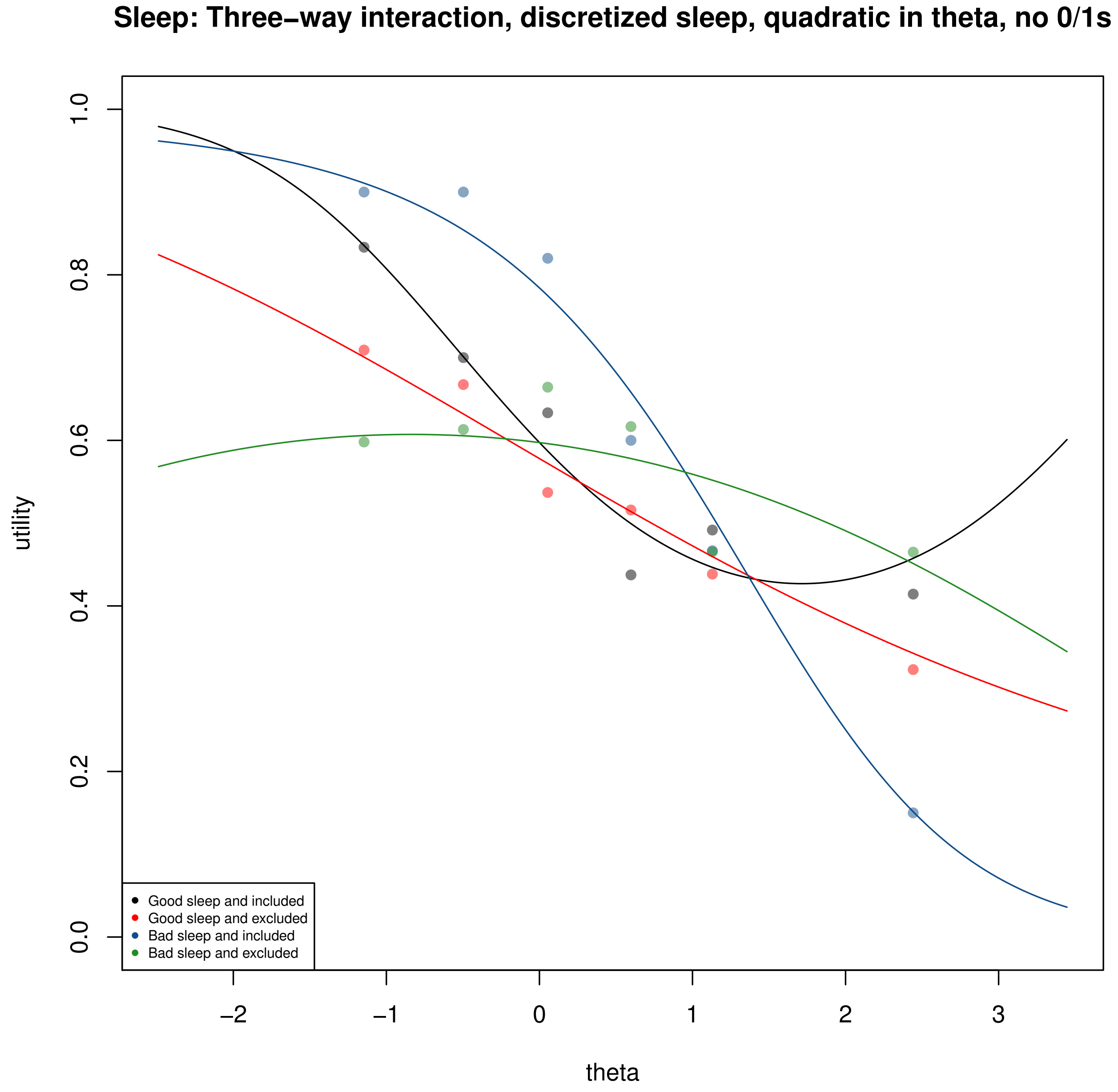


Figure A.25: *Utility curves and sample means for a version of the model for sleep utilities as a quadratic function of theta, violates-SG, and discretized participant sleep quality, with no 0s or 1s in the data.* Sample mean utilities and the conditional mean estimated for four groups: those who sleep well and are included by *violates-SG* (black), those who sleep well and are excluded by *violates-SG* (red), those who sleep poorly and are included by *violates-SG* (blue), and those who sleep poorly and are excluded by *violates-SG* (green). The conditional means come from the beta regression model in equation (6.3) expanded to include quadratic terms for the theta regressor, and where PROMIS scores on sleep are discretized as in the main text, and utilities of 0 and 1 are removed (i.e., it represents the beta portion of a ZOIB model).

would provide a model at the extreme of non-linearity in participant PROMIS scores and theta. We could then implement a procedure similar to the non-parametric testing of model miss-specification (Shalizi, n.d.), where we compare the performance of a proposed continuous model with that of the full factor model via simulation, where the full factor model stands in for the non-parametric model in the standard case (e.g., a kernel regression). This would allow a systematic and faster search for the best parametric form than the one implemented above for the sleep example.

One challenge of implementing this strategy is that the number of factor combinations in the full factor model is large (e.g., 204 for sleep), with many unobserved in the data. Thus, the model matrix is rank deficient. A method currently does not exist for the **betareg** package for determining a linearly independent set of covariates, so one will need to be implemented (e.g., by finding the eigendecomposition of the Gram matrix of the model matrix, using its eigenvectors with positive eigenvalues as the new set of covariates, and then transforming the model output into a recognizable form). Another challenge is ensuring that the method produces a p-value that behaves like the original method does, in the non-beta regression case.

Once implemented, this should allow us to more easily repeat the analysis completed for sleep and *violates-SG* with the other criteria and other domains. If the best-fitting model is non-linear in the theta or in the PROMIS scores, it could be challenging to interpret the results of the three-way interaction. For the moment, we believe our analysis of the sleep and social examples in the main text is illustrative, but further work is needed to determine whether it generalizes to the other domain-criterion pairs.

References

Arrow, K. J. (1951). *Social Choice and Individual Values*. New York, NY: John Wiley & Sons. 5, 6, 15

Asaria, M., Griffin, S., & Cookson, R. (2015). Distributional Cost-Effectiveness Analysis: A Tutorial. *Medical Decision Making*, *36*(1), 8–19. doi: 10.1177/0272989X15583266 5, 19

Badia, X., Fernandez, E., & Segura, A. (1995). Influence of socio-demographic and health status variables on evaluation of health states in a Spanish population. *The European Journal of Public Health*, *5*(2), 87–93. Retrieved from https://academic.oup.com/eurpub/article-lookup/doi/10.1093/eurpub/5.2.87 doi: 10.1093/eurpub/5.2.87 127

Badia, X., Monserrat, S., Roset, M., & Herdman, M. (1999). Feasibility, validity and test-retest reliability of scaling methods for health states: The visual analogue scale and the time trade-off. *Quality of Life Research*, *8*, 303–310. 127

Baird, J. C., & Noma, E. J. (1978). *Fundamentals of Scaling and Psychophysics*. John Wiley & Sons. 45, 51, 55

Basu, A., & Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, *27*, 112–127. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17409362 doi: 10.1177/0272989X06297393 19

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 54

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, *11*(5), 447–456. doi: 10.1002/hec.688 127

Bleichrodt, H., & Johannesson, M. (1997). Standard gamble, time trade-off and rating scale: Experimental results on the ranking properties of QALYs. *Journal of Health Economics*, *16*(2), 155–175. doi: 10.1016/S0167-6296(96)00509-7 127

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. Springer. 45, 51, 53, 55, 61, 137

Borg, I., Groenen, P. J. F., & Mair, P. (2012). *Applied multidimensional scaling*. Springer. 51, 140

Botteman, M. F. (2009). Health economics of insomnia therapy: Implications for policy. *Sleep Medicine*, *10*(SUPPL. 1), S22–S25. Retrieved from http://dx.doi.org/10.1016/j.sleep.2009.07.001 doi: 10.1016/j.sleep.2009.07.001 117

Botteman, M. F., Ozminkowski, R. J., Wang, S., Pashos, C. L., Schaefer, K., & Foley, D. J. (2007). Cost Effectiveness of Long-Term Treatment with Eszopiclone for Primary Insomnia in Adults: A Decision Analytical Model. *CNS Drugs*, *21*(4), 319–334. 117

Brazier, J., Roberts, J., & Deverill, M. (2002). The Estimation of a Preference-Based Measure of Health From the SF-12. *Journal of Health Economics*, *21*(9), 271–292. 21

CDC. (2016). *Chronic Disease Overview.* Retrieved from https://www.cdc.gov/chronicdisease/overview/ 134

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*(SUPPL. 1), 133–141. doi: 10.1007/s11136-007-9204-6 22

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., . . . Hays, R. D. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. Retrieved from http://dx.doi.org/10.1016/j.jclinepi.2010.04.011 doi: 10.1016/j.jclinepi.2010.04.011 22

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. *Medical Care*, *45*(5), 3–11. 22

Chan, K. K. W., Xie, F., Willan, A. R., & Pullenayegum, E. (2016). Underestimation of Variance of Predicted Health Utilities Derived from Multiattribute Utility Instruments: The Use of Multiple Imputation as a Potential Solution. *Medical Decision Making*, 0272989X16650181.

Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27216582 doi: 10.1177/0272989X16650181 42

Charness, G., & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. 17

Collins, F. S., & Riley, W. T. (2016). NIH's transformative opportunities for the behavioral and social sciences. *Science Translational Medicine*, *8*(366ed14). 22

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical Psychology: An Elementary Introduction*. Prentice-Hall. 22

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*(2), 1–24. doi: 10.18637/jss.v034.i02 45, 70, 73

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. 18

Culyer, A. J. (2006). The bogus conflict between efficiency and vertical equity. *Health Economics*, *15*, 1155–1158. doi: 10.1002/hec.1158 18, 19

D'Aspremont, C., & Gevers, L. (2002). Social welfare functionals and interpersonal comparability. In K. J. Arrow, A. K. Sen, & K. Suzumura (Eds.), *Handbook of social choice and welfare* (Vol. 1, pp. 461–541). 9

Davis, A. (2017a). *Applied Data Analysis* (Tech. Rep.). Carnegie Mellon University. 102

Davis, A. (2017b). *Why are my beta regression results biased? - Cross Validated.* Retrieved 2017-07-11, from https://stats.stackexchange.com/questions/290519/why-are-my-beta-regression-results-biased/ 152, 156

Davis, A., & Fischhoff, B. (2014). Communicating uncertain experimental evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 261–74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23895447 doi: 10.1037/a0033778 122

de Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1–30. 55

de Leeuw, J., & Meulman, J. (1986). A special Jackknife for Multidimensional Scaling. *Journal of Classification*, *3*(1), 97–112. doi: 10.1007/BF01896814 54

Deschamps, R., & Gevers, L. (1978). Leximin and utilitarian rules: A joint characterization. *Journal* *of Economic Theory*, *17*(2), 143–163. doi: 10.1016/0022-0531(78)90068-6 119, 120

Devlin, N. J., Hansen, P., Kind, P., & Williams, A. (2003). Logical inconsistencies in survey respondents' health state valuations - A methodological challenge for estimating social tariffs. *Health Economics*, *12*(7), 529–544. doi: 10.1002/hec.741 44, 50, 127

Devlin, N. J., Hansen, P., & Selai, C. (2004). Understanding health state valuations: A qualitative analysis of respondents' comments. *Quality of Life Research*, *13*(7), 1265–1277. doi: 10.1023/B:QURE.0000037495.00959.9b 44, 45

Dewitt, B., Davis, A., Fischhoff, B., & Hanmer, J. (2017). An Approach to Reconciling Competing Ethical Principles in Aggregating Heterogeneous Health Preferences. *Medical Decision Making*, 0272989X1769699. Retrieved from `http://journals.sagepub.com/doi/10.1177/0272989X17696999` doi: 10.1177/0272989X17696999 2, 4, 22, 41, 44, 126

Dewitt, B., Fischhoff, B., Davis, A., & Broomell, S. B. (2015). Environmental risk perception from visual cues: the psychophysics of tornado risk perception. *Environmental Research Letters*, *10*(12), 124009. Retrieved from `http://stacks.iop.org/1748-9326/10/i=12/a=124009?key=crossref.81ace82405fc4426a8d997503a5899a3` doi: 10.1088/1748-9326/10/12/124009 51, 55, 61

Dolan, P., Shaw, R., Tsuchiya, A., & Williams, A. (2005). QALY maximisation and people's preferences: a methodological review of the literature. *Health Economics*, *14*(2), 197–208. Retrieved from `http://doi.wiley.com/10.1002/hec.924` doi: 10.1002/hec.924 5, 16, 19, 117

Dolders, M. G. T., Zeegers, M. P. A., Groot, W., & Ament, A. (2006). A meta-analysis demonstrates no significant differences between patient and population preferences. *Journal of Clinical Epidemiology*, *59*(7), 653–664. doi: 10.1016/j.jclinepi.2005.07.020 123

Edwards, W. (1954). The Theory of Decision Making. *Psychological Bulletin*, *51*(4), 380–417. 15, 45

Ellsberg, D. (1954). Classic and Current Notions of "Measurable Utility". *The Economic Journal*, *64*(255), 528–556. 71

Embretson, S., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Inc. 22

Engel, L., Bansback, N., Bryan, S., Doyle-Waters, M. M., & Whitehurst, D. G. T. (2016). Exclusion Criteria in National Health State Valuation Studies: A Systematic Review. *Medical Decision Making*, *36*(7), 798–810. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26209475 doi: 10.1177/0272989X15595365 35, 41, 44, 45, 47, 49, 50, 53, 63, 122, 125, 126

EuroQol Group. (1990). EuroQol – A new facility for the measurement of health related quality of life. *Health Policy*(16), 199–208. doi: 10.1016/0168-8510(90)90421-9 21

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17641137 doi: 10.1177/0272989X07304449 47, 136

Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., . . . Boyle, M. (2002). Multiattribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. *Medical Care*, *40*(2), 113–128. doi: 10.1097/00005650-200202000-00006 5, 21, 24, 27, 28, 30, 32, 36, 44, 48, 50, 63, 126, 130, 133, 134

Feeny, D., Krahn, M., Prosser, L. A., & Salomon, J. A. (2016). Valuing Health Outcomes – Online Appendices. In P. J. Neumann, G. D. Sanders, L. B. Russell, J. E. Siegel, & T. G. Ganiats (Eds.), *Cost-effectiveness in health and medicine* (Second ed., pp. 167–199). New York: Oxford University Press. Retrieved from http://www.chepa.org/research-papers/valuing-health-outcomes/online-appendices 21

Field, M., & Gold, M. R. (Eds.). (1998). *Summarizing population health: Directions for the development and application of population metrics*. Washington, D. C.: National Academy Press. 16, 19

Fischhoff, B. (1984). Setting standards: A systematic approach to managing public health and safety risks. *Management Science*, *30*(7), 823–843. doi: 10.1287/mnsc.30.7.823 14

Fischhoff, B. (1991). Value Elicitation: Is There Anything in There? *American Psychologist*, *46*(8), 835–847. 16

Fischhoff, B. (2010). Judgment and decision making. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 724–735. Retrieved from http://doi.wiley.com/10.1002/wcs.65 doi: 10.1002/wcs.65 45

Fischhoff, B. (2015). The realities of risk-cost-benefit analysis. *Science*, *350*(6260), aaa6516–aaa6516.

Retrieved from http://www.sciencemag.org/cgi/doi/10.1126/science.aaa6516 doi: 10.1126/science.aaa6516 5, 19, 129

Fischhoff, B., & Kadvany, J. (2011). *Risk: A Very Short Introduction*. New York: Oxford University Press. 15, 19, 45

Fischhoff, B., & Morgan, G. (2009). The Science and Practice of Risk Ranking. *Horizons*, *10*(3), 40–47. 19

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences*, *9*, 127–152. doi: 10.1007/BF00143739 45

Fryback, D. G. (2010). *Measuring health-related quality of life.* Washington, DC. 41

Fryback, D. G., Palta, M., Cherepanov, D., Bolt, D., & Kim, J.-S. (2010). Comparison of 5 Health-Related Quality-of-Life Indexes Using Item Response Theory Analysis. *Medical Decision Making*, *30*(1), 5–15. doi: 10.1177/0272989X09347016 21, 22

Furlong, W., Feeny, D., Torrance, G. W., Goldsmith, C. H., DePauw, S., Zhu, Z., . . . Boyle, M. (1998). *Multiplicative Multi-Attribute Utility Function for the Health Utilities Index Mark 3 (HUI3) System: A Technical Report.* 5, 24, 27, 32, 126, 130

Gafni, A. (1994). The Standard Gamble Method: What Is Being Measured and How It Is Interpreted. *Health Services Research*, *29*(2). 28, 47

Gerhards, S. A. H., Evers, S. M. A. A., Sabel, P. W. M., & Huibers, M. J. H. (2011). Discrepancy in rating health-related quality of life of depression between patient and general population. *Quality of Life Research*, *20*(2), 273–279. doi: 10.1007/s11136-010-9746-x 123

Gershon, R. C., Rothrock, N., Hanrahan, R., Bass, M., & Cella, D. (2010). The use of PROMIS and Assessment Center to deliver Patient-Reported Outcome Measures in clinical research. *Journal of Applied Measurement*, *11*(3), 304–314. doi: 10.1037/a0013262.Open 22, 27, 47, 134

Gold, M. R., Siegel, J. E., Russell, L. B., & Weinstein, M. C. (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press. Retrieved from https://books.google.com/books?id=HWttErwnBHsC&pgis=1 4, 19

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*, 5–48. doi: 10.1007/BF01896809 51, 52, 137, 138, 151

Grootendorst, P., Feeny, D., & Furlong, W. (2000). Health Utilities Index Mark 3: Evidence of Construct Validity for Stroke and Arthritis in a Population Health Survey. *Medical Care*, *38*(3), 290–299. 32

Hanmer, J., & Dewitt, B. (2017). *PROMIS-Preference (PROPr) Score Construction – A Technical Report.* Retrieved from janelhanmer.pitt.edu/PROPr.html 22, 24, 25, 27, 36, 42, 44, 46, 47, 126, 136

Hanmer, J., Feeny, D., Fischhoff, B., Hays, R. D., Hess, R., Pilkonis, P. A., . . . Yu, L. (2015). The PROMIS of QALYs. *Health and Quality of Life Outcomes*, *13*, 122. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4530486&tool=pmcentrez&rendertype=abstract doi: 10.1186/s12955-015-0321-6 17, 44

Hays, R. D., Bjorner, J., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. *Quality of Life Research*, *18*. 134

Hays, R. D., Spritzer, K. L., Thompson, W. W., & Cella, D. (2015). U.S. General Population Estimate for “Excellent” to “Poor” Self-Rated Health Item. *J Gen Intern Med*, *30*(10), 1511–1516. doi: 10.1007/s11606-015-3290-x 134

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., . . . Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, *20*(10), 1727–1736. doi: 10.1007/s11136-011-9903-x 21, 134

Hirose, I. (2015). *Moral Aggregation*. Oxford University Press. 5

Hogg, K., Kimpton, M., Carrier, M., Coyle, D., Forgie, M., & Wells, P. (2013). Estimating quality of life in acute venous thrombosis. *JAMA Internal Medicine*, *173*(12), 1067–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23689427 doi: 10.1001/jamainternmed.2013.563 19

Kaplan, R. M., & Anderson, J. P. (1995). The Quality of Well-Being Scale: rationale for a single quality of life index. *Quality of Life: Assessement and Application*, 51–77. doi: 10.1007/978-94-011-2988-6_3 21

Kaplan, R. M., Tally, S., Hays, R. D., Feeny, D., Ganiats, T. G., Palta, M., & Fryback, D. G. (2011). Five preference-based indexes in cataract and heart failure patients were not equally responsive to change. *Journal of Clinical Epidemiology*, *64*(5), 497–506. Retrieved from http://dx.doi.org/10.1016/j.jclinepi.2010.04.010 doi: 10.1016/j.jclinepi.2010.04.010 21

Keeney, R. L., & Raiffa, H. (2003). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York, NY: John Wiley & Sons. 5, 17, 21, 22, 24, 27, 28, 33, 34, 36, 47, 130, 131

Kind, P., & Dolan, P. (1995). The Effect of Past and Present Illness Experience on the Valuations of Health States. *Medical Care*, *33*(4). 127

Koebberling, V. (2006). Strength of preference and cardinal utility. *Economic Theory*, *27*(2), 375–391. doi: 10.1007/s00199-005-0598-5 71

Konow, J. (2003). Which Is the Fairest One of All? A Positive Analysis of Justice Theories. *Journal of Economic Literature*, *41*(4), 1188–1239. doi: http://www.jstor.org/stable/3217459 17

Krabbe, P. F., Tromp, N., Ruers, T. J., & van Riel, P. L. (2011). Are patients' judgments of health status really different from the general population? *Health and Quality of Life Outcomes*, *9*(1), 31. Retrieved from http://www.hqlo.com/content/9/1/31 doi: 10.1186/1477-7525-9-31 123

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly*, *82*(4), 661–687. doi: 10.1111/j.0887-378X.2004.00327.x 19

Lamers, L. M., Stalmeier, P. F. M., Krabbe, P. F. M., & Busschbach, J. J. V. (2006). Inconsistences in TTO and VAS Values for EQ-5D Health States. *Medical Decision Making*, *26*(2), 173–181. 45

Lancsar, E., & Louviere, J. (2006). Deleting 'irrational' responses from discrete choice experiments: A case of investigating or imposing preferences? *Health Economics*, *15*(8), 797–811. doi: 10.1002/hec.1104 45

Lichtenstein, S., & Slovic, P. (Eds.). (2006). *Construction of preferences*. New York: Cambridge University Press. 16

Lilford, R., Girling, A., Braunholtz, D., Gillett, W., Gordon, J., Brown, C. A., & Stevens, A. (2007). Cost-Utility Analysis When Not Everyone Wants the Treatment: Modeling Split-Choice Bias. *Medical Decision Making*, *27*(1), 21–26. Retrieved from http://mdm.sagepub.com/cgi/doi/10.1177/0272989X06297099 doi: 10.1177/0272989X06297099 19

Liu, F., & Kong, Y. (2015). zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. *The R Journal*, *7*(2), 34–51. Retrieved from https://journal.r-project.org/archive/2015/RJ-2015-019/index.html 71, 93, 99, 102, 151, 152

Luce, R. D., & Suppes, P. (1965). Preferences, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 249–410). New York: John Wiley & Sons. 22

Martinson, B. C., Anderson, M. S., & de Vries, R. (2005). Scientists behaving badly. *Nature*, *435*(7043), 737–738. Retrieved from http://www.nature.com/doifinder/10.1038/435737a doi: 10.1038/435737a 122

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Structure*, *405*(2), 442–451. doi: 10.1016/0005-2795(75)90109-9 52

McCabe, C., Claxton, K., & Culyer, A. J. (2008). The NICE cost-effectiveness threshold: What it is and what that means. *PharmacoEconomics*, *26*(9), 733–744. doi: 10.2165/00019053-200826090-00004 117

McNamee, P., & Seymour, J. (2005). Comparing generic preference-based health-related quality-of-life measures: advancing the research agenda. *Expert Review of Pharmacoeconomics & Outcomes Research*, *5*(5), 567–582. 21

McNaughton, C. D., Cavanaugh, K. L., Kripalani, S., Rothman, R. L., & Wallston, K. A. (2015). Validation of a Short, 3-Item Version of the Subjective Numeracy Scale. *Medical Decision Making*, *35*(8), 932–936. Retrieved from http://mdm.sagepub.com/cgi/doi/10.1177/0272989X15581800 doi: 10.1177/0272989X15581800 35, 47, 136

Menzel, P., Dolan, P., Richardson, J., & Olsen, J. A. (2002). The role of adaptation to disability and disease in health state valuation: A preliminary normative analysis. *Social Science and Medicine*, *55*(12), 2149–2158. doi: 10.1016/S0277-9536(01)00358-6 105, 118, 123

Morgan, M. G., Fischhoff, B., Bostrom, A., & Atman, C. J. (2002). *Risk Communication: A Mental Models Approach*. Cambridge University Press. 17

Mulhern, B., Rowen, D., Snape, D., Jacoby, A., Marson, T., Hughes, D., . . . Brazier, J. (2014).

Valuations of epilepsy-specific health states: a comparison of patients with epilepsy and the general population. *Epilepsy & Behavior*, *36C*, 12–17. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24836527 doi: 10.1016/j.yebeh.2014.04.011 73, 123

Neumann, P. J., Goldie, S. J., & Weinstein, M. C. (2000). Preference-Based Measures in Economic Evaluation in Health Care. *Annu. Rev. Public Health*, *21*, 587–611. 44, 73, 106

Neumann, P. J., Saunders, G. D., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (Eds.). (2016). *Cost-Effectiveness in Health and Medicine* (Second ed.). New York: Cambridge University Press. 4, 117, 129, 132

Nord, E. (2014). Cost-Value Analysis of Health Interventions: Introduction and Update on Methods and Preference Data. *PharmacoEconomics*, *33*(2), 89–95. doi: 10.1007/s40273-014-0212-4 104, 117

Nord, E., Pinto, J. L., Richardson, J., Menzel, P., & Ubel, P. (1999). Incorporating societal concerns for fairness in numerical valuations of health programmes. *Health Economics*, *8*(1), 25–39. 5, 19

Owens, D. K., & Shekelle, P. G. (2013). Quality of Life, Utilities, Quality-Adjusted Life-years, and Health Care Decision Making. *JAMA Internal Medicine*, *173*(12), 1073–1074. 19

Paolino, P. (2001). Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables. *Political Analysis*, *9*(4), 325–346. Retrieved from http://pan.oxfordjournals.org/cgi/doi/10.1093/oxfordjournals.pan.a004873 70

Paulden, M., O'Mahony, J. F., Culyer, A. J., & McCabe, C. (2014). Some Inconsistencies in NICE's Consideration of Social Values. *PharmacoEconomics*, *32*(11), 1043–1053. doi: 10.1007/s40273-014-0204-4 18

Peeters, Y., & Stiggelbout, A. M. (2010). Health State Valuations of Patients and the General Public Analytically Compared: A Meta-Analytical Comparison of Patient and Population Health State Utilities. *Value in Health*, *13*(2), 306–309. 106, 123

PROMIS. (2015). *Applied Cognition – Abilities.* 27, 47, 134

Rawls, J. (1971). *A Theory of Justice*. Oxford: Oxford University Press. 5, 17

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric Evaluation and Calibration of Health-Related Quality of Life Item Banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, *45*(1). 22, 41

Roberts, K. (1980). Interpersonal Comparability and Social Choice Theory. *The Review of Economic Studies*, *47*(2), 421–439. 5, 6, 9, 10, 11, 12, 119, 120

Roberts, K. (2005). *Social Choice Theory and the Informational Basis Approach.* 7

Rowen, D., Mulhern, B., Banerjee, S., Tait, R., Watchurst, C., Smith, S. C., . . . Brazier, J. E. (2015). Comparison of General Population, Patient, and Carer Utility Values for Dementia Health States. *Medical Decision Making*, *35*(1), 68–80. Retrieved from http://mdm.sagepub.com/cgi/doi/10.1177/0272989X14557178 doi: 10.1177/0272989X14557178 123

Samsa, G., Edelman, D., Rothman, M. L., Williams, G. R., Lipscomb, J., & Matchar, D. (1999). Determining clinically important differences in health status measures: A general approach with illustration to the Health Utilities Index Mark II. *PharmacoEconomics*, *15*(2), 141–155. doi: 10.2165/00019053-199915020-00003 32

Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahn, M., . . . Ganiats, T. G. (2016). Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. *JAMA*, *316*(10), 1093–1103. Retrieved from http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.12195 doi: 10.1001/jama.2016.12195 73, 105, 121

Savage, L. J. (1972). *The Foundations of Statistics* (Second Rev ed.). New York: Dover Publications, Inc. 22

Schwalm, A., Feng, Y. S., Moock, J., & Kohlmann, T. (2015). Differences in EQ-5D-3L health state valuations among patients with musculoskeletal diseases, health care professionals and healthy volunteers. *European Journal of Health Economics*, *16*(8), 865–877. doi: 10.1007/s10198-014-0636-y 123

Sen, A. (1970). *Collective Choice and Social Welfare*. San Francisco: Holden-Day. 5, 6, 9

Sen, A. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, *6*(4), 317–344. doi: 10.2307/2264946 15

Shalizi, C. R. (n.d.). *Advanced Data Analysis from an Elementary Point of View*. Cambridge, England: Cambridge University Press. Retrieved from http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV 32, 70, 102, 172

Shepard, R. N. (1980). *Multidimensional scaling, tree-fitting, and clustering.* (Vol. 210) (No. 4468). doi: 10.1126/science.210.4468.390 45

Smithson, M., Budescu, D. V., Broomell, S. B., & Por, H.-H. (2012). Never say "not": Impact of negative wording in probability phrases on imprecise probability judgments. *International Journal of Approximate Reasoning*, *53*(8), 1262–1270. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0888613X12000953 doi: 10.1016/j.ijar.2012.06.019 70

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. doi: 10.1037/1082-989X.11.1.54 45, 69, 70, 71, 87, 156

Snedecor, S. J., Botteman, M. F., Bojke, C., Schaefer, K., Barry, N., & Pickard, A. S. (2009). Cost-Effectiveness of Eszopiclone for the Treatment of Adults with Primary Chronic Insomnia. *Sleep*, *32*(6), 817–824. Retrieved from http://www.journalsleep.org/ViewAbstract.aspx?pid=27481 doi: 10.1093/sleep/32.6.817 117, 124

Stern, P. C., & Fineberg, H. V. (Eds.). (1996). *Understanding Risk: Informing Decisions in a Democratic Society*. National Academy Press. 15

Torrance, G. W. (1986). Measurement of Health State Utillities for Economic Appraisal. *Journal of Health Economics*, *5*, 1–30. 8

Torrance, G. W., Boyle, M. H., & Horwood, S. P. (1982). Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research*, *30*(6), 1043–1069. 5, 22, 24, 71

Torrance, G. W., Feeny, D., & Furlong, W. (2001). Visual Analog Scales. *Medical Decision Making*, *21*(329). Retrieved from http://mdm.sagepub.com/content/21/4/329.short doi: 10.1177/0272989X0102100408 28, 47, 127

Torrance, G. W., Feeny, D., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). Multiattribute utility function for a comprehensive health status classification system: Health Utilities Index Mark 2. *Medical Care*, *34*(7), 702–722. 5, 8, 19, 21, 24, 28, 36, 130, 133, 134

Tosh, J., Brazier, J., Evans, P., & Longworth, L. (2012). A Review of Generic Preference-Based Measures of Health-Related Quality of Life in Visual Disorders. *Value in Health*, *15*(1), 118–127. Retrieved from http://dx.doi.org/10.1016/j.jval.2011.08.002 doi: 10.1016/j.jval.2011.08.002 21

Tourangeau, R., & Plewes, T. J. (Eds.). (2013). *Nonresponse in Social Science Surveys: A Research Agenda*. The National Academies Press. 41

Versteegh, M., & Brouwer, W. (2016). Patient and general public preferences for health states: A call to reconsider current guidelines. *Social Science & Medicine*, *165*, 66–74. Retrieved from http://dx.doi.org/10.1016/j.socscimed.2016.07.043 doi: 10.1016/j.socscimed.2016.07.043 105, 118, 119, 123, 124

Volk, R. J., Cantor, S. B., Cass, A. R., Spann, S. J., Weller, S. C., & Krahn, M. (2004). Preferences of husbands and wives for outcomes of prostate cancer screening and treatment. *J Gen Intern Med*, *19*(4), 339–348. doi: 10.1111/j.1525-1497.2004.30046.x[doi]\rJGI30046[pii] 19

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton University Press. 21, 22, 28, 47

Warrens, M. J. (2008). On association coefficients for 2 x 2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *73*(4), 777–789. doi: 10.1007/s11336-008-9070-3 51, 52

Wilson, I. B., & Cleary, P. D. (1995). Linking clinical variables with health-related quality of life: A conceptual model of patient outcomes. *JAMA*, *273*(1), 59–65. doi: 10.1001/jama.1995.0352025007503 4, 44

Wittenberg, E., Goldie, S. J., Fischhoff, B., & Graham, J. D. (2003). Rationing decisions and individual responsibility for illness: Are all lives equal? *Medical Decision Making*, *23*, 194–211. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12809318 doi: 10.1177/0272989X03253647 17

Wittenberg, E., & Prosser, L. A. (2011). Ordering errors, objections and invariance in utility survey responses: A framework for understanding who, why and what to do. *Applied Health*

Economics and Health Policy, *9*(4), 225–241. doi: 10.2165/11590480-000000000-00000 41, 49, 126

Yaari, M. E., & Bar-Hillel, M. (1984). On dividing justly. *Social Choice and Welfare*, *1*(1), 1–24. doi: 10.1007/BF00297056 17

Yu, L., Buysse, D. J., Germain, A., Moul, D. E., Stover, A., Dodds, N. E., . . . Pilkonis, P. A. (2011). Development of short forms from the PROMIS sleep disturbance and Sleep-Related Impairment item banks. *Behavioral sleep medicine*, *10*(1), 6–24. doi: 10.1080/15402002.2012.636266 117

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, *75*(6), 579–652. 52