Statistical Analysis of Likert Data on Attitudes

Kristin Nicole Javaras

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2004

Abstract

Researchers interested in measuring people's underlying attitudes towards an object (e.g., abortion) often collect Likert data by administering a survey. Likert data consist of surveyees' responses to statements about the object, where responses fall into ordered categories running from 'Strongly agree' to 'Strongly disagree' or into a 'Don't Know / Can't Choose' category. Two examples of Likert data are used for illustrative purposes. The first dataset was collected by the author from American and British graduate students at Oxford University and contains items measuring underlying abortion attitudes. The second dataset was taken from British and American responses to the 1995 National Identity Survey (NIS) and contains items measuring underlying national pride and immigration attitudes.

A model for Likert data and underlying attitudes is introduced. This model is more principled than existing models. It treats people's underlying attitudes as latent variables, and it specifies a relationship between underlying attitudes and responses that is consistent with attitudinal research. Further, the formal probability model for responses allows people's interpretation of the response categories to differ. The model is fitted by maximising an appropriate likelihood.

Variants of the model are used to analyse Likert data in three contexts; in each, the method using our model compares favourably to existing methods. First, the model is used to visualise the structure underlying the abortion attitude data. This method of visualization produces more sensible plots than analogous multivariate data visualization methods. Second, the model is used to select the statements whose responses (in the abortion attitude data) best reflect underlying abortion attitudes. Our method of statement selection more closely adheres to attitude researchers' stated aims than popular methods based on sample correlations. Third, the model is used to investigate how underlying national pride varies with nationality in the NIS data and also how underlying abortion attitude varies with gender, religious status, and nationality in the abortion attitude data. Unlike methods currently used by social scientists to model the relationship between attitudes and covariates, our method controls for the effects of differing response category interpretation. As a result, inferences about group differences in underlying attitudes are more robust to group differences in response category interpretation.

Personal acknowledgements

First and foremost, I would like to thank my superviser, Brian Ripley. Without his wisdom, guidance, and technical and moral support, this work simply would not exist. In addition, I am grateful for the assistance provided by other members of Oxford's Department of Statistics: Ruth Ripley, David Firth, and the department's computing and administrative staffs. I would also like to thank the Rhodes Trust for funding this work and, more generally, providing me with the opportunity to reside abroad in Britain, which piqued my interest in national differences in response category interpretation. Last, I am indebted to Alexander Rau and George and Barbara Javaras for their kindness and patience throughout the creation of this work.

Data acknowledgements

The national pride and immigration data utilised in this paper were documented and made available by the ZENTRALARCHIV FUER EMPIRISCHE SOZIALFORSCHUNG, KOELN. The data for the 'ISSP' were collected by independent institutions in Britain and America. The British data were collected as part of SCPR's (Social and Community Planning Research) 1995 British Social Attitudes survey under principal investigators Roger Jowell, Lindsay Brook, Alison Park, Katarina Thomson, and Caroline Bryson. The American data were collected as part of NORC's (National Opinion Research Center) 1996 General Social Survey under principal investigators James A. Davis and Tom W. Smith. Neither the original collectors nor the ZENTRALARCHIV bear any responsibility for the analyses or interpretation presented here.

Contents

Chapter 1

Introduction

Researchers in the social sciences and marketing are often interested in measuring people's attitudes towards objects such as abortion, their nation, or a consumer brand. Here, we use *attitude* to refer to an underlying construct hypothesised by psychologists, rather than to its observable manifestations which are referred to as attitudes in common parlance. More specifically, we define an attitude as a 'psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavour' (Eagly and Chaiken, 1998, p. 269). This definition implies that attitudes can be represented as points on an evaluative continuum that runs from extremely anti-object to extremely pro-object. Sometimes, researchers are interested in the locations of people along this continuum for their own sake. However, more often, researchers are interested in seeing how people's attitudes vary with certain background and behavioural covariates.

People's attitudes cannot be measured directly, and researchers disagree over whether and how they can be measured indirectly using people's emotions, thoughts, or behaviours, which are observable manifestations of attitudes (Bohner, 2001, p. 241). Those researchers who believe attitude measurement possible have proposed numerous techniques for doing so. Fishbein and Ajzen (1972 and 1975) provide a catalogue of attitude measurement techniques existing at that time, whereas Churchill (1999, Chapter 9) and Erwin (2001) provide excellent and recent overviews of attitude measurement techniques. One caveat applies to all attitude measurement techniques: Since an attitude is only a hypothetical construct, it can be measured at most at an interval level. In fact, there is considerable disagreement over whether even interval-level measurement is possible with some attitude measurement techniques.

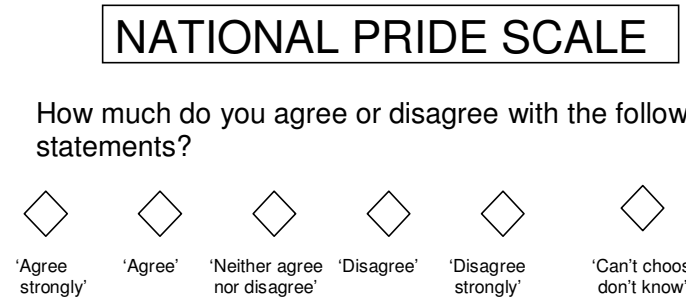


Figure 1.1: National Pride Scale. *These Likert items were used to measure general national pride in the National Identity Survey, which was administered in 1995 as part of the International Social Survey Programme.*

The attitude measurement technique on which we focus here is the Likert scale. Likert data have their origins in the scale proposed by Likert (1932), an example of which can be seen in Figure 1. Here, we use *scale* to refer to a relatively small number of questions (or items) selected to measure people's attitude towards a single object. In the scale proposed by Likert, each item involves choosing a response category— 'Agree strongly,' 'Agree,' 'Neither agree nor disagree,' 'Disagree,' or 'Disagree strongly'— to reflect one's level of agreement with a statement about the object. Today, some scales contain variations of the items proposed by Likert. These variations, all of which we term *Likert items*, may have more or fewer than five response categories spanning the agreement continuum, and might also have a 'Don't Know/Can't Choose' category. We refer to any scale containing Likert items as a Likert scale, regardless of the way in which those items are selected. Last, we use the term *Likert data* to refer to responses to Likert items, regardless of whether those items belong to a (final) scale or to an initial pool.

In this thesis, we propose new methods for the statistical analysis of Likert data.

Unlike existing methods, ours take a principled approach.

1.1 Requirements for methods of analysing Likert data

The method used to analyse Likert data will depend on the questions being asked. However, every method involves (albeit sometimes only implicitly) a *measurement model*, which is a model for the relationship between the Likert data and attitudes. Coombs (1964) stresses that a measurement model is "actually a theory about behaviour, admittedly on a miniature level" (p. 5). Further, he insists that "while building theory about more complex behavior it behooves us not to neglect the foundations on which the more complex theory rests." We share Coombs' view that the measurement model should reflect an appropriate theory. In the context of Likert data on attitudes, this means that the measurement model should reflect current theories of attitudes and attitude formation in several ways that we now enumerate.

A measurement model for Likert data should represent attitudes appropriately. The way we define an attitude implies that it can be represented as points along an (unobserved, underlying) evaluative continuum. Thus, in our measurement model, a person's attitude towards an object should be represented by a single parameter that takes one of a continuous set of values and affects the person's Likert responses.

The measurement model should also represent the statements appropriately. Studies investigating how attitude statements are processed (e.g., Judd and Kulik, 1980; Pratkanis, 1989) suggest that, in a person's mind, the statements fall along a continuum which is bipolar, at least when the attitude object is controversial (see Bohner, 2001, p. 244). Thus, in a measurement model, the statements about a particular object should be located along the relevant evaluative continuum. Further, the statements' locations should be consistent with their content.

Next, the measurement model should appropriately model the relationship between a person's attitude and his responses to the statements. It seems reasonable to model agreement as an (inverse) function of the distance between a person's attitude and a statement location that is assumed to be constant across all persons. We will refer to this type of relationship as an "unfolding model" because it has its origins in the unfolding process (a.k.a. ideal point process) proposed by Coombs (1964) as a description of how subjects arrive at preference orderings of stimuli. In an unfolding process, each person is represented as a point, each stimulus is represented as a point

that is perceived the same by all persons, and each person prefers (perhaps with some error) stimuli less distant from him.

Last, the measurement model should be appropriate for the way in which Likert data measure a person's responses to the statements. In Likert data, responses are expressed using categories that have three characteristics: they are mostly ordinal, possibly contain a 'Don't Know / Can't Choose' category, and may be interpreted differently by different people. As an example of this last characteristic, some people may use the 'Disagree strongly' category to reflect only anathema towards a statement, whereas others might use it to reflect anything less that whole-hearted agreement. We will refer to this phenomenon as *differing response category interpretation*.[1] A measurement model—in addition to being intended for responses that are ordinal and may contain a 'Don't Know / Can't Choose' category—should allow for differing response category interpretation. This is a particular concern when the data include persons from different cultures since there is considerable evidence that response category interpretations differ considerably between different cultures (Churchill, 1999, p. 450-451).

1.2 Overview of the thesis

Existing methods of analysing Likert data are not entirely appropriate for the task, in large part because the measurement models they employ do not meet the aforementioned requirements. In Chapter 2, we describe some of the measurement models used in these existing methods, and we explain how each fails to address some of the aforementioned requirements, or else does so in an unprincipled and *ad hoc* manner.

In Chapter 3, we introduce a new measurement model that addresses, in a principled manner, all (but one) of the above requirements for measurement models. In particular, our model allows for differing response category interpretation, which existing measurement models do not.

In Chapter 4, we describe how a variant of our measurement model can be employed to visualise the structure underlying Likert data. We use this method to visualise synthetic Likert data and Likert data on abortion attitudes, and we compare the results to those produced by an analogous multivariate data visualization method.

In Chapter 5, we describe how another variant of our model can be employed to select items (from an inital pool) for inclusion in a scale. We use this method to

[1] Rossi et al. (2001) refer to this phenomenon as "scale usage heterogeneity."

perform item analysis on a simulated dataset and on the abortion attitude data, and we compare the results to those produced by a popular method of item analysis.

In Chapter 6, we describe how a third variant of our model can be employed to investigate how attitudes vary with background and behavioural covariates, while controlling for the effects of differing response category interpretation. We conduct several simulation experiments to compare the performance of this method to that of a simple but popular method that does not allow for differing response category interpretation. We then use our method to investigate how national pride varies with (British or American) nationality, while adjusting for national differences in response category interpretation. Finally, we use our method to investigate how abortion attitudes vary with (British or American) nationality, gender, and religious status, while adjusting for national and gender differences in response category interpretation.

1.3 Description of datasets

The methods we propose are intended for Likert data with items that all share the same response categories. Here, we analyse two examples of data fitting this description.

1.3.1 Abortion attitude dataset

The abortion dataset is taken from a survey of 141 students. The surveyees were required to be (i) either British or American, (ii) raised primarily in Britain or the United States (respectively), and (iii) graduate students at the University of Oxford during the period April - May 2003. A random sampling mechanism was not used to select surveyees from the population fulfilling these three requirements; thus, care should be taken in generalizing the surveyees' abortion attitudes to all American and British graduate students at Oxford.

The surveyees responded to an interactive web-based survey containing a large number of questions related to abortion. These include five background questions, nine questions regarding behaviours related to abortion attitudes, and 50 Likert items on abortion attitudes. These fifty Likert items comprise an item pool from which a finalized Likert scale on abortion could be developed. The 50 statements[2] in this pool can be viewed in Table 1.1. For each surveyee, the 50 statements were presented in a

[2]Many of these statements are taken (often after slight modification) from the abortion statement pool tested on University of South Carolina undergraduates by Roberts et al. (2000).

random order, with the spelling of certain words (e.g.,'fetus' or 'foetus') determined by the person's response to a background question on nationality. Further, the six response categories (1 ='Agree strongly,' 2 ='Agree,' 3 ='Neither agree nor disagree,' 4 ='Disagree,' 5 ='Disagree strongly,' 99 ='Don't Know/Can't Choose') were arranged vertically on the computer screen in either an ascending or a descending order that varied randomly from statement to statement.

The abortion dataset contains 140 surveyees' responses to the survey questions. (We removed one surveyee with an abnormally high number of 'Don't Know/Can't Choose' responses.) Some information on the composition of the 140 person sample can be seen in Figure 1.2, which presents the marginal response category frequencies for four background questions. Note that the dataset is, by accident, almost perfectly balanced by gender and by nation. Summary information for the Likert items can be seen in Figure 1.3, which presents their marginal response category frequencies.

1.3.2 National identity dataset

This dataset is taken from the National Identity Survey (NIS), which was administered in many countries in 1995 as part of the International Social Survey Programme (ISSP). A description of the background and organizational aspects of the ISSP can be found at `http://www.gesis.org/en/data_service/issp/introduction.htm`. The ISSP website also contains the codebook for the 1995 NIS, which describes the sampling mechanism used in each country, lists the questions asked and possible responses, and presents the marginal response frequency (by country) for each question. Although the survey contained over twenty Likert items, we include in our NIS dataset only eleven Likert items measuring general national pride and immigration attitudes; these items can be seen in Table 1.2. Further, we include only British and American surveyees in our dataset, more specifically the 807 British and 998 American surveyees who had no missing or 'Don't Know / Can't Choose' responses to the national pride and immigration items. Figure 1.4 presents the marginal frequencies for these surveyees, by nation, for each of the national pride and immigration items.

Table 1.1: Fifty statements pertaining to abortion attitudes

S-1:	'No man or woman has the right to decide if a fetus should be aborted.'
S-2:	'Abortion is a threat to our society.'
S-3:	'Abortion is inhumane.'
S-4:	'Abortion is murder.'
S-5:	'There is no situation in which abortion is justified.'
S-6:	'Abortion is immoral.'
S-7:	'Abortion violates the unborn child's fundamental right to life.'
S-8:	'Abortion involves taking a life unjustly.'
S-9:	'Abortion could destroy the sanctity of motherhood.'
S-10:	'Abortion is the destruction of one life for the convenience of another.'
S-11:	'Abortion is a sin against God.'
S-12:	'Having the option to legally terminate a pregnancy encourages promiscuous behavior.'
S-13:	'Having an abortion is far worse than having an unwanted child.'
S-14:	'Having an abortion is a risk to a woman's physical health.'
S-15:	'Having an abortion is a risk to a woman's mental health.'
S-16:	'Even if one believes that there may be some exceptions, abortion is still generally wrong.'
S-17:	'Abortion is basically immoral, except when the woman's physical health is in danger.'
S-18:	'Abortion should be illegal except in extreme cases involving incest or rape.'
S-19:	'Abortions after the first three months should be illegal.'
S-20:	'Abortion is unacceptable, except when there is evidence that the fetus has severe defects.'
S-21:	'Partial birth abortions should be illegal.'
S-22:	'Abortion, if legal, should be strictly regulated.'
S-23:	'I believe that abortion is wrong in some or all situations, but I still think that it should be a matter of personal choice.'
S-24:	'Abortion should be permitted, but should never be used simply due to its convenience.'
S-25:	'Abortion, in general, should be legal, but should never be used as a conventional method of birth control.'
S-26:	'There are some cases where abortion is justified, but there are also some cases where it is not.'
S-27:	'It's impossible to make an airtight case either uniformly for or uniformly against abortion.'
S-28:	'My feelings about abortion are very mixed.'
S-29:	'I find myself agreeing with arguments both for and against abortion.'

continued on following page

Table 1.1: Fifty statements pertaining to abortion attitudes (continued)

continued from previous page

S-30: 'I personally have not resolved how I feel about abortion.'
S-31: 'If abortion were not legal, (illegal) abortions would still be performed.'
S-32: 'Sometimes I am in favor of a woman's right to abortion, but at other times I am not.'
S-33: 'I cannot wholeheartedly support either side of the abortion debate.'
S-34: 'Abortion should generally be a woman's prerogative, but it should not be permitted in every case.'
S-35: 'Regardless of my personal views about abortion, I do believe that others should have the legal right to choose for themselves.'
S-36: 'A woman should have control over what is happening to her own body by having the option to choose abortion.'
S-37: 'Only the woman who is pregnant can decide whether an abortion is warranted.'
S-38: 'Abortion is a matter of personal choice.'
S-39: 'Abortion should be legal under any circumstances.'
S-40: 'The government should never prohibit a woman from having an abortion.'
S-41: 'Restrictions should never be placed on a woman's right to an abortion.'
S-42: 'Outlawing abortion violates a woman's civil rights.'
S-43: 'Legal abortions pose less risk to the woman's mental and physical health than illegal abortions.'
S-44: 'I believe that abortion is generally wrong, but I think that it is necessary for it to be legal in today's society.'
S-45: 'If abortion became illegal, there would be negative consequences for society.'
S-46: 'It is better to have an abortion than an unwanted child.'
S-47: 'Abortion should be a socially acceptable method of birth control.'
S-48: 'Abortion is an acceptable means of dealing with an unwanted pregnancy.'
S-49: 'Abortion is a reasonable alternative if a woman feels that having a baby might ruin her life.'
S-50: 'Abortion should be available on demand.'

N.B.: American spelling is used in the above statements.

Table 1.2: Eleven statements pertaining to national pride and immigration attitudes

S-1:	'I would rather be a citizen of [my country] than of any other country in the world.'
S-2:	'There are some things about [my country] today that make me feel ashamed of [my country].'
S-3:	'The world would be a better place if people from other countries were more like the people in [my country].'
S-4:	'Generally [my country] is a better country that most other countries.'
S-5:	'People should support their own country even if the country is in the wrong.'
S-6:	'Immigrants increase crime rates.'
S-7:	'Immigrants are generally good for [own country's] economy.'
S-8:	'Immigrants take jobs away from people who were born in [own country].'
S-9:	'Immigrants make [own country] more open to new ideas and cultures.'
S-10:	'Refugees who have suffered political repression in their own country should be allowed to stay in [my own country].'
S-11:	'[Own country] should take stronger measures to exclude illegal immigrants.'

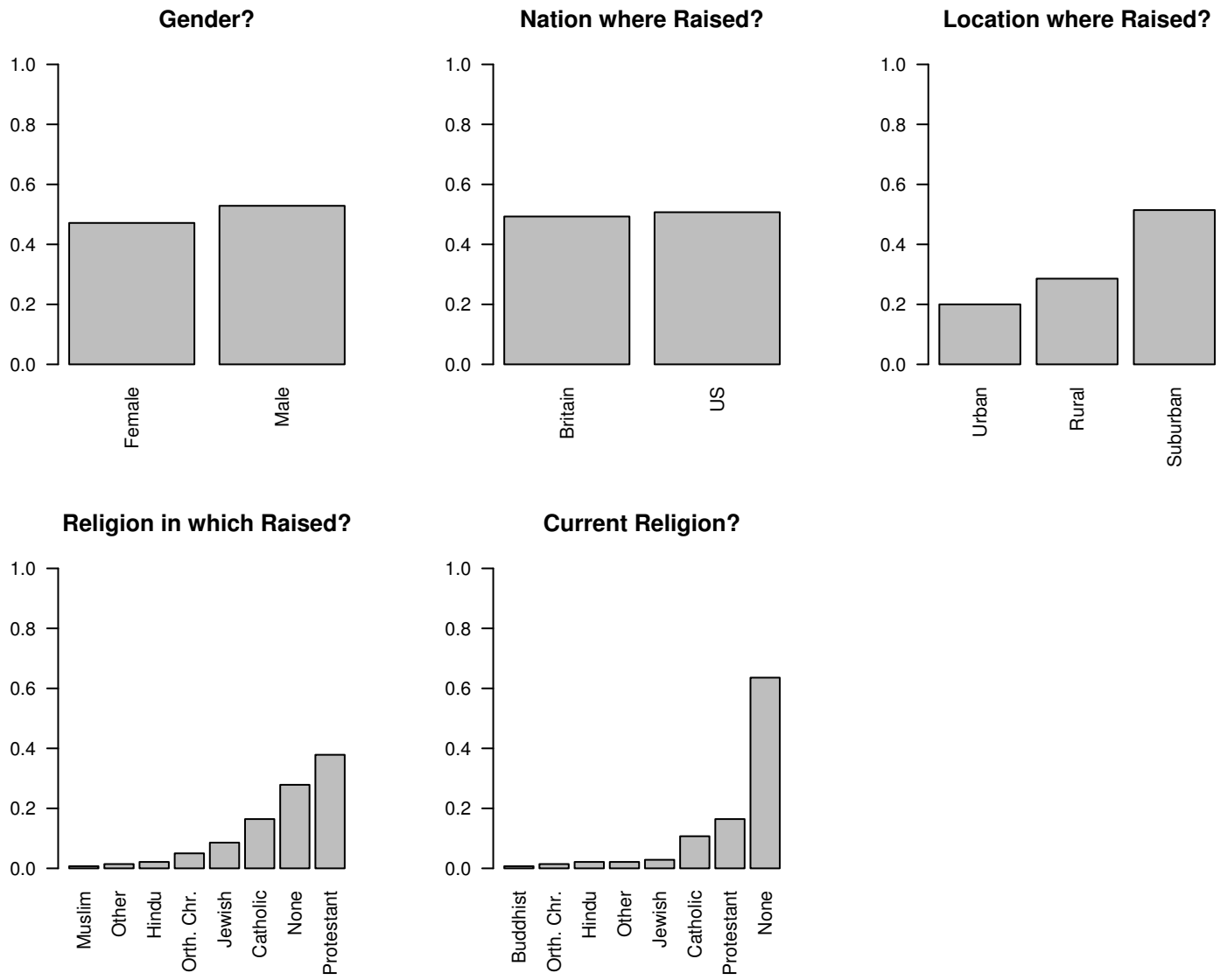


Figure 1.2: Background Information for Abortion Attitude Surveyees. *Each plot shows the marginal response category frequencies for a background question, calculated for* 140 *surveyees.*

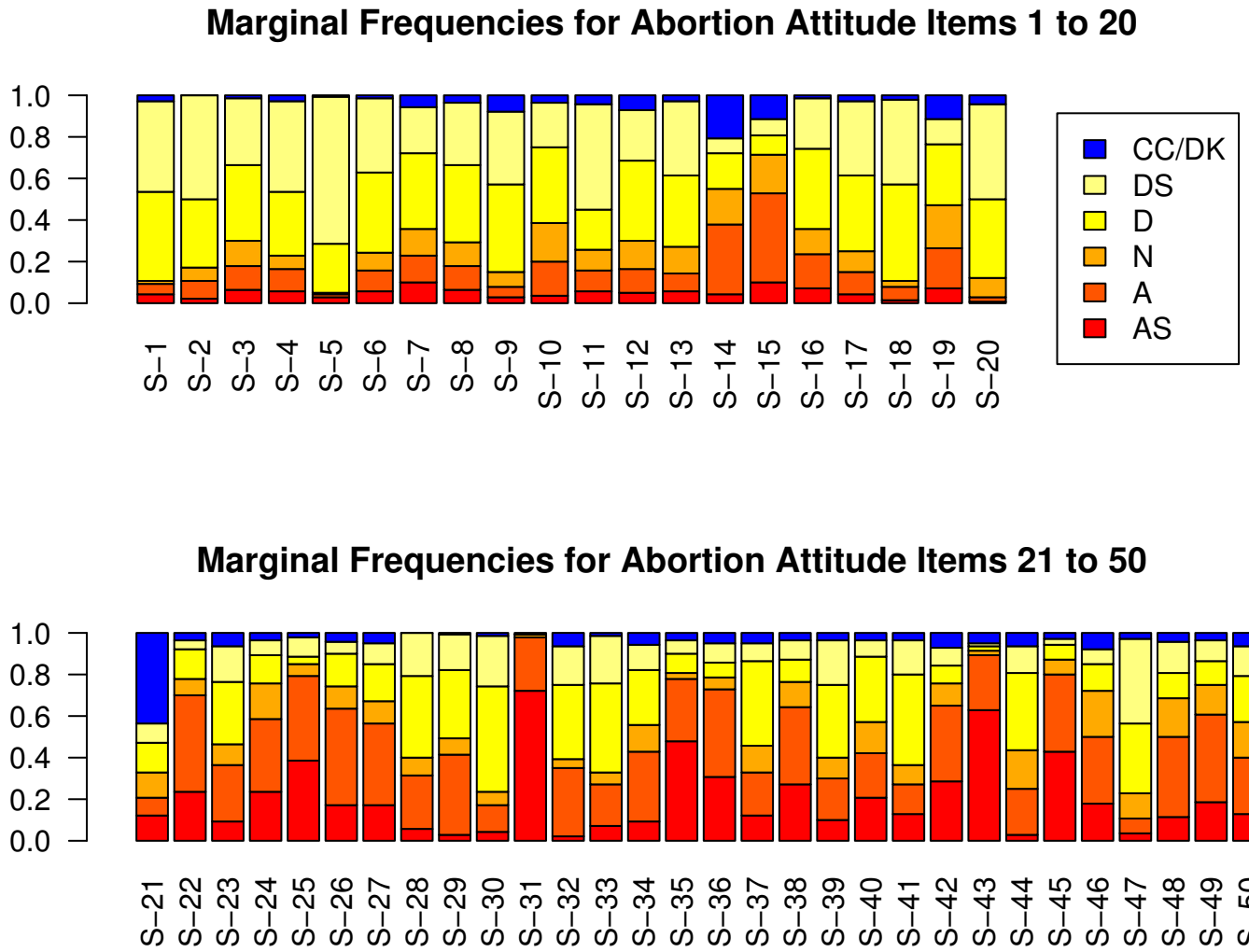


Figure 1.3: Responses for Abortion Attitude Items. *Each stacked bar shows the marginal response category frequencies for one of the abortion attitude items, calculated for* 140 *surveyees.*

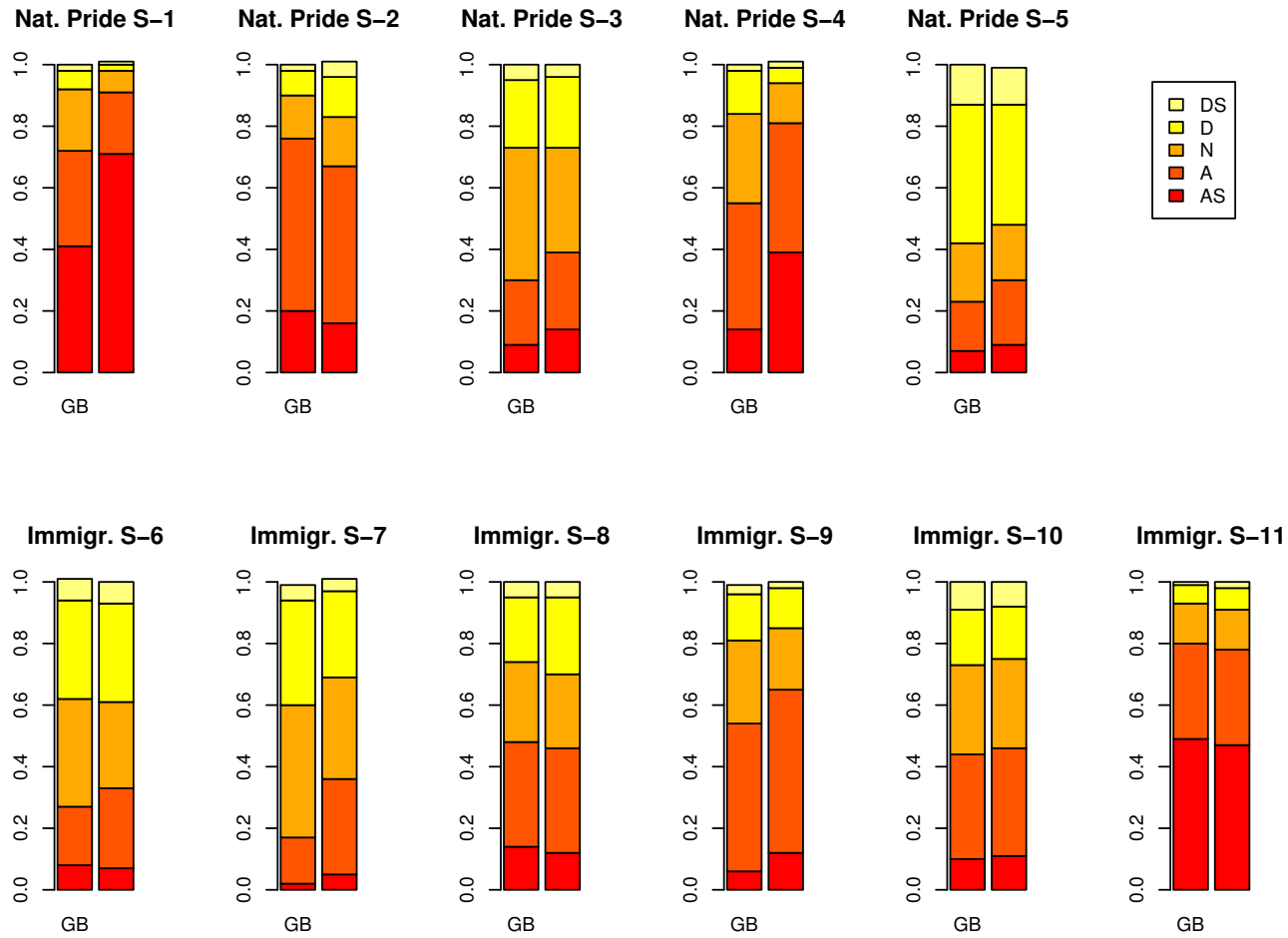


Figure 1.4: Responses for National Pride and Immigration Items. *Each plot contains two stacked bars showing the marginal response category frequencies, by nation, for one of the national pride or immigration items in the NIS dataset.*

Chapter 2

Existing measurement models for Likert data

Although the simple measurement model proposed by Likert remains popular, both multivariate data visualization type models and latent variable models have been employed as measurement models more recently. In this chapter, we overview several of the most frequently used models. Among these are Likert's model and a model developed for visualizing multivariate data; the remainder are latent variable models, which represent attitudes as underlying, continuous variables. In addition to describing each model, we note how it fails to meet some of our requirements, or else does so in an unprincipled manner.

We note that although Likert data are generated as ordinal direct response data, they are not treated as such by all of the following measurement models. *Direct response data* is generated by asking the respondent to indicate how much he endorses each of a number of stimuli. With Likert data, endorsement of (i.e., agreement with) the stimuli (i.e, statements) is measured at an ordinal level. Some of the following measurement models (e.g., latent trait models for ordinal variables) do treat the data as ordinal direct response data. However, some of the other measurement models (e.g., factor analysis, principal components analysis, and Likert's model) treat the data as quantitative direct response data. Last, one model (the unfolding model for rank data) treats the data as *rank order data*, which is generated by presenting the respondent with all stimuli at once and asking him to rank them in terms of descending or ascending endorsement.

Before describing these models, we introduce some notation. We begin with the Likert data, which consist of n persons' responses to J Likert items. We use $\mathbf{Y}$ to refer to the matrix containing the Likert data. Similarly, Y_i refers to the vector of person

i's responses to the J items, and Y_{ij}, which we term an *observed response*, refers to person i's response to item j. For the data we consider, Y_{ij} falls into one of K ordered agreement categories or into a 'Don't Know / Can't Choose' category. The items in $\mathbf{Y}$ can be partitioned into S mutually exclusive and exhaustive *item sets*, where a set contains all items intended to measure attitudes towards one particular object. I_s will refer to item set s, where $s = 1, ..., S$. Further, $S(j)$ will be used to indicate the index of the set to which item j belongs, and J^s will refer to the number of items in I_s. From now on, we will use θ_i^s to refer to person i's underlying attitude towards object s, and the vector Y_i^s will refer to person i's responses to the items in set s.

2.1 Likert's measurement model

The earliest and simplest measurement model for Likert data is the one proposed by Likert (1932). It requires that, first, each item in I_s be classified as either favourable or unfavourable towards object s. Then, the response categories must be quantified using consecutive integer scores,[1] with the scores running in opposite directions for favourable and unfavourable statements. We refer to the quantified data as $\mathbf{Y}^*$. The relationship between θ_i^s and person i's (quantified) responses to the items in I_s is then

$$\hat{\theta}_i^s = \sum_{j \in I_s} Y_{ij}^*, \tag{2.1}$$

where $\hat{\theta}_i^s$ is referred to as the *total score*.

The ease of implementing Likert's model makes it an appealing choice. Further, the model does meet some of our requirements for a measurement model. It represents attitudes appropriately, and, by reversing the direction of the scores for favourable and unfavourable statements, implicitly reflects an unfolding process.

However, Likert's model fails to meet our other requirements. First, the way that it represents statements is not entirely appropriate. Likert's model differentiates only between favourable and unfavourable statements and assumes that they are all similarly extreme. This is not appropriate for most sets of items, which typically contain statements of varying extremity. Further, some sets contain moderate statements that cannot be classified as unfavourable or favourable, making it impossible to use Likert's model. Second, Likert's model ignores the characteristics of Likert responses. It

[1] Most researchers would probably use the consecutive integers scores $1, \ldots, K$, but the choice of scores does not affect the ordering of people's attitude estimates.

assumes that they are measured at an interval (rather than ordinal) level, does not allow for 'Don't Know / Can't Choose' responses, and assumes that all people interpret the response categories identically. Some researchers attempt to address this final shortcoming by normalizing the quantified responses for a person before summing them; we refer to the resulting sum as the *adjusted total score.* However, when there is only a small number of items available, the mean and standard deviation estimates on which the normalization relies will not have the desired statistical properties. Further, this *ad hoc* approach to dealing with differences in response category interpretation assumes that the response data are continuous and from an elliptically symmetric distribution (Rossi et al., 2001). For Likert data, the first assumption is obviously untrue, and the second assumption is likely to be untrue.

Aside from failing to meet many of our requirements, Likert's model suffers from other limitations when used to perform certain types of analysis, due to the fact that it is not a formal probability model (see Chapter 6).

2.2 Unfolding model for rank data

This model belongs to the class of multivariate data visualization techniques. In the context of Likert data, the model seeks to locate the persons and statements in an underlying Euclidean space so that the *order* of the distances between them best matches the *order* of the Y_{ij}s within rows of $\mathbf{Y}$. Despite being intended for rank data, the model does meet almost all of our requirements for a measurement model. Unfortunately, as we discuss in Chapter 4, the way it is formulated leads to performance problems, especially when the number of items is small. In addition, the model is not a formal probability model, which means it is not appropriate for some types of analysis performed here (e.g., scale development and modelling the relationship between attitudes and covariates).

2.3 Using factor analysis and principal components analysis

Both factor analysis and principal components analysis (PCA) are frequently used to analyse Likert data. We focus here on the former, specifically on using the one-

dimensional normal linear factor model (NLFM)[2] as a measurement model for Likert data after they have been quantified:

$$Y_i^{s*} = \mathbf{\Lambda}\theta_i^s + \xi_i, \tag{2.2}$$

where Y_i^{s*} is person's i's (quantified) responses to the items in I_s; where $\mathbf{\Lambda}$ is a J by 1 matrix of parameters pertaining to the items; where θ_i^s has a standard normal distribution; and where ξ_i contains the J error terms for person i, is orthogonal to θ_i^s, and has a $N(\mathbf{0}, \mathbf{\Psi})$ distribution, with $\mathbf{\Psi}$ a diagonal matrix of specific variances.

Although this model does represent attitudes as a continuous, underlying parameter, it does not meet any of our other requirements (not surprisingly, since it was not formulated as an attitude measurement model). First, the one-dimensional NLFM does not reflect an unfolding process because agreement with a statement either increases or decreases with attitudes (depending on the sign of $\mathbf{\Lambda}_{j,1}$). Second, in practice, $\hat{\mathbf{\Lambda}}$ does not usually order the statements in a manner that seems consistent with their content.[3] Third and finally, the NLFM is not intended for responses that have the three characteristics of Likert responses. Although quantifying the ordinal Likert data makes it possible to use the NLFM, the quantification process is ad hoc and arbitrary since the spacing of categories for ordinal data is, by definition, unknown. Further, even after the data are quantified, they remain categorical and still unsuited for the NLFM, which is intended for continuous variables.

2.4 Using item response theory models

Within the past decade, researchers have developed item response theory approaches to attitude measurement with Likert data.[4] These approaches use latent trait models that reflect an ideal point or unfolding process (Coombs, 1964). Unfolding models can be distinguished from monotone models, which reflect a dominance process (Coombs, 1964). Latent trait models of either type represent items with a parameter that, for Likert data, can be thought of as the statement's location. However, the relationship between the latent variable, statement locations, and responses takes different forms

[2] See Bartholomew and Knott, 1999, Chapter 3, for an overview of the NLFM.

[3] Oftentimes, a plot of the two-dimensional NLFM or PCA solutions will locate the statements around a horseshoe in a manner consistent with their content. This empirical result accords with the theoretical work of Davison (1977) on PCA with data generated from metric unfolding models.

[4] Roberts (1995) describes parametric approaches for data with K agreement categories. Johnson (2001) describes parameteric and non-parametric approaches for data with two agreement categories.

in the two types of models. In an unfolding latent trait model for ordinal variables, the expected response function[5] is unimodal in the latent variable and peaks at the statement location (Luo, 2001). In a monotone latent trait model for ordinal variables, the expected response function is monotonic in the latent trait.

2.4.1 Unfolding latent trait models for ordinal variables

These models, which were developed for attitude measurement with Likert data, include Andrich's (1996) and Rost and Luo's (1997) generalised hyperbolic cosine model (GHCM), Roberts and Laughlin's (1996) graded unfolding model (GUM), and Roberts et al.'s (2000) generalised graded unfolding model (GGUM).[6] All three are formal probability models with one latent variable. Further, all three adopt the same approach—that of *unobserved response categories*—to create an unfolding structure. In this approach, each item's K observed, ordered response categories unfold into $2*(K-1)+1$ (or $2*K$) unobserved, ordered response categories according to a prescribed mapping. Figure 2.1 illustrates how this mapping works for one of the national pride items introduced in Chapter 1. The probabilities of the unobserved categories are modelled (as a function of the latent trait) using a monotone model for ordinal variables. The probability of each observed category is then obtained by summing the probabilities of the corresponding unobserved categories.

These models fulfill almost all of our requirements. For one, they are designed to reflect an unfolding process, as can be seen from their *item category response functions* (*ICRF*s), which model the probability of category responses as a function of the latent trait. The ICRFs for a five-category Likert item are presented in Figure 2.2. The curve for 'Disagree strongly' clearly reflects an unfolding process: The probability of strongly disagreeing with a statement increases as an individual is farther from the statement's location.

Unfortunately, however, these models do not allow for 'Don't Know / Can't Choose' responses, or differing response category interpretation. Moreover, the models' structure—specifically, the way in which they induce an unfolding structure—does not lend itself to incorporating differing response category interpretation.

[5]Here, the expected response function for a particular item is a weighted sum of the item's K category probability functions; the weights equal consecutive integer scores, with the highest score used for the 'Strongly agree' category.

[6]Luo (2001) developed a general framework for a class of probabilistic, unfolding, unidimensional latent trait models for ordered data. This class includes the GHCM and GUM as special cases.

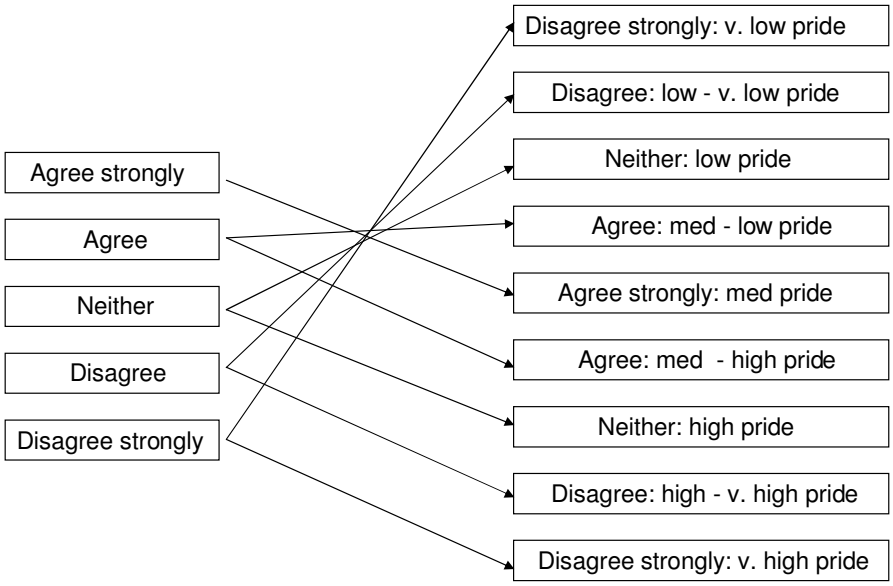


Figure 2.1: Unfolding Mapping of Observed Response Categories. *This mapping is illustrated for a Likert item from the national pride scale. Five ordered, observed response categories that span the agreement continuum unfold into nine ordered, unobserved response categories that span the national pride continuum.*

2.4.2 Monotone latent trait models for ordinal variables

Although these models were not developed for attitude measurement, they have been used to analyse Likert data. They can be divided into two classes depending on the approach they take to modelling the category probabilities. These classes are referred to as *divide-by-total models* and *difference models* in Thissen and Steinberg's (1986) classification.

2.4.2.1 Difference models

A popular difference model is what we will term the *Underlying Variable Model for Ordinal Variables* (UVMOV), which has logit and probit variants.[7] This model can be motivated from an ICRF perspective or from an underlying variable, factor analysis perspective.[8] In the underlying variable perspective, each observed ordinal response

[7]The ordinal-logit variant of the UVMOV with one latent trait also known as Samejima's (1969) *Graded Response Model*.

[8]These two perspectives result in different approaches to fitting the UMVOV. See Bartholomew et al. (2002, Chapter 8), J oreskog and Moustaki (2001), and for the relationship and differences between

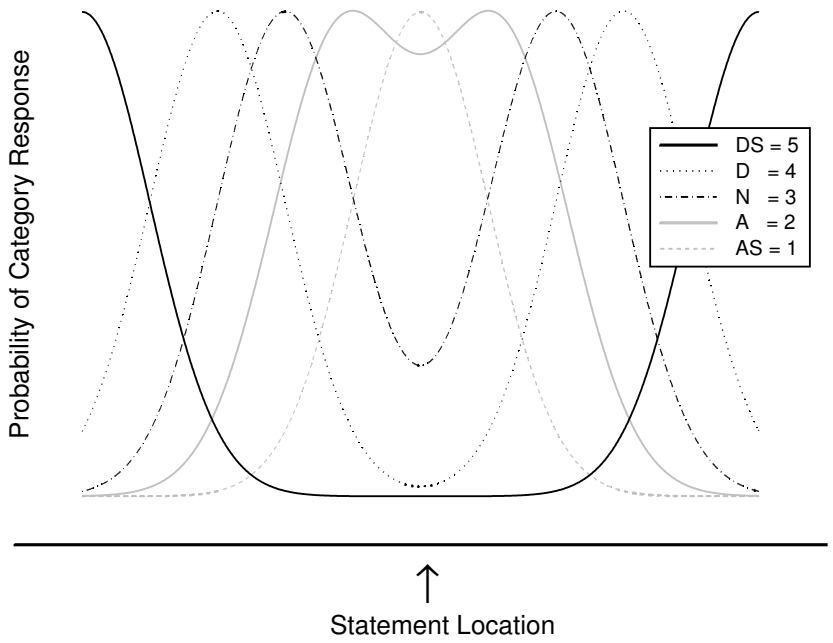


Figure 2.2: Unfolding ICRFs. *These ICRFs for a Likert item with $K = 5$ are generated by an unfolding latent trait models for ordinal variables. The bottom axis represents the evaluative continuum.*

($\mathbf{Y}_i^s$) is a coarsened version of an unobserved continuous response, which we refer to as $\mathbf{Y}_i^{s*}$. The $\mathbf{Y}_i^{s*}$s are then modelled in a manner similar to (2.2), except that the error terms can have a logistic distribution (ordinal-logit variant) instead of a normal distribution (ordinal-probit variant).

Although the UVMOV does represent attitudes appropriately and is intended for ordinal responses, it fails to meet some of our other requirements. First, the model cannot handle 'Don't Know / Can't Choose' responses, and it does not allow response category interpretation to differ.[9] Second, since the UVMOV is a monotone model, the relationship between the latent variable (i.e., attitude) and the responses does not reflect an unfolding process. Figure 2.3 shows how the UVMOV does model the relationship

the two perspectives.

[9]The model could be modified to incorporate differing response category interpretation, by allowing people to have different thresholds (for coarsening their unobserved continuous responses). This approach is used in the measurement model proposed in the next chapter.

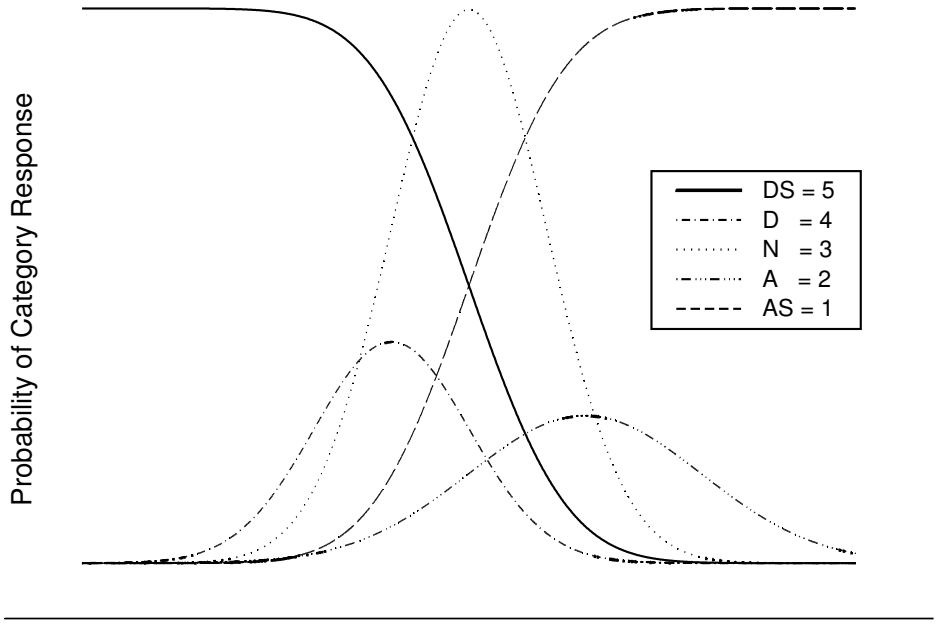


Figure 2.3: Monotone ICRFs. *These ICRFs for a Likert item with $K = 5$ are generated by a monotone latent trait models for ordinal variables. The bottom axis represents the evaluative continuum.*

between the latent variables and the category probabilities. Obviously, the relationship depicted in this plot would be appropriate only when the statement is very extreme compared to the population surveyed (e.g., if the statement were located off the right-hand side of the plot).

2.4.2.2 Divide-by-total models

A popular divide-by-total model for ordinal variables is the partial credit model (PCM) (Masters, 1982; Masters and Wright, 1984), alternatively known as the polytomous Rasch model. Many other well-known divide-by-total models are variations on the PCM. For instance, Andrich's (1978) rating scale model (RSM) is a restricted version

of the PCM, and Muraki's (1992) generalized partial credit model (GPCM) generalizes the PCM.

Following in this vein, we propose a new model that modifies the PCM for use with Likert items that can be classified (*a priori*) as pro-object or anti-object. In this model, the probability that a person with latent attitude θ_i^s selects category k of item j is

$$P\left(Y_{ij} = k \mid \theta_i^s\right) = \frac{\exp\left[k\theta_i^s + c_{jk}\right]}{\sum_{s=1}^{K} \exp\left[s\theta_i^s + c_{js}\right]} \tag{2.3}$$

if statement j is anti-object and

$$P\left(Y_{ij} = k \mid \theta_i^s\right) = \frac{\exp\left[(K+1-k)\theta_i^s + c_{jk}\right]}{\sum_{s=1}^{K} \exp\left[(K+1-s)\theta_i^s + c_{js}\right]} \tag{2.4}$$

if statement j is pro-object. In both equations, $k = 1$ corresponds to greatest agreement and $k = K$ corresponds to greatest disagreement. In addition, we assume that local independence holds for this model; thus, the probability of person i's response pattern is the product (across items) of the probabilities specified in (2.3) or (2.4).

Figure 2.4 shows how the category probabilities (for a pro-statement and for an anti-statement) vary with θ_i^s in this model. Note that the ordering of the categories along the evaluative continuum is reversed for the pro-object and anti-object statements. Further, note that these plots represent realistic scenarios only for pro-object and anti-object statements located to the right and left, respectively, of the population surveyed.

An interesting property of this model is that the total score from Likert's measurement model is sufficient for θ_i^s. (It is easy to show that conditioning on the total score removes θ_i^s from person i's contribution to the likelihood.) In fact, the new PCM-like model was specifically formulated to have this property. Its connection to Likert's measurement model, combined with the observations made in the previous paragraph, give us insight into Likert's measurement model, suggesting that it implicitly specifies very extreme locations for the statements.

Unfortunately, as was the case for the UVMOV, the new PCM-like model would not be appropriate for Likert data containing any non-extreme statements.

2.5 Conclusions

Clearly, existing methods of analysis are limited by their measurement models, which are not fully appropriate for Likert data.[10] In response to this dearth, we now introduce a measurement model that fulfills all of our requirements (save one) in a principled manner.

[10]There are also other reasons why these methods are not entirely appropriate for their task. These additional reasons are discussed in Chapters 4, 5, and 6.

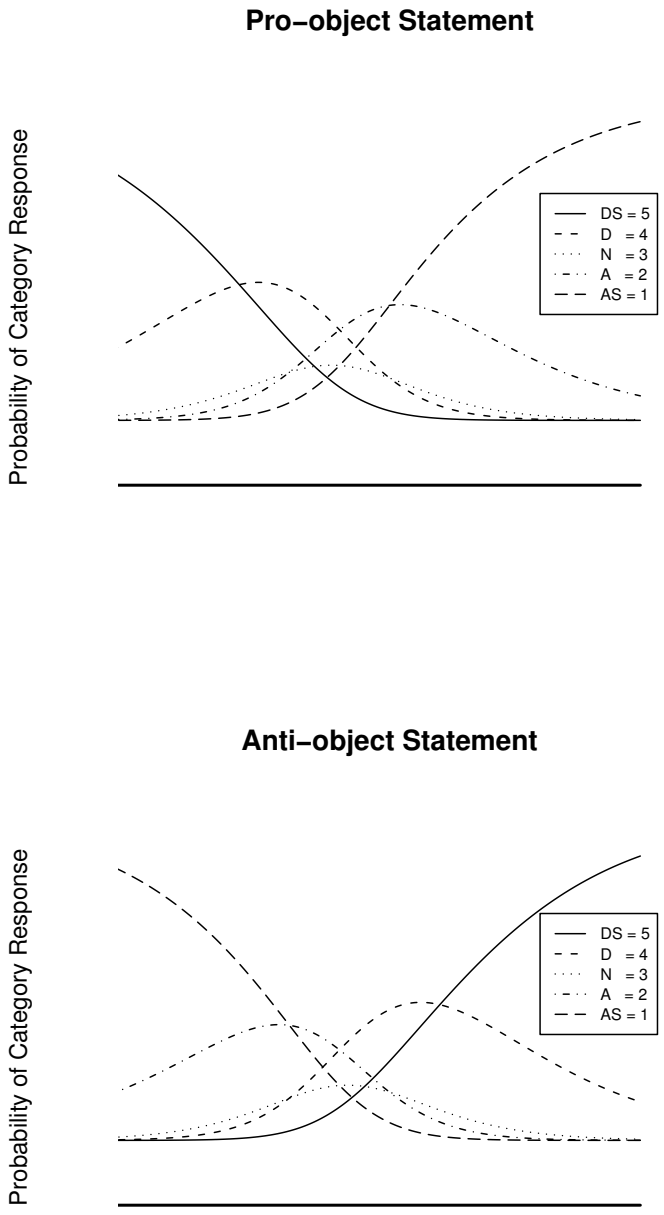


Figure 2.4: Likert ICRFs. *The ICRFs for a pro-object item and for an anti-object item, each with $K = 5$. The curves are generated by a PCM-like model formulated so that the total score is sufficient for θ_i^s. The bottom axis in the plot represents the evaluative continuum.*

Chapter 3

A new measurement model for Likert data

In this chapter, we introduce a new measurement model for Likert data, the *Unfolding Latent Trait Model for Ordered Direct Responses* (*ULTMODR*). The ULTMODR is a formal probability model for ordinal responses that, unlike other latent trait models, allows for differing response category interpretation. (Unfortunately, the model cannot incorporate 'Don't Know / Can't Choose' responses without modification). Further, the model relates the latent traits, which represent attitudes, to responses in a manner that reflects an unfolding process.

Note that the ULTMODR, though developed for Likert responses, is appropriate for other types of direct responses measured at an ordinal level.

3.1 Motivation behind the model

The ULTMODR combines a response structure with an unfolding latent structure. Separating the response structure and latent structure makes it possible to model differing response category interpretation in the response structure and more than one attitude (i.e., latent trait) in the latent structure. The two-part structure reflects the fact that people's responses might differ because of differing underlying attitudes, differing response category interpretations, or both. Obviously, we cannot hope to entirely separate the effects of these two phenomena on people's responses since they are both unobserved and, further, may not be independent. Of course, the model's ability to separate the two effects will depend on the data. As a trivial example, consider Likert data with only one item, in which case attitude and response category interpretation

would be hopelessly confounded. Obviously, the model will be better able to separate the two effects when there are more items. However, as we will see in Chapter 6, it is preferable that these items come from more than one set and that they be located in different places along the evaluative continuum.

Different variants of the ULTMODR's latent structure can be formulated depending on the particular data and questions being investigated. In the following discussion, we describe the simplest of these variants. It is intended for data containing only one set of items, and it assumes that only one latent trait (i.e., attitude) underlies them.

3.2 The ULTMODR

The formulation of the ULTMODR adheres to certain principles common to most latent variable models. First, the ULTMODR makes the assumption that the people are independent. In addition, it assumes local independence. This means that the items are (conditionally) independent given the parameters specific to each person; these parameters include each person's latent trait values (i.e., his attitudes) and possibly certain person-specific response category interpretation parameters. The assumption of local independence allows us to separately model responses to each item instead of the joint responses to all items.

The latent structure of the ULTMODR is a Euclidean space in which both persons and statements are located. The *person location* for person i will be denoted θ_i, and the *statement location* for item j will be denoted β_j. Note that the statement locations are objective in the sense that they do not depend on the persons. In the simplest variant of the ULTMODR, the Euclidean space is unidimensional and corresponds to the evaluative continuum for object s; in this variant, $\theta_i = \theta_i^s$ and $\beta_j = \beta_j^s$.

The latent structure is connected to the response structure by the assumption that Y_{ij} (probabilistically) increases with the Euclidean distance between the relevant person and statement locations. This assumption results in a model that reflects an unfolding process. We will refer to the Euclidean distance between θ_i and β_j as d_{ij}.

We now consider the model's response structure, which is a mapping from the d_{ij}s to the joint distribution of the persons' observed responses. This mapping occurs via the underlying-variable-coarsened-by-thresholds approach, which is equivalent to the cumulative probability approach to modelling ordinal variables (see Agresti, 2002,

Section 7.2.3). Here, the underlying response for person i on item j (denoted Y_{ij}^*) is the following function of d_{ij}:

$$Y_{ij}^* = \pm d_{ij} + \varepsilon_{ij}, \tag{3.1}$$

where ε_{ij} is assumed to have a logistic distribution with centre 0 and scale equal to $1/1.7$ or, alternatively, a normal distribution with mean 0 and variance 1; and where Y_{ij}^* is referred to as person i's *unobserved response* to item j and lies along an *agreement continuum* whose direction is determined by the $\pm$ sign.[1] The agreement continuum is partitioned (or coarsened) by thresholds into ordered categories. The ordered threshold set used for coarsening is

$$-\infty = c_0^i \leq c_1^i \leq \ldots \leq c_{K-1}^i \leq c_K^i = \infty. \tag{3.2}$$

As usual, a response of $Y_{ij} = k$ is observed if and only if $c_{k-1}^i \leq Y_{ij}^* \leq c_k^i$. Note that the superscript i in c_k^i indicates that the threshold set can be person-specific (although it might rather be group-specific or common to all people, as we discuss below). This personalisation of the thresholds between categories reflects differing response category interpretation, and it is possible because the thresholds are assumed to be the same for all items.

The ICRF for category k of item j in the simple variant is then

$$\begin{aligned} P\left(Y_{ij} = k\right) &= P\left(c_{k-1}^i \leq Y_{ij}^* \leq c_k^i\right) \quad (k = 1, \ldots, K) & (3.3) \\ &= P\left(\mp|\theta_i^s - \beta_j^s| + c_{k-1}^i \leq \varepsilon_{ij} \leq \mp|\theta_i^s - \beta_j^s| + c_k^i\right), & (3.4) \end{aligned}$$

where $S(j) = s$. These ICRFs have forms similar to those illustrated in Figure 2.2.

Finally, the likelihood for the data is

$$L = \prod_{i=1}^{n} L(Y_i^s) = \prod_{i=1}^{n} \prod_{j \in I_s} \prod_{k=1}^{K} I\left(Y_{ij} = k\right) P\left(Y_{ij} = k\right), \tag{3.5}$$

with the first and second multiplications allowed by the assumed properties of person independence and local independence, respectively.

[1] For example, if a positive sign is used, it is oriented in terms of decreasing agreement, which means that 'Agree strongly' should be treated as the lowest category.

3.2.1 Different cases of the response structure

The ULTMODR models differing response category interpretation by allowing the thresholds in (3.2) to vary. In this section, we describe different assumptions that can be made about the way in which this variation occurs.

In the following discussion, Z denotes a *group variable*, which is a demographic variable (e.g., nation) where response category interpretation varies more between this variable's groups than within them. Here, we assume that Z is categorical with its categories labeled $0, \ldots, G-1$, where G is the total number of groups and group 0 is the reference group.[2]

We enumerate the response structure cases in terms of increasingly stringent assumptions. In Case 1, the threshold set is allowed to differ for each person. In Case 2, the set is allowed to differ between Z groups but is the same for all people within a given Z group. Lastly, in Case 3, the threshold set is the same for all people. In Cases 1 and 2, the threshold set could be allowed to vary (across persons or groups) in an unrestricted manner or in a restricted manner that involves shifting and scaling the threshold set (for each person or group). In Case 1, we allow only restricted variation for reasons of computational convenience. However, in Case 2, we allow the set to vary across groups in either an unrestricted manner (Case 2a) or a restricted manner (Case 2b).

1. *Person-specific Response Structure*: Each person is allowed to interpret the response categories differently. In formal terms, this means that a person-specific threshold set, $\{c_1^i, \ldots, c_{K-1}^i\}$, is used in the model's response structure. For person i,

$$c_k^i = (\sigma^i)^{-1} c_k + \tau^i \text{ for } k = 1, \ldots, K$$

where $\{c_1, \ldots, c_{K-1}\}$ is the common threshold set; and where τ^i and σ^i are referred to as *person interpretation parameters* and are used to shift and scale the common threshold set for each person. In particular, τ^i describes the centre of where person i's d_{ij}s map onto the agreement continuum. This parameter can be thought of as a quantification of *acquiescence*, a tendency to use more agreeable response categories regardless of the questions being asked. Analogously,

[2] If we suspect that more than one nominal categorical background variable affects response category interpretation, we could create a group variable by crossing the levels of all variables expected to influence response category interpretation.

σ^i describes the spread with which person i's d_{ij}s map onto the agreement continuum. This parameter can be thought of as a quantification of *extremity*, which is a greater tendency to use outer response categories regardless of the questions asked.

2. *Group-specific Response Structure:* All members of each Z group share the same interpretation of the response categories, but this interpretation can differ between groups. In formal terms, this means that a group-specific threshold set, $\{c_1^{(g)}, \ldots, c_{K-1}^{(g)}\}$, is used in the model's response structure.

   (a) *Unrestricted:* The group-specific threshold set is allowed to vary across groups in an unrestricted manner.

   (b) Restricted: The group-specific threshold set varies across groups in a restricted (shifted and scaled) manner. More specifically, for group g,

   $$c_k^{(g)} = (\sigma^{(g)})^{-1} c_k + \tau^{(g)} \text{ for } k = 1, \ldots, K,$$

   where $\{c_1, \ldots, c_{K-1}\}$ is the common threshold set; and where $\tau^{(g)}$ and $\sigma^{(g)}$ are referred to as *group interpretation parameters*. These parameters are analogous to the person interpretation parameters discussed above.

3. *Common Response Structure*: All people share the same response category interpretation. In formal terms, this means that a common threshold set, $\{c_1, \ldots, c_{K-1}\}$, is used for all persons.

3.2.2 Incorporating 'Don't Know / Can't Choose' Responses

The ULTMODR, as presented above, is intended for data containing ordinal responses. However, as noted in Chapter 1, Likert data often contains 'Don't Know / Can't Choose' responses. To incorporate these into analysis using the ULTMODR, we could modify the data, treating the 'Don't Know / Can't Choose' responses as missing, or else modify the method. The former approach is easier to implement, but may introduce bias into the analysis, especially when the proportion of 'Don't Know / Can't Choose' responses is high.

One way of modifying the data is to omit any 'Don't Know / Can't Choose' responses. We could remove persons with any 'Don't Know / Can't Choose' responses prior to fitting the model to the data. Alternatively, we could ignore any 'Don't Know /

Can't Choose' responses during the model fitting process by omitting the corresponding terms from the likelihood.

Another way to modify the data it to replace the 'Don't Know/Can't Choose' responses with ordered agreement responses. For instance, we could adopt a (multiple) imputation approach in which any 'Don't Know / Can't Choose' responses would be replaced with imputed values falling into one of the ordered categories. Alternatively, if the ordered response categories contain a middle 'Neither agree nor disagree' type category, then the 'Don't Know / Can't Choose' responses could be recoded as this middle category.

Last, we could modify the ULTMODR, which involves specifying an ICRF for the 'Don't Know / Can't Choose' category. For example, we might use the ICRF employed for the ordered response categories, but with the thresholds c_k^i and c_{k-1}^i replaced by two additional thresholds ($c_{DK,u}^i$ and $c_{DK,l}^i$, respectively), where no ordering constraints are imposed on these additional thresholds. Alternatively, we could look to the numerous item response theory models (see van der Linden and Hambleton, 1997) for a different ICRF appropriate for the 'Don't Know / Can't Choose' category.

3.3 Comparison to existing latent variable models

In this section, we describe how the ULTMODR relates to other models, it order to put it into some context.

The simplest variant of the ULTMODR possesses the attribute that characterises latent trait models for ordinal variables (Luo, 2001, Theorem 1): The expected response function for item j is a unimodal function of the latent trait, θ_j^s, with its mode at the item location, β_j^s. (See the next section for a proof).

However, the ULTMODR differs in several aspects from other unfolding latent trait models for ordinal variables. For one, the latent structure can be multidimensional in the ULTMODR, whereas the other models are unidimensional and their structure does not make it easy to incorporate additional attitudes (latent traits). Another difference centres around the way in which the models create ICRFs that reflect an unfolding process. In the other models, an unfolding structure is induced by assuming unobserved response categories whose probabilities are monotonic in the latent trait. In the ULTMODR, on the other hand, an unfolding structure is induced by using a non-monotone function (absolute value) to directly model the probabilities of the observed response

categories. In other words, in the ULTMODR, the latent structure unfolds instead of the response structure. As a result, the response structure can be allowed to vary for persons or groups of persons. It is much less feasible to incorporate differing response category interpretations into the other models because it is their response categories that unfold.

The way that the ULTMODR induces unfolding does resemble the approach used in other types of unfolding models, such as the one developed for rank data (see Chapter 4).[3] Generally speaking, these models locate stimuli (e.g., statements) in a Euclidean latent space, and then treat each person's responses as a (deterministic or probabilistic) function of the distance between his ideal point (i.e., person location) and the stimuli locations. These unfolding models, like the ULTMODR, more explicitly reflect Coombs' (1964) description of an unfolding process than do models like the GGUM and GHCM.

Analogously, the ULTMODR's response structure does resemble those monotone latent trait models for ordinal variables that use a underlying-variable-coarsened-by-thresholds approach (e.g., the UVMOV). The ULTMODR's response structure particularly resembles Shi and Lee's (1998) Bayesian approach to fitting an UVMOV-like model, where the thresholds are treated as person-specific random effects. However, in Shi and Lee's approach, the thresholds vary across persons in an unrestricted manner, whereas the ULTMODR models threshold variation using a shifting parameter and a scaling parameter.

In fact, the ULTMODR models threshold variation in a manner very similar to Rossi et al.'s (2001) hierarchical model for differing response category interpretation. In some senses, the ULTMODR is a variation on Rossi et al.'s model, a variation in which the unobserved continuous variables are modelled as a function of underlying latent traits (attitudes). For the sake of convenience, we adopt some of Rossi et al.'s notation when describing the ULTMODR and some of their constraints when fitting the ULTMODR.

3.4 Model fitting and identifiability

Having introduced the ULTMODR, we now describe a frequentist approach to fitting it. This approach treats any person-level parameters (i.e., person locations and, if rele-

[3] Cox and Cox (2001, Chapter 8) and Marden (1995, Section 10.4) overview unfolding models for rank, pairwise comparison, and dissimilarity data.

vant, the person interpretation parameters) as random effects, and all other parameters (and hyperparameters) as fixed effects.

The distributions of random effects are chosen for computational convenience. We assume that θ_i comes from a normal distribution, denoted $g_1(\theta_i)$, with hyperparameters that may be fixed or estimated. If Case 1 is used, we assume that the person interpretation parameters come from a bivariate normal distribution, denoted $g_2(\tau^i, \ln(\sigma^i))$, that has mean vector φ and covariance matrix Λ, which are estimated. We might allow this distribution to depend on person i's Z group, specifically by allowing each group's distribution to have different hyperparameters (now referred to as $\varphi^{(g)}$ and $\Lambda^{(g)}$). (We will refer to the scenario where $g_2(\tau^i, \ln(\sigma^i))$ has group-specific hyperparameters as Case 1a, and the scenario where all groups share common hyperparameters as Case 1b.) We note that the distribution of person locations and the distribution of person interpretation parameters are assumed to be independent. Although we could easily imagine a scenario where underlying attitudes and response category interpretation are not independent, we choose not to model their relationship in order to simplify the model.

Before fitting the model, we must impose constraints on some parameters to make sure that it is identified. Of course, the particulars of identifiability will depend on the case and variant of the ULTMODR and on the particular questions being investigated. However, we can still note here some general principles of identifiability. Beginning with the latent structure, the person locations are additively confounded with the statement locations, as can be seen in (3.3). We could deal with this identifiability problem by setting to $\mathbf{0}$ the mean of $g_1(\theta_i)$ for at least some people.[4] Alternatively, we could fix the item locations to pre-specified values. Turning to the response structure, in Case 2b, the $\tau^{(g)}$s are additively confounded with the thresholds, and the $\sigma^{(g)}$s are multiplicatively confounded with the thresholds. These problems are resolved by constraining $\tau^{(0)} = 0$ and $\sigma^{(0)} = 1$, respectively. Similar identifiability problems arise in Case 1b and are resolved by setting $E(\tau_i)$ to 0 and $E(\sigma_i^2)$ to 1, which translates into the constraints $\varphi_1 = 0$ and $\varphi_2 = -\lambda_{2,2}$. (In Case 1a, the constraints $\varphi_1^{(g)} = 0$ and $\varphi_2^{(g)} = -\lambda_{2,2}^{(g)}$ are used only for group 0's distribution.) Last, there may be confusion (i.e., imperfect confounding) between the latent structure and the response structure in certain analyses using the ULTMODR. In response, we may want to set the variance of θ_i to a pre-specified value if the item locations are not fixed.

[4]Note that, even with this constraint, the item locations and person locations will be identified only up to a change of sign.

The fixed-effects parameters are estimated by maximizing the likelihood, which is produced by integrating out the random effects from the product of the conditional distribution of the data (given the random effects) and the marginal distribution of the random effects. We will refer to the likelihood as the *marginal likelihood (ML)* as is commonly done in item response theory literature; note that it is marginal only in the sense that the latent traits have been integrated out, unlike in the *joint likelihood* where they are treated as fixed effects and estimated simultaneously with the other model parameters. For the simple variant of the ULTMODR, the marginal likelihood takes the form:

$$ML = \prod_{i=1}^{n} \left\{ \int g_1(\theta_i^s) \cdot \prod_{j \in I_s} \prod_{k=1}^{K} I\left(Y_{ij} = k\right) P\left(Y_{ij} = k\right) d\theta_i^s \right\}, \tag{3.6}$$

or, for Cases 1a and 1b,

$$\begin{aligned} ML \;=\; & \prod_{i=1}^{n} \int \cdots \int \left\{ g_1(\theta_i^s) \cdot g_2(\tau^i, \ln(\sigma^i)) \cdot \right. \\ & \left. \prod_{j \in I_s} \prod_{k=1}^{K} I\left(Y_{ij} = k\right) P\left(Y_{ij} = k\right) \right\} d(\theta_i^s) d(\tau^i) d(\ln(\sigma^i)). \end{aligned} \tag{3.7}$$

(Marginal) MLEs will be used even for any variance parameters that are estimated (e.g., the $\sigma^{(g)}$s, if Case 2b is used, or the variances of any random effects distributions). For these parameters, we might consider using an alternative and more complex method, like REML, that avoids under-estimating the variances. However, since our primary interest lies in the item and person locations, using MLEs for the variances will be sufficient for our purposes.

The person locations are then estimated using the mean of the conditional distribution of the random effects (given the data), with the fixed effects parameters set equal to their MLEs.

3.5 Assessing goodness-of-fit

Assessing goodness-of-fit for the ULTMODR is difficult because the number of cells in the J-way contingency is typically large, making sparsity a problem. For instance, the NIS dataset has only 1805 persons, but there are $11^5 = 161,051$ cells. As a result,

most response patterns are not observed, and very few response patterns (only eleven) are repeated.

To assess overall goodness-of-fit, we can use the maximum value of $\ln(ML)$ for the ULTMODR. Because the standard asymptotic theory is not appropriate for ungrouped data, we will not assess $\ln(ML)$ in terms of deviance (which equals $-2\ln(ML)$ for ungrouped multinomial data). Instead, we could interpret $\ln(ML)$ in terms of logarithmic scoring (Good, 1983), which sums the negative log probability of the event that occurred. Under this interpretation, $\ln(ML)$ can be translated into an average probability: On average, the probability that a person selects the item-category that he did is $exp\left\{\ln(ML)/(n*J)\right\}$. Alternatively, we could interpret $\ln(ML)$ by comparing it to the analogous value for various proportional odds models (for all the items). One example is a model with no covariates but different category cut-offs for each item. This model, which is equivalent to a product-multinomial model, assumes that there are differences in the items but not in the persons. Another example is a model with (fixed) effects for persons and items but the same category cut-offs for all items. This model assumes that there are differences in the items and in the persons, but does not distinguish between person differences in attitudes and person differences in response category interpretation.

In addition to looking at overall goodness-of-fit, we can assess how well the expected probabilities (predicted by the ULTMODR) match the observed probabilities for the univariate and bivariate margins. To compare the expected and observed probabilities, we calculate signed Pearson residuals for the relevant margins. The signed univariate Pearson residual for category k of item j is

$$\chi^2_{j(k)} = \text{sign}\left(p_{j(k)} - \hat{p}_{j(k)}\right) \cdot \frac{n\left(p_{j(k)} - \hat{p}_{j(k)}\right)^2}{\hat{p}_{j(k)}}, \tag{3.8}$$

where $p_{j(k)}$ is the (observed) proportion of respondents who select category k for item j; and where $\hat{p}_{j(k)}$ is the expected probability of selecting category k for item j, under the fitted ULTMODR. Similarly, the signed bivariate Pearson residual for category k of item j and category m of item l is

$$\chi^2_{j(k)l(m)} = \text{sign}\left(p_{j(k)l(m)} - \hat{p}_{j(k)l(m)}\right) \cdot \frac{n\left(p_{j(k)l(m)} - \hat{p}_{j(k)l(m)}\right)^2}{\hat{p}_{j(k)l(m)}}, \tag{3.9}$$

where $p_{j(k)l(m)}$ and $\hat{p}_{j(k)l(m)}$ are the bivariate analogues of $p_{j(k)}$ and $\hat{p}_{j(k)}$. For ULTMODR cases 2a, 2b, and 3, the expected univariate and bivariate probabilities are

calculated using

$$\hat{p}_{j(k)} = \int P\left(\mp|\theta_i^s - \hat{\beta}_j| + \hat{c}_{k-1}^i \leq \varepsilon_{ij} \leq \mp|\theta_i^s - \hat{\beta}_j| + \hat{c}_k^i\right) g_1(\theta_i^s) d\theta_i^s$$

and

$$\begin{aligned}\hat{p}_{j(k)l(m)} &= \int P\left(\mp|\theta_i^s - \hat{\beta}_j| + \hat{c}_{k-1}^i \leq \varepsilon_{ij} \leq \mp|\theta_i^s - \hat{\beta}_j| + \hat{c}_k^i\right) \\ &\quad P\left(\mp|\theta_i^s - \hat{\beta}_l| + \hat{c}_{m-1}^i \leq \varepsilon_{il} \leq \mp|\theta_i^s - \hat{\beta}_l| + \hat{c}_m^i\right) g_1(\theta_i^s) d\theta_i^s,\end{aligned}$$

respectively, where $\hat{\beta}_j$ and $\hat{c}_k$ are the MLEs for those parameters. (For cases 1a and 1b, analogous expressions are used.) The sum of the unsigned univariate or bivariate residuals[5] cannot be used to formally test the fit of the ULTMODR because the model was not fitted using either of those marginal frequencies. Bartholomew et al. (2002) and Jöroreskog and Moustaki (2001) give heuristic arguments supporting two rules of thumb for deciding whether the sum of the bivariate residuals indicates poor fit. However, we will use the residuals in (3.8) and (3.9) simply to see which frequencies are most poorly predicted by the ULTMODR, in order to get insight into the model and the data.

3.6 A small simulation experiment

We performed a small simulation experiment using the simplest ULTMODR variant. The experiment was designed with two purposes in mind: (i) To see whether we should model response category interpretation if we are mainly interested in determining the order of the items, and (ii) To see what range of $log(ML)$ values indicates a well-fitting model.

First, $N = 100$ datasets were generated from Case 1a of the simplest ULTMODR variant. Recall that this case allows response category interpretation to differ by person. The datasets contained $n = 140$ persons (the same as the abortion attitude dataset) and $J = 6$ items comprising a Likert scale. In the latent structure, the item locations were fixed at $\beta_1 = -4$, $\beta_2 = -3$, $\beta_3 = -2$, $\beta_4 = 2$, $\beta_5 = 3$, and $\beta_6 = 4$. These values were chosen because Likert scales often contain a small number of statements ranging from very to extremely pro-object and the same number of statements ranging

[5]The $\chi^2_{j(k)}$ and $\chi^2_{j(k)l(m)}$ statistics are signed versions of the univariate and bivariate GF-Fit statistics proposed by Jöreskog and Moustaki (2001) in the context of the UVMOV.

from very to extremely anti-object. For each person, θ_i was generated from a $N(0,1)$ distribution. In the response structure, the thresholds were fixed at $c_1 = 0.5$, $c_2 = 1.25$, $c_3 = 1.75$, and $c_4 = 3.0$. For each person, τ_i and $\ln(\sigma_i)$ were generated from a bivariate normal distribution whose hyperparameters were chosen from experience with actual Likert data. (Their values were $\phi_1 = 0$, $\phi_2 = -0.16$, $\lambda_{1,1} = 0.36$, $\lambda_{1,2} = 0.22$, and $\lambda_{2,2} = 0.16$.)

Then, Case 3 of the simplest ULTMODR variant was fit to each of the simulated datasets. The $\hat{\beta}_j$s ordered the items correctly for every dataset, suggesting that the ULTMODR can recover the true item order without modelling differing response category interpretation. For each dataset, we calculated $exp\{log(ML)/(140 \cdot 6)\}$ and found that the average probability (that a person selected the response category that he did) ranged between 0.25 and 0.30. Since response category interpretation is not allowed to differ in Case 3, we would expect these values to be on the lower side. For this reason, we also fit Case 1a to each dataset. The resulting average predicted probabilities ranged between 0.30 and 0.34. Thus, for an actual dataset, we should not be surprised to see average predicted probabilities lower than 0.30.

Out of curiosity, we fit the GGUM with $\alpha_i = 1$ and $\tau_{ik} = \tau_k$ (see Roberts, 2000) and the one dimensional normal linear factor model to each simulated dataset. Note that both models assume the same response category interpretation for all persons. The item locations estimated for the GGUM ordered the items correctly for every dataset. On the other hand, the loadings estimated for the NLFM never ordered the items anywhere near their true ordering. Note, however, that in some situations, plotting the loadings from the two dimensional NLFM can give us an idea of the true item ordering. An example can be seen in Figure 3.1. The plot contains the estimated two dimensional loadings for a simulated dataset with twenty items evenly spaced between -4 and 4. Note that the true item ordering can be roughly recovered by reading around the horseshoe formed by the items.

3.7 Proof of the Unimodality of the Expected Item Response Function

The *expected response function* for item j is

$$E(Y_{ij} \,|\, \theta_i) = \sum_{k=1}^{K} k \cdot P\left(Y_{ij} = k\right), \tag{3.10}$$

where $k = 1, \ldots, K$.

Theorem 1 *In the simplest ULTMODR variant, the expected response for item j, where $S(j) = s$, is a unimodal function of θ_i^s, with the single mode occurring at $\theta_i^s = \beta_j^s$.*

Proof. Without loss of generality, we assume that category K represents 'Agree strongly,' which means that a negative sign is used in Equation (3.1). Thus,

$$E(Y_{ij} \,|\, \theta_i) = \sum_{k=1}^{K} k \cdot P\left(c_{k-1}^i + \left|\theta_i^s - \beta_j^s\right| \leq \varepsilon_{ij} \leq c_k^i + \left|\theta_i^s - \beta_j^s\right|\right). \tag{3.11}$$

Obviously,

$$E(Y_{ij} \,|\, \theta_i) = \sum_{k=1}^{K} k \cdot \left\{P\left(\varepsilon_{ij} \leq c_k^i + \left|\theta_i^s - \beta_j^s\right|\right) - P\left(\varepsilon_{ij} \leq c_{k-1}^i + \left|\theta_i^s - \beta_j^s\right|\right)\right\}.$$

Simplifying the right-hand side yields

$$E(Y_{ij} \,|\, \theta_i) = 1 + \sum_{k=2}^{K} P\left(\varepsilon_{ij} \geq c_{k-1}^i + \left|\theta_i^s - \beta_j^s\right|\right). \tag{3.12}$$

Each of the $K - 1$ probability terms in (3.12) increases as the expression $c_{k-1}^i + \left|\theta_i^s - \beta_j^s\right|$ decreases. Since this expression is a U-shaped function of θ_i^s with a minimum at β_j^s, each term in the summation in (3.12) is thus a unimodal function of θ_i^s with the single mode located at β_j^s. Clearly, summing these terms (and adding 1) to obtain $E(Y_{ij} \,|\, \theta_i)$ results in a function that is unimodal in θ_i^s with the single mode located at β_j^s. ■

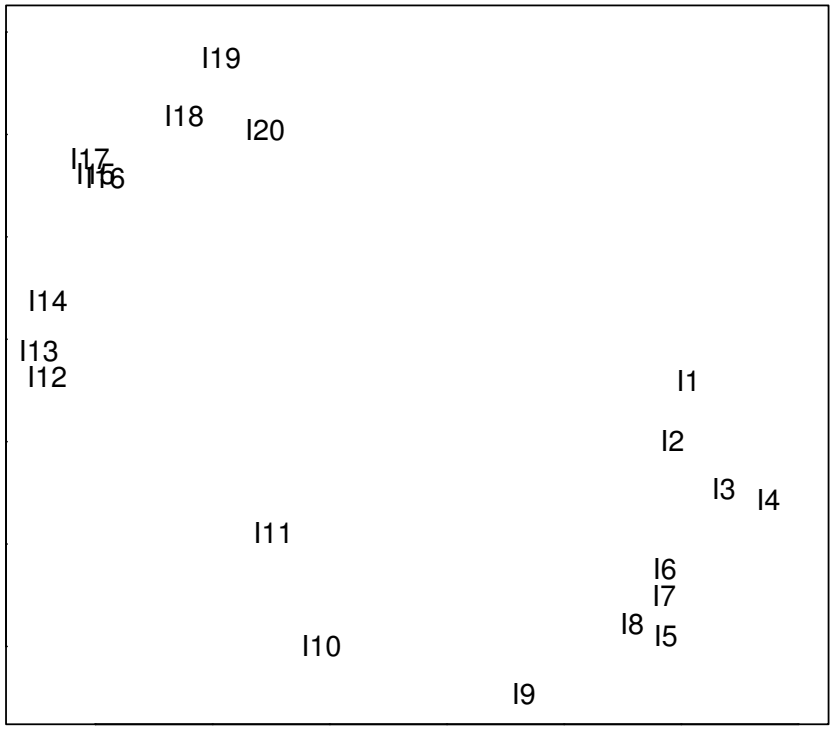


Figure 3.1: NLFM Item Loadings for a Dataset Simulated from the ULTMODR. *The plot shows* $\hat{\lambda}_{j,2}$ *vs* $\hat{\lambda}_{j,1}$ *for the two dimensional normal linear factor model. The model was fit to a dataset with* 140 *persons and* 20 *items that was generated from Case 3 of the simplest ULTMODR variant. The locations of the twenty items were fixed at equally spaced intervals from -4 to 4.*

Chapter 4

Visualisation

Following in the tradition of Exploratory Data Analysis (Tukey, 1977), we would like to visualise Likert data prior to performing any formal statistical analysis. Specifically, we would like to locate both the persons and the statements as points in the same plot. This chapter focuses on a visualisation technique that we term *multidimensional unfolding analysis* (*MUA*), which does just that.

We describe two methods that are appropriate for performing MUA on Likert data. The first, an existing method, uses the unfolding model for rank data introduced in Section 2.2. The second, a new method, uses a variant of the ULTMODR developed for the purpose of visualization. Since this method uses a formal probability model, confidence regions can be obtained for the person and statement locations. Both methods are employed to visualize the Likert items in three synthetic datsets and in the abortion attitude dataset.

4.1 Overview of multidimensional unfolding analysis

Multidimensional unfolding analysis locates members from two sets or *modes* in a Q-dimensional Euclidean space in such a way that the distances between objects from different modes best match observed two-way dissimilarities between objects from different modes.

MUA has its origins in the concept of unfolding formulated by Coombs' (1950) in his original MUA method.[1] The concept was developed for a situation where judges order objects in terms of decreasing preference. The judges and objects are assumed

[1]Note that our definition of MUA encompasses some methods that do not truly reflect Coomb's implementation of the unfolding concept.

to be located along a continuum (in Coombs' original formulation; in subsequent formulations by Coombs and others, objects and judges are located in multidimensional spaces). Folding the continuum at a judge's location generates a ranking of the objects that depends on their relative distances from the judge's location. Unfolding refers to performing the reverse process: Given observed object rankings for each of the judges, locating the judges and objects along the continuum. Of course, it is not usually possible to locate the objects and judges in such a way that the generated ranking for each judge perfectly matches his observed ranking.

The concept of unfolding—and, more generally, MUA—can be applied to Likert data. In the terms used by Coombs, Y_i can be viewed as a person's ranking of the statements (with ties). In the terms used in our MUA definition, Y_{ij} can be viewed as a measure of the dissimilarity between person i and statement j. This dissimilarity is measured at an ordinal level using categories whose interpretation may differ for different people. Of the many methods[2] encompassed by our definition of MUA, we focus on two that seem appropriate for dissimilarities with those two qualities. (In this chapter, any 'Don't Know / Can't Choose' responses in the Likert data must be omitted, recoded, or imputed).

In our presentation of these methods, we use $\theta_i = [\ \theta_{i,1} \quad \theta_{i,2} \quad \ldots \quad \theta_{i,Q}\ ]^T$ and $\beta_j = [\ \beta_{j,1} \quad \beta_{j,2} \quad \ldots \quad \beta_{j,Q}\ ]^T$ to refer to the locations of the persons and statements, respectively, in the Q-dimensional Euclidean latent space. (Typically, $Q = 2$.) As usual, d_{ij} will denote the Euclidean distance between person i's location and statement j's location.

4.2 An existing method of performing MUA

The *existing method* comes from the multivariate data visualisation context. It uses the unfolding model for rank data, which we recall seeks to locates the persons and statements so that the *order* of the d_{ij}s best matches the *order* of the Y_{ij}s within rows of $\mathbf{Y}$. The model is fit by algorithms operationally similar to ordinal MDS in the sense that they use a loss function to assess the similarity between distances and dissimilarities.[3]

[2] See Cox and Cox, 2001, Chapter 8 for descriptions of some methods.

[3] See Borg and Groenen (1997, Chapter 14) for an excellent presentation of this and other loss-function-based methods of performing MUA.

Although some algorithms use stress-based loss functions,[4] the algorithm ALSCAL (Young and Lewyckyj, 1979), which we use in the following applications, employs (a normalised version of) the s-stress loss function:

$$\text{s-stress} \quad = \quad \frac{1}{n} \sum_i \sum_j (d_{ij}^2 - d_{ij}^{*2})^2,$$

where the d_{ij}^*s for each i are a (weak) monotonic transformation of the Y_{ij}s for each i:

$$Y_{ij} < Y_{ij'} \Rightarrow d_{ij}^* \leq d_{ij'}^*. \tag{4.1}$$

ALSCAL allows ties in the Y_{ij}s to be preserved (Kruskal's secondary approach) or ignored (Kruskal's primary approach) by the d_{ij}^*s:

$$\begin{aligned} Y_{ij} = Y_{ij'} &\Rightarrow d_{ij}^* = d_{ij'}^* \quad \text{(secondary approach)} \\ Y_{ij} = Y_{ij'} &\Rightarrow \text{irrelevant} \quad \text{(primary approach)} \end{aligned}$$

We select the primary approach because having dissimilarity categories means a person will have the same response for statements towards which he doesn't react identically. In order to prevent degenerate configurations from being optimal, ALSCAL uses a normalised version of (raw) s-stress:

$$\text{s-stress(1)} = \frac{1}{n} \sum_i \frac{\sum_j (d_{ij}^2 - d_{ij}^{*2})^2}{\sum_j d_{ij}^4}. \tag{4.2}$$

This loss function does not take a value of zero for the degenerate configuration with all statements and persons in one location. However, it does take a value of zero for the degenerate configuration with the persons in one location and the statements located in a circle around them.[5]

Unfortunately, the existing method does not always perform well in practice, due to the limited amount of information contained in $\mathbf{Y}$.[6] We are trying to locate $n + J$ objects, but know only some of the dissimilarities between them: Table 4.1 show that

[4]The programs SSAR-II (Guttman-Lingoes series, 1973) and MINIRSA (Roskam, 1979) employ algorithms with stress-based loss functions. MINIRSA stands for Michicagna-Israel-Netherlands-Integrated Rectangular Smallest space Analysis. It is a modification of SSAR-II by Roskam, and it is available as part of the MDS(X) software package.

[5]The program MINIRSA avoids the second type of degenerate configuration by dividing stress by the quantity $\sum_j (d_{ij} - \bar{d}_i)^2$. The resulting function does not take a value of zero for either degenerate configuration.

[6]See Borg and Groenen (1997, Chapter 14) for an excellent discussion of the drawbacks of the existing method.

a large proportion of the dissimilarities are unknown and that, worse still, the missingness occurs in a systematic way. Further, not comparing the observed dissimilarities across rows means that we have even less information available to us. In the existing method, we have only $n \cdot [J(J-1)/2]$ order constraints, compared to the $(n+J)(n+J-1)/2$ that would be available if we had a complete dissimilarity matrix where comparisons could be made across rows. As a result, the existing method suffers from problems of indeterminancy. Thankfully, these are less severe when there are more statements.

Table 4.1: Full $n+J$ by $n+J$ dissimilarity matrix

	Persons				Statements			
Persons	?	?	?	?	Y_{11}	...	...	Y_{1J}
	?	?	?	?	...	...	...	...
	?	?	?	?	...	...	...	...
	?	?	?	?	Y_{n1}	...	...	Y_{nJ}
Statements	Y_{11}	...	...	Y_{n1}	?	?	?	?
	...	...	...	...	?	?	?	?
	...	...	...	...	?	?	?	?
	Y_{1J}	...	...	Y_{nJ}	?	?	?	?

4.3 A new method of performing MUA

As an alternative to the existing method, we introduce a new method that uses a *multidimensional variant* of the ULTMODR's latent structure along with the common threshold case (3) of its response structure. The ICRFs for this variant and case of the ULTMODR are

$$P\left(Y_{ij}=k\right)=P\left(\mp\sqrt{\sum_{q=1}^{Q}(\theta_{iq}-\beta_{jq})^2+\ c_{k-1}}\leq\varepsilon_{ij}\leq\mp\sqrt{\sum_{q=1}^{Q}(\theta_{iq}-\beta_{jq})^2+\ c_{k}}\right). \tag{4.3}$$

Our algorithm for estimating the model's statement and person locations follows the approach described in Section 3.4. The statement location and thresholds are estimated first by maximizing the marginal likelihood in equation (3.6), with $g_1(\theta_i)$ assumed to be a standard Q-variate normal distribution. Fixing the mean vector and

covariance matrix of $g_1(\theta_i)$ to pre-specified values resolves the translational and dilational invariance issues inherent in MUA. However, it does not resolve the rotational unidentifiability of the statement locations; as a result, their estimates will depend on the starting values used. Last, the person locations are estimated by the mean of their conditional distribution (given the data), with the fixed-effects parameters set equal to their MLEs (perhaps after rotation in the case of the statement locations).

Although our new MUA method would not perform well if the number of items were very small, it still performs well in moderate situations where the existing method does not. However, this gain comes at the expense of suitability: Our new method, though developed for Likert data, does not allow the interpretation of dissimilarities to differ between rows in the dissimilarity matrix.

Another advantage of the new method is that confidence regions can be obtained for each location. This is possible because the method uses a formal probability model to locate persons and statements.

To find a confidence region for each statement location, we use an approach that involves sampling from the large-sample approximation to the posterior distribution of the fixed-effects parameters (under a vague prior).[7] We set the covariance matrix of this multivariate normal distribution equal to the inverse of the observed Fisher information, evaluated at the MLEs of the fixed-effects parameters. (Due to certain types of invariance (e.g., rotational) in the statement locations, it may be necessary to set to zero one or more eigenvalues in the spectral decomposition of the covariance matrix.) In each of the N samples from the approximate distribution, the statement locations are subjected to the same rotation as the MLEs. Since the realizations of the Q coordinates for statement location j come from a normal distribution (even after rotation), the sample covariance matrix for the N realizations of those coordinates can be used to plot a confidence ellipse for the statement's location.

To find a confidence region for each person location, we use the fact that, for fixed values of the statement locations, the posterior distribution of the person locations factors into separate distributions for each person. The contour lines of the distribution for person i can then be used to plot a confidence region around his estimated location.

[7] We note that importance sampling could have been used to sample from the exact posterior distribution of the fixed-effects parameters (under a vague prior).

4.4 Application: Visualising synthetic datasets

The new and existing MUA methods were used to create two-dimensional plots of three synthetic datasets. All three were created under the assumption that a single attitude continuum underlies the persons and statements. Further, each was designed so that the performance of the methods could be easily assessed.

Here, one may question why we used two dimensions to visualize the data when a single continuum underlies them? There are two answers to this question. First, the data for some statements and some persons are not be consistent with one evaluative continuum. (In actual datasets, this can happen because different people interpret statements differently, or because certain statements unintentionally reflect an additional dimension.) Thus, plotting the data in more than one dimension can help us detect which statements and/or persons have data that are explained most poorly by one evaluative continuum. Second, even in the absence of the first problem, the nature of Likert data means that one-dimensional unfolding analysis can fail at ordering the statements sensibly. For instance, a strongly anti-object statement and a strongly pro-object statement might be placed near each other because their data appear similar (most people disagree with both of them).[8]

For each of the synthetic datasets, we used the SPSS implementation of ALSCAL to create the existing method plot and an R function that implements the fitting algorithm described in Section 4.3 to create the new method plot. For the latter, we also calculated 95% confidence regions for the statement and person locations using an R function that implements the approaches described in Section 4.3. Details pertaining to the use of ALSCAL and the R functions are contained below in Section 4.7.

4.4.1 Data with ordered persons and ordered statements

The first dataset contains fifteen persons' responses to ten statements (see rows 1-15 and columns 1-10 of Table 4.2). In creating the dataset, statements S-1 through S-10 were assumed to be ordered along the attitude continuum. Then, the responses for each person were created so that it would be clear which statement he should be located near: Each person strongly agrees with one statement and finds other statements

[8]This result becomes more likely when K is small and when most people's attitudes are moderate compared to the statements.

Table 4.2: Synthetic datasets

		Straightforward statements										Confusing statements	
		S-1	**S-2**	**S-3**	**S-4**	**S-5**	**S-6**	**S-7**	**S-8**	**S-9**	**S-10**	**S-11**	**S-12**
Unconfused persons	**P-1**	SD	SD	D	N	A	SA	A	A	N	D	A	SD
	P-2	D	N	A	SA	A	A	N	N	D	SD	N	D
	P-3	A	SA	A	A	A	N	D	D	SD	SD	D	A
	P-4	SD	SD	SD	D	D	D	N	A	A	SA	N	SD
	P-5	SD	D	N	A	SA	A	N	D	SD	SD	N	D
	P-6	N	N	A	SA	A	N	N	N	D	D	N	A
	P-7	D	D	N	A	A	A	SA	A	N	D	A	A
	P-8	SD	SD	D	N	A	SA	A	N	D	SD	A	A
	P-9	SA	A	N	D	SD	SD	SD	SD	SD	SD	SD	SD
	P-10	SD	SD	SD	SD	D	D	D	N	A	SA	A	D
	P-11	SD	SD	D	D	D	N	A	A	SA	A	A	N
	P-12	SD	SD	D	N	N	A	A	SA	A	N	N	A
	P-13	SD	SD	SD	N	N	A	SA	A	A	N	N	SD
	P-14	D	D	N	A	SA	A	A	N	D	SD	SD	D
	P-15	D	A	SA	A	A	N	N	D	SD	SD	SD	N
Confused persons	**P-16**	N	A	SA	A	N	D	N	A	SA	A		
	P-17	D	N	A	SA	A	A	SA	A	N	D		
	P-18	SA	SA	SA	SA	A	A	N	D	SD	SD		
	P-19	D	N	A	SA	SA	SA	A	N	D	SD		

increasingly (though not strictly) more disagreeable moving away from that statement. The persons can be ordered based on their data (albeit with some ties).

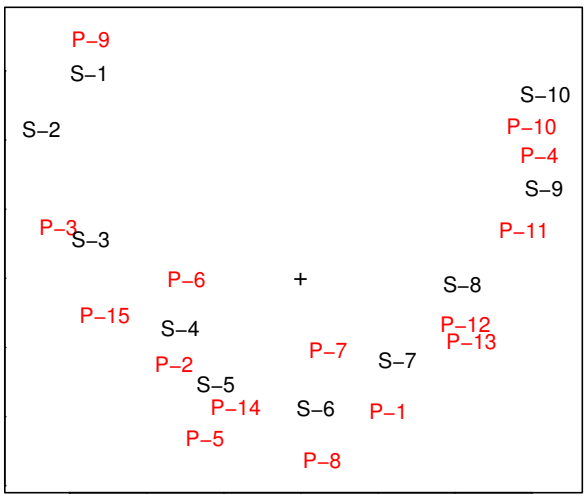
The MUA plots in Figure 4.1 reveal that both methods result in very similar configurations for the statements and persons. Both succeed in making apparent the true order of statements S-1 to S-10, by locating them in order around a horseshoe.[9] In addition, both methods succeed in locating each person (with the exception of P-3) nearest the statement with which he strongly agrees. We note that the standard normal prior used for the person locations in the new plot has the effect of pulling them towards the origin.

Confidence regions for the person and statement locations in the new plot are shown in Figure 4.2. A glance at the upper plot reveals that the more extreme statements have larger confidence regions. This is not surprising since most people strongly disagree with those statements, which means their location can be pushed further away from the origin without substantially changing the model's fit. In general, there is considerable overlap amongst the confidence regions for the statement locations, which we would expect given the small sample size. As for the person locations, their confidence regions are roughly the same size, but a little larger for those persons with more extreme locations.

4.4.2 Data with some confusing statements

To create the second synthetic dataset, two confusing statements were added to the first synthetic dataset (see rows 1-15 and columns 1-12 of Table 4.2). These additional statements do not have an obvious place in the statement ordering because different people view their locations differently. Some people think that statement S-11 falls between S-6 and S-7, whereas others think that it falls between S-9 and S-10. Similarly, some people think that statement S-12 falls between S-1 and S-2, and others think that it falls between S-5 and S-6.

[9]The horseshoe shape has also been noted in some applications of ordinal MDS where a single continuum does underlie the data. The original and most famous of these applications is the seriation of graves using grave good abundance matrices (Kendall, 1971). Various explanations have been proposed for why a continuum appears as a horseshoe when using ordinal MDS. Shepard (1974) suggests that the horseshoe effect occurs because the monotonic transformation of dissimilarities in ordinal MDS permits points along a continuum to be mapped to a semicircle. Alternatively, Kendall (1971) suggests that the horseshoe effect is an artefact (no pun intended!) of a bounded dissimilarity metric, where the largest dissimilarity is used for fairly dissimilar and for extremely dissimilar objects. Both of these explanations apply when MUA is performed on Likert data.



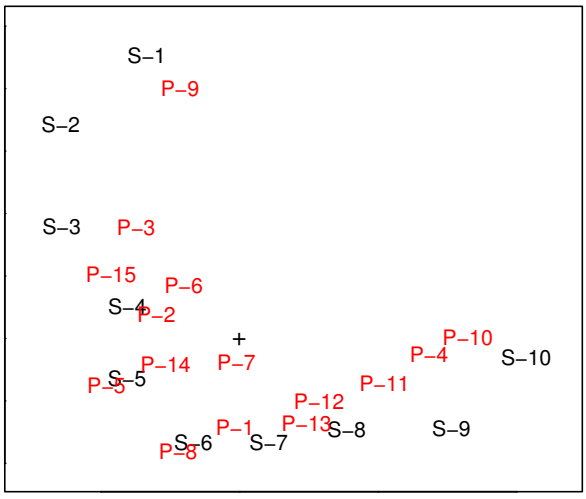


Figure 4.1: MUA of synthetic dataset 1 using the existing method (top) and new method (bottom). *Points corresponding to statements begin with "S" and appear in black. Points corresponding to persons begin with "P" and appear in red. In both plots, the origin is marked with a cross.*

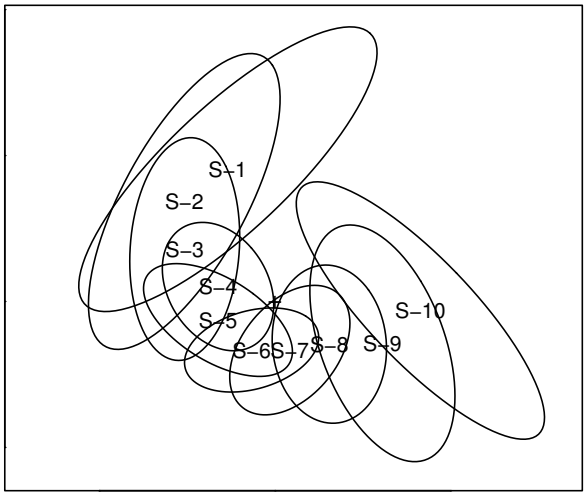


Figure 4.2: Confidence regions for the statement and person locations in the new MUA plot of synthetic dataset 1.

The plots in Figure 4.3 reveal that both methods still produce statement configurations that make apparent the ordering of statements S-1 to S-10: With neither method is the configuration significantly altered by the presence of two confusing statements. As for those two statements, the new method does a better job of conveying that they are different because it places them inside the horseshoe formed by the other statements. Though the existing method places S-12 inside the statement horseshoe, it makes it appear as if statement S-11 clearly falls in between S-9 and S-10. Both methods perform reasonably well at locating each person nearest the statement with which he strongly agrees, but not as well as when there are no confusing statements.

The confidence regions for the statement and person locations in the new plot can be seen in Figure 4.4. Note that the confidence regions for statements S-11 and S-12, though not large in an absolute sense, are reasonably large for centrally located statements. Note also that the confidence regions for the person locations are larger than when there are no confusing statements.

4.4.3 Data with some confused persons

To create the third synthetic dataset, four confused persons were added to the first synthetic dataset (rows 1-19 and columns 1-10 of Table 4.2). All four additional people strongly agree with more than one statement, making it unclear which statement each should be located near. Persons P-16 and P-17 strongly agree with two non-adjacent statements; these statements are far apart for person P-16, but fairly close for person P-17. Persons P-18 and P-19, on the other hand, strongly agree with several adjacent statements.

The MUA plots in Figure 4.5 reveal that, for both methods, the configuration of statement locations does not change much with the inclusion of several confused persons. In addition, the locations for the unconfused persons change only slightly (relative to the statement locations) when confused persons are included. As for the additional persons, both methods signal that there is something different about P-16 by locating it in the center of the statement horseshoe. This is only somewhat true for P-17, since his location is no more central than those of various unconfused persons (e.g., P-6). Neither method makes it apparent that there is something different about P-18 and P-19: Both simply locate each person nearest the statement(s) in the middle of those with which he strongly agrees.

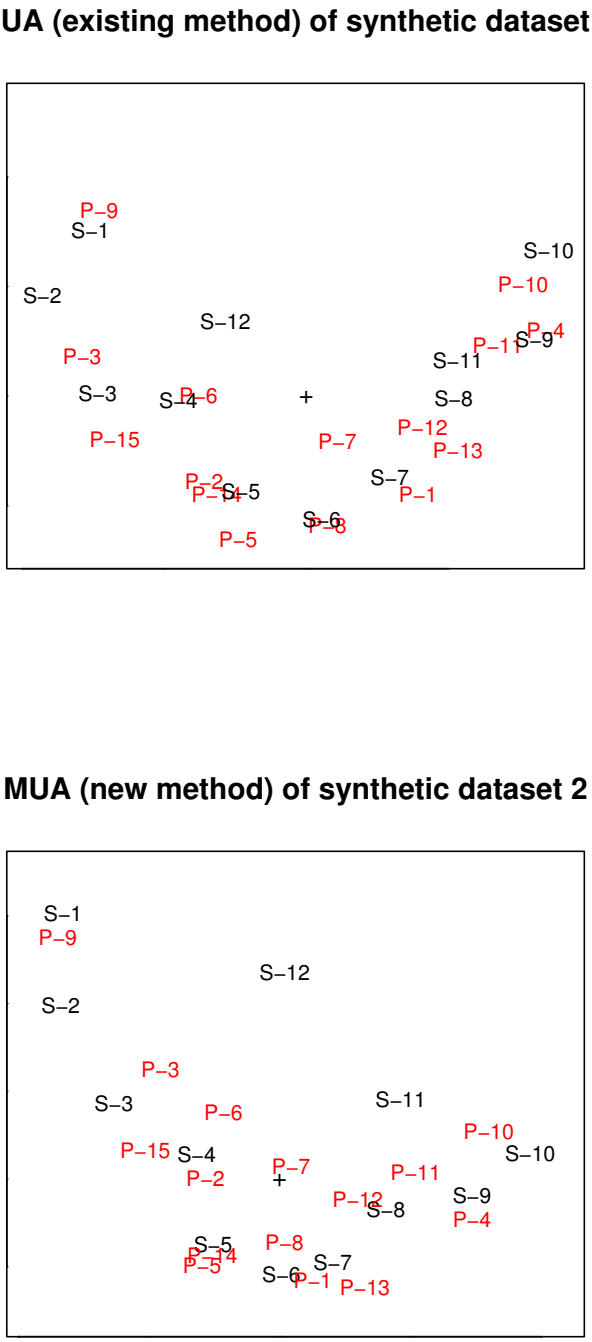
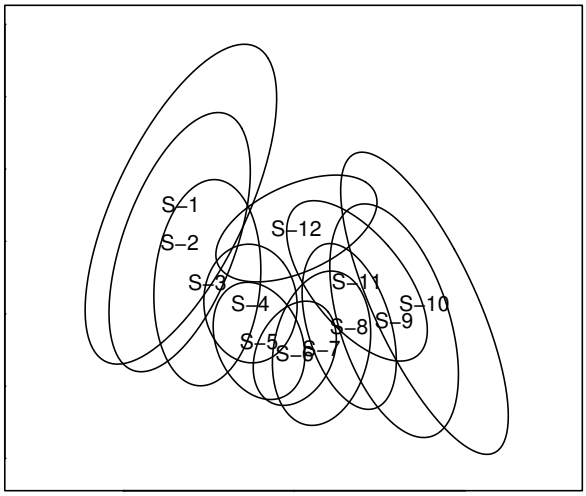


Figure 4.3: MUA of synthetic dataset 2 using the existing method (top) and new method (bottom). *Points corresponding to statements begin with "S" and appear in black. Points corresponding to persons begin with "P" and appear in red. In both plots, the origin is marked with a cross.*



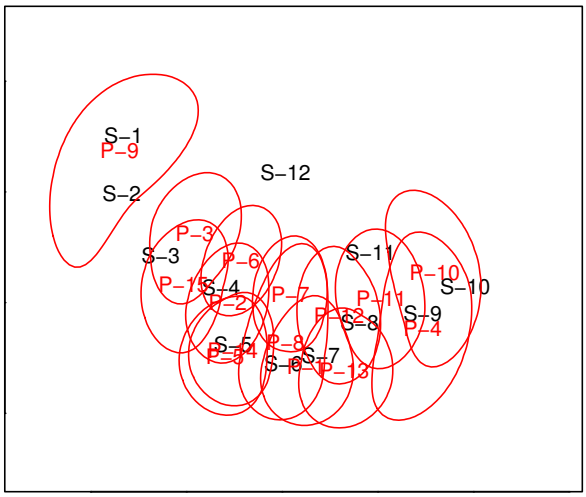


Figure 4.4: Confidence regions for the statement and person locations in the new MUA plot of synthetic dataset 2.

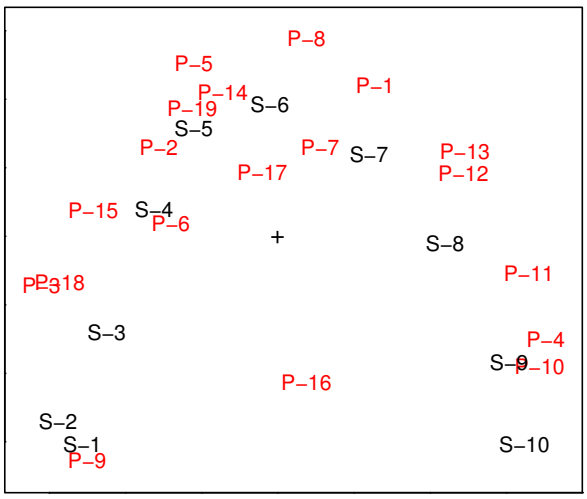
Figure 4.6 displays confidence regions for the statement and person locations estimated by the new method. In the upper plot, the confidence regions for the statement locations are not noticeably larger than when when there are no confused persons. The lower plot contains confidence regions only for each confused person (P-16 to P-19) and, for comparison, his nearest neighbour. Note that the confidence region for each confused person is no larger than the comparable region.

4.5 Application: Visualising abortion attitude data

We used both methods to perform MUA on the Likert items in the abortion attitude data. However, we first removed four items (14, 15, 19, and 21) that had more than 10% 'Don't Know/Can't Choose' responses; for the remaining 46 items, we recoded any 'Don't Know/Can't Choose' responses as 'Neither agree nor disagree.'

First, we used the existing method to perform MUA on the reduced, recoded data. More specifically, we used the SPSS implementation of ALSCAL (see Section 4.7 for details). We did not use the default ALSCAL starting configuration of statement and person locations because it resulted in a plot that did not make sense for the data. Strangely, the default starting configuration did result in a reasonable plot if any four items were omitted. Thus, we ran ALSCAL (from its default starting configuration) on the abortion data with the last four items removed since most people find them disagreeable. The final person locations from these 42 items were then used as starting values when ALSCAL was run on all 46 items (ALSCAL was allowed to calculate its own starting values for the statement locations). The resulting plot can be seen in Figure 4.7. The final s-stress value corresponding to this plot was unfortunately not included in the ALSCAL output.

Second, we used the new method to perform MUA on the data, using the aforementioned R code that implements the fitting algorithm in Section 4.3; details appear in the final section of this chapter. The algorithm was started from various initial values that were both random and based on experience with the model. Depending on the initial values used, the algorithm converged at different local maxima. The statement location and threshold values for the largest of the maxima found were retained as estimates. The estimates of the statement locations were then subjected to a varimax rotation (Kaiser, 1958) before estimating the person locations. The plot of the statement and person location estimates can be seen in Figure 4.8. The $log(ML)$ value corresponding



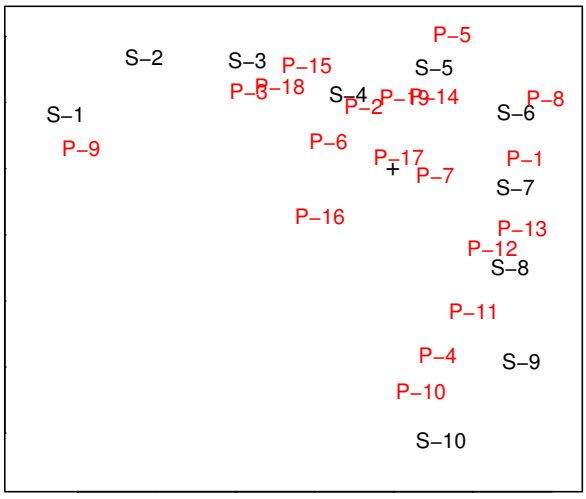
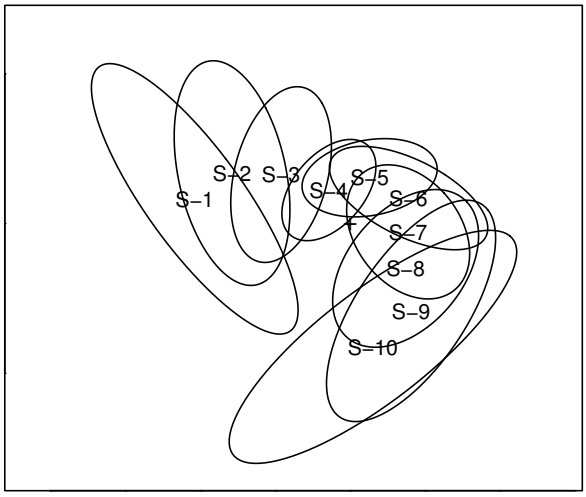


Figure 4.5: MUA of synthetic dataset 3 using the existing method (top) and new method (bottom). *Points corresponding to statements begin with "S" and appear in black. Points corresponding to persons begin with "P" and appear in red. In both plots, the origin is marked with a cross.*



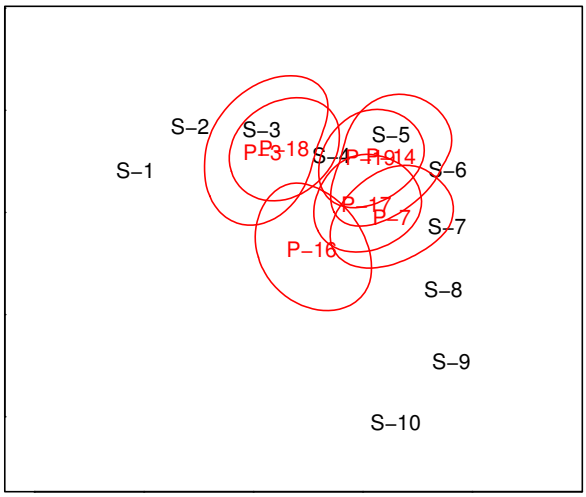


Figure 4.6: Confidence regions for the statement and person locations in the new MUA plot of synthetic dataset 3.

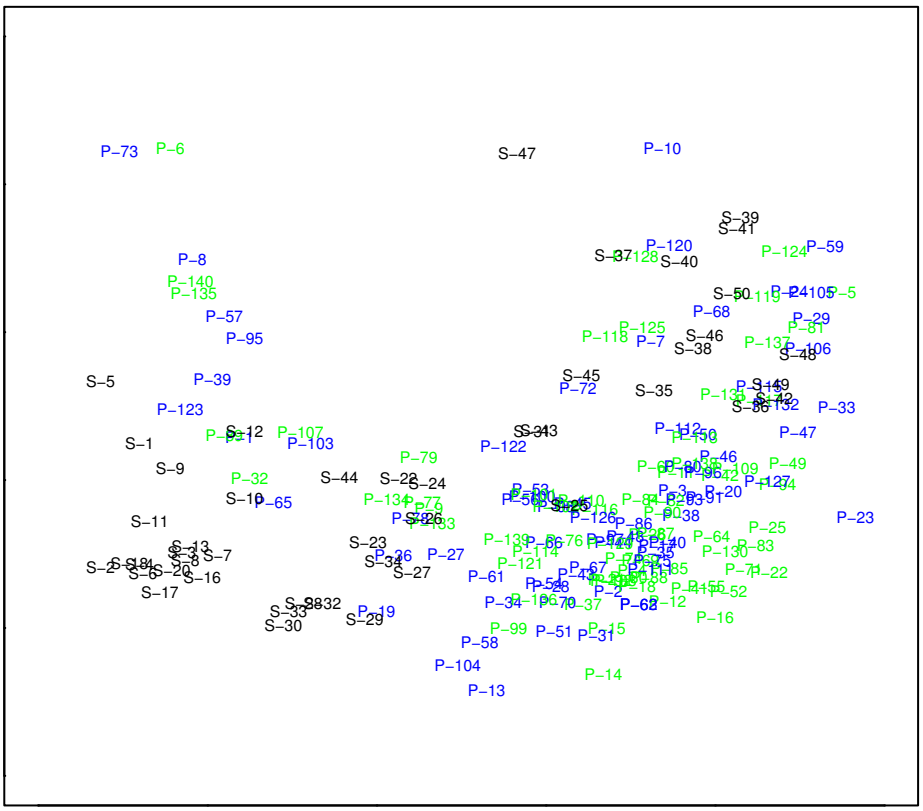


Figure 4.7: Existing MUA Method. *The plot was produced using ALSCAL with a user-supplied starting configuration. Points corresponding to statements begin with "S" and appear in black. Points corresponding to persons begin with "P" and appear in green for British respondents and blue for American respondents.*

to this plot is -6940. Interpreted in terms of log probability scores (Good, 1983), this value corresponds to an average probability (that a person would choose the category that he did for a statement) of $0.34 = \exp(-6940/(140 \cdot 46))$. Since this probability is bigger than 0.20 $(1/K)$, it seems that the multidimensional variant and simplest case of the ULTMODR performs fairly well at predicting people's responses to the statements. However, we are more interested in how well the relationships (between statements and persons) depicted by the plots match those present in the data.

We compare the fit of the new and existing plots by calculating three statistics that assess how well, on average, the within-person rankings of the distances in the plot match the person's rankings of the statements. The three statistics are:

1. S-stress(1) from equation (4.2), where the d_{ij}^{*2}s are the predicted values from an isotonic regression[10] of d_{ij}^2 on δ_{ij}. The value of s-stress is 0.097 for the existing plot and 0.057 for the new plot.

2. A normalized version of stress,

$$\text{stress} = \frac{1}{n} \sum_i \frac{\sum_j (d_{ij} - d_{ij}^*)^2}{\sum_j d_{ij}^2}, \tag{4.4}$$

   where the d_{ij}^*s are the predicted values from an isotonic regression of d_{ij} on δ_{ij}. The value of stress is 0.043 for the existing plot and 0.020 for the new plot.

3. The average Kendall's correlation,

$$\bar{\tau}_b = \Sigma_{i=1}^n \tau_b^i, \tag{4.5}$$

   where τ_b^i is Kendall's τ_b correlation between the d_{ij}s and the δ_{ij}s for person i. The value of $\bar{\tau}_b$ is 0.62 for the existing plot and 0.66 for the new plot.

All three statistics suggest that the observed rankings of the statements are better matched by the new plot. This is somewhat surprising since the existing method explicitly seeks to find distances whose within-person ordering matches the within-person ordering of the dissimilarities as closely as possible. However, the existing method assesses this match using a loss function based on s-stress. As a result of the

[10]The regression was performed using the isoreg() function in the R library modreg. This function adopts weak monotonicity constraints and the primary approach to ties.

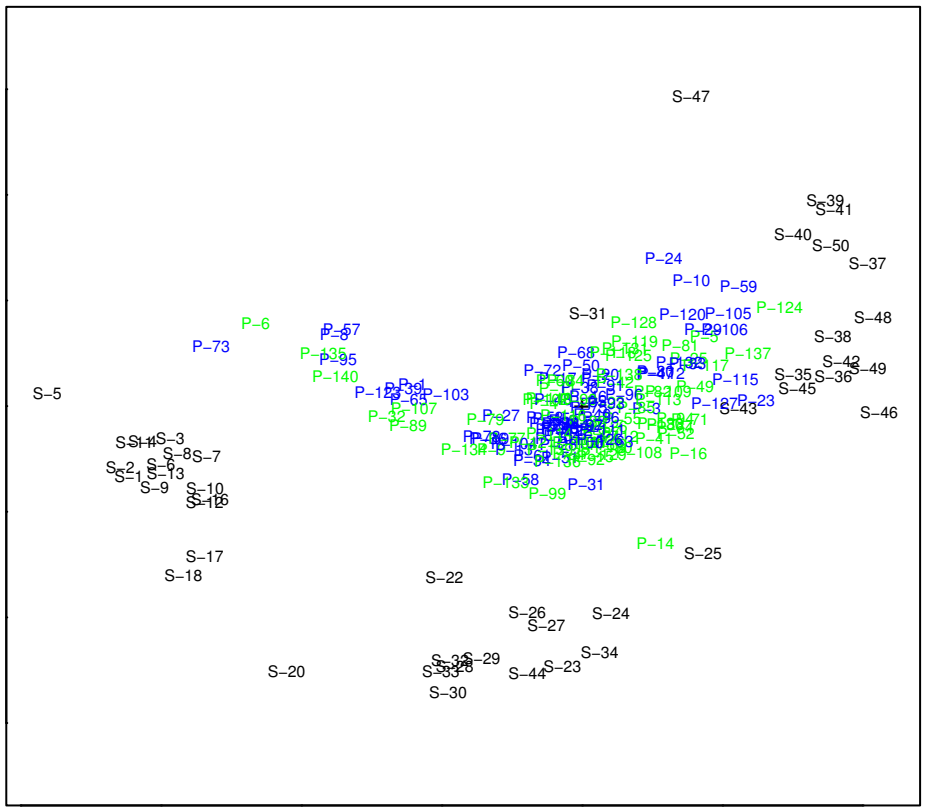


Figure 4.8: New MUA Method. *The plot was produced using R code written by the author, the details of which are in Section 4.7.2. Points corresponding to statements begin with "S" and appear in black. Points corresponding to persons begin with "P" and appear in green for British respondents and blue for American respondents.*

outer square in its formula, less weight is placed on fitting the data values indicating agreement, which we suspect may be the source of the existing method's poorer performance.

Next, we interpret the statement locations in the new and existing plots. In both plots, the statements are located roughly around a horseshoe that runs from S-47 (extremely pro-abortion) to S-5 (extremely anti-abortion). This horseshoe traces a continuum that, judging from the content of the statements, seems to correspond to the evaluative continuum for abortion attitudes. The statements' relative positions within the horseshoe differ in the two plots. Focusing on the new method plot only, the horseshoe can be divided into three clusters: An anti-abortion attitude cluster on the right-hand side, an ambiguous attitude cluster in the lower middle, and a pro-abortion attitude cluster on the upper left-hand side. Only S-20 cannot be assigned to one of these clusters. It lies between the anti-abortion cluster and the ambiguous cluster, as one might expect for a double-barreled statement with one moderate clause and one anti-abortion clause. S-31 clearly departs from the horseshoe shape, not surprisingly given that it asks people for their opinion on a factual rather than ethical question. If we want to form an abortion attitude scale, we will want to omit S-31 and S-20.

As for the person locations, in both plots they are also located along a horseshoe that, judging from each person's 46 responses, seems to correspond to the evaluative continuum representing abortion attitudes. For example, the person marking the left end of the horseshoe— P-73, a Catholic American female raised Catholic in an urban environment—strongly agreed with the anti-abortion statements, and strongly disagreed with almost all other statements. On the other hand, one of the people marking the left end of the horseshoe—P-10, a non-practicing American male raised Catholic in an urban environment—strongly agreed with the pro-abortion statements and strongly disagreed with almost all other statements. Although the persons' relative locations within this horseshoe differ in the two plots, the British and American locations are interspersed throughout the horseshoe in both plots. This suggests that the distribution of abortion attitudes is similar within the British and American groups in our sample. Similarly, although the shape and placement of the person horseshoe differs in the two plots,[11] it is located closer to the pro-abortion and ambiguous statements than to the anti-abortion statements in both plots. This suggests that our sample has, on average, fairly liberal attitudes towards abortion.

[11]The fact that the new method locates most persons near the origin is not surprising when we recall that a $N(\mathbf{0}, \mathbf{I})$ prior distribution is assumed for the person locations.

Last, we calculated confidence regions for all of the statement locations and two of the person locations in the new method plot. These regions are shown in Figures 4.9 and 4.10, respectively. The regions were calculated using the aformentioned R function that implements the approaches described in Section 4.3; the details are discussed in the last section of this chapter. The confidence regions for the statement locations suggest that the location of S-5 is the most uncertain. This is not surprising given that everyone disagrees strongly with it. As for the confidence regions for persons 14 and 31,[12] they indicate that there is more uncertainty in person 14's location.

4.6 Conclusions

We have introduced a new method for performing multidimensional unfolding analyis on Likert items. The new method is based on a multidimensional variant of the ULT-MODR, and it is an alternative to a popular existing method operationally similar to ordinal MDS.

The new and existing MUA methods seemed to perform well for all datasets, although both methods encountered some problems with local maxima for the abortion attitude items. These generally positive results are not surprising since all four datasets contained a fairly large number of items.

If we are interested in getting a sense of the general structure underlying Likert data, then we prefer to use the new MUA method. This is the case because the plots it produces are generally less messy and thus more clearly convey the relationships between statements and the relationships between persons and statements. This can be seen by comparing the abortion attitude plots in Figure 4.7 and 4.8 above or the two MUA plots in the following chapter. In the former plots, the new method plot makes it easier to tell that there are three types of abortion statements and that the statements are ordered (around a horseshoe). Further, because the persons are located inside this horseshoe, it is easier to compare the distributions of British and American attitudes.

Another advantage of the new method is that it allows us to obtain confidence regions for the statement and person locations, which the existing method does not because it is not based on a formal probability model. The second synthetic data

[12]These particular persons were chosen because they are located near each other, making it possible to compare their regions without having to adjust for the effect of extremity.

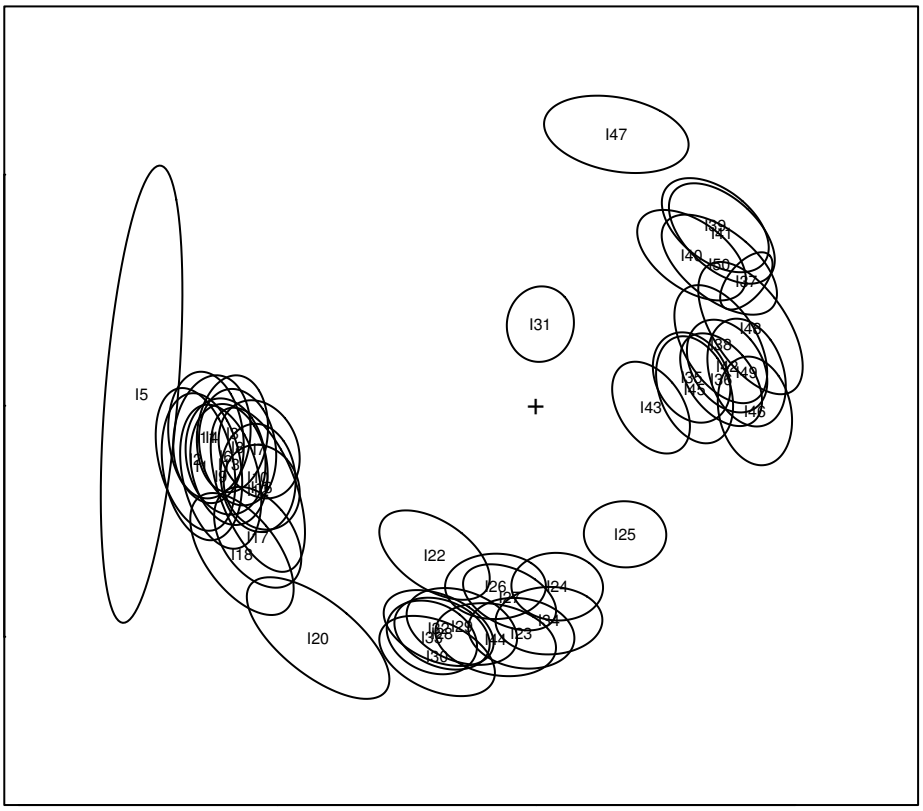


Figure 4.9: Confidence Regions for the Statement Locations Produced Using the New Method. *The plot contains the same statement locations as Figure 4.2, with confidence ellipses drawn around each statement location. These ellipses were obtained using the approach described in Section 4.3, and the details of their calculation are described in Section 4.7.2*. The origin is marked by an "X."

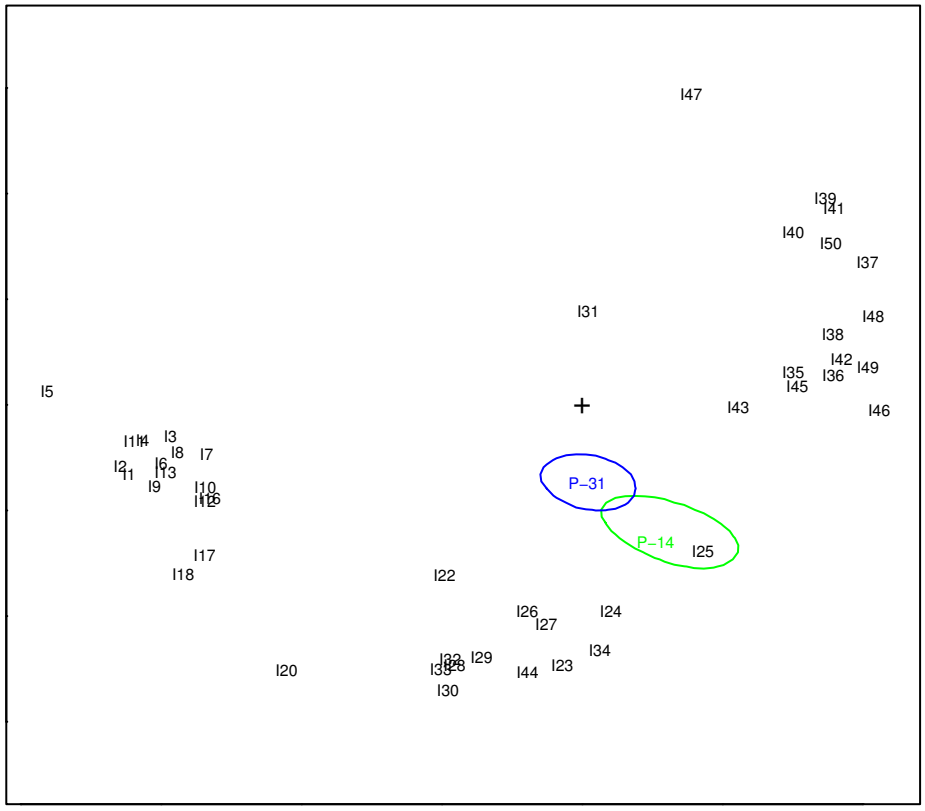


Figure 4.10: Confidence Regions for Two Person Locations Produced Using the New Method. *The plot contains the same statement locations as Figure 4.2, with confidence regions drawn around person locations 1 and 14. These regions were obtained using the approach described in Section 4.3, and the details of their calculation are described in Section 4.7.2.*

application suggests that the confidence regions for statement locations can assist us in identifying statements that are interpreted differently.[13]

The second synthetic data application also suggests that the new method statement locations themselves, not just their confidence regions, make it easier to identify confusing statements. However, in the synthetic data application in the following chapter on item analysis, the existing method locates more of the confusing statements inside the statement horseshoe. Thus, if we are interested in identifying anomalous items, then it is unclear which MUA method we should use.

4.7 Further details of the applications

In this section, we describe how we created the abortion attitude plots above. Unless noted in a footnote, the details are the same for the synthetic datasets.

4.7.1 The existing method

The following SPSS syntax creates the plot seen in Figure 4.7. ALSCAL performs loss-function-based multidimensional unfolding analysis when its "Rectangular" dataset option is selected. To get ALSCAL to perform the existing method of MUA, one also has to select the "Ordinal Measurement" and "Row Conditional Dissimilarities" options.

```
ALSCAL
  VARIABLES = v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v16 v17 v18
  v20 v22 v23 v24 v25 v26 v27 v28 v29 v30 v31 v32 v33 v34 v35 v36 v37
  v38 v39 v40 v41 v42 v43 v44 v45 v46
  /SHAPE = RECTANGULAR
  /LEVEL = ORDINAL (UNTIE)
  /CONDITION = ROW
  /MODEL = EUCLID
  /PRINT = HEADER
  /PLOT
  /CRITERIA = CONVERGE(.001) STRESSMIN(.005) ITER(30)
              CUTOFF(0) TIESTORE(5000) DIMENS(2,2)
  /OUTPUT = SPSSCORD.SAV .

ALSCAL
  VARIABLES = v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13 v16 v17 v18
  v20 v22 v23 v24 v25 v26 v27 v28 v29 v30 v31 v32 v33 v34 v35 v36 v37
```

[13] Disappointingly, the third synthetic data application do not suggest that the confidence regions for person locations can be used to identify confused persons.

```
v38 v39 v40 v41 v42 v43 v44 v45 v46 v47 v48 v49 v50
/FILE = SPSSCORD.SAV ROWCONF(INITIAL)
/SHAPE = RECTANGULAR
/LEVEL = ORDINAL (UNTIE)
/CONDITION = ROW
/MODEL = EUCLID
/PRINT = HEADER
/PLOT
/CRITERIA = CONVERGE(.001) STRESSMIN(.005) ITER(30)
            CUTOFF(0) TIESTORE(5000) DIMENS(2,2) .
```

4.7.2 The new method

This section describes the R code used to find estimates and confidence regions for the statement and person locations in the multidimensional variant and common case of the ULTMODR.

Fixed-point Gauss-Hermite quadrature (with a product rule) was used to approximate the integrals that appear in the marginal likelihood and the posterior distributions, respectively. More specifically, for each person, the integrand of his two-dimensional integral was evaluated at the lattice points of a two-dimensional grid formed by crossing 21 Gauss-Hermite abscissas. The person's integral was then approximated using a weighted sum of the 21^2 evaluations, with weights equal to the product of the corresponding Gauss-Hermite weights.

The fixed effects (i.e., statement locations and thresholds) were estimated by maximising the logarithm of the marginal likelihood in equation (4.3), with each of the n integrals approximated in the manner described above. The marginal likelihood was maximised with respect to the statement locations and the differences in the thresholds. (Box constraints were used to keep the differences positive to ensure that the thresholds remain ordered during the optimisation process.) Maximisation was performed using *R*'s optim() function with the "L-BFGS-B" method, which implements the algorithm introduced by Byrd et al. (1995) for optimisation with box constraints. The initial values for the threshold differences were $c_1 = 2$, $c_2 - c_1 = 1.5$, $c_3 - c_2 = 0.5$, and $c_4 - c_3 = 1.5$; these values were based on experience with the simplest variant of the model. Multiple sets of initial values were used for the statement locations: These values were both random and based on experience with the model. Depending on the

initial values used for the statement locations, the algorithm converged at different local maxima.[14] The initial statement locations that resulted in the largest maxima were, for the first dimension, spaced equally between -5 and 5 according to statement number and, for the second dimension, generated from a $N(0, 0.3)$ distribution. The final statement location and threshold values corresponding to the largest maxima were retained as estimates. The estimates of the statement locations were then subjected to a varimax rotation before plotting them.

The person locations were estimated with the thresholds and statement locations fixed at their MLEs (after varimax rotation in the case of the statement locations). Thus, the conditional distribution of the person locations (given the data) could be factored into separate distributions for each person. The location for each person was estimated by the mean of his distribution, with the two-dimensional integral approximated in the manner described above.

Confidence regions for the statement locations were calculated from $N = 100$ samples from a multivariate normal distribution with mean vector equal to the MLEs for the statement locations (before rotation) and thresholds and covariance matrix equal to the modified inverse of the observed information matrix, evaluated at the MLEs of the statement locations and thresholds.[15] Modification of the matrix was necessary because the observed information matrix should and did have one eigenvalue near zero due to the invariance of the marginal likelihood with respect to orthogonal rotations of the statements. Further, in the abortion attitude dataset, there was a second very small negative eigenvalue that probably resulted from some sort of scaling invariance in the likelihood. (This did not happen for any of the synthetic datasets.) All near zero eigenvalues were set to zero before inverting them to obtain the eigen-decomposition of the covariance matrix; the infinite eigenvalues that resulted were then set to zero since we did not care which rotation (or scaling) was generated. The modified matrix was then used to generate the $N = 100$ samples; the statement locations in each sample were subjected to a varimax rotation. Next, for each statement, the two-dimensional sample covariance matrix was calculated from the 100 rotated realizations of its two coordinates. Finally, we plotted an ellipse (around the estimated statement location)

[14] No problems with local maxima arose for any of the synthetic datasets.

[15] The observed information matrix was calculated by optim(), which returns a numerical approximation of the Hessian matrix at the solution found. Although optim() actually returns the Hessian matrix of the unconstrained problem, the box constraints are not active in the solutions found for our datasets.

that represented the 95% contour of the bivariate normal distribution with a covariance matrix equal to the sample covariance matrix.

A confidence region for each person location (given the statement locations) was obtained using his posterior distribution, with the integral approximated in the manner described above. The plotted confidence region for each person was the 95% contour for this distribution.

Chapter 5

Item analysis

In this chapter, we focus on selecting items (from an initial pool[1]) for inclusion in a Likert scale designed to measure attitudes towards object s. This process is known as *item analysis* and typically uses data from a pilot survey of the Likert items in the pool; we refer to this data as *screening sample data.* Ideally, we would like to select those items that best reflect attitudes towards object s. Obviously, assessing which items do so is difficult since we don't know people's attitudes (or else we wouldn't be trying to measure them!).

We describe how a variant of the ULTMODR can be used to perform item analysis in a manner that reflects researchers' stated aims. This new method is applied to a simulated dataset and to the Likert items in the abortion attitude data, and the results are compared to those produced using a popular method of item analysis.

5.1 Existing methods of item analysis

Researchers are interested in creating an attitude scale that is both valid and consistent. Bollen (1989) defines validity and consistency and discusses several methods used to assess them. Alternatively, Mueller (1986) and Spector (1990) discuss validity and consistency in the particular context of attitude scales.

Most methods for creating a Likert scale from screening sample data focus on selecting items that are internally consistent.[2] In these methods, the data are first quantified using the procedure described in Section 2.1, and internally consistent items are

[1] Mueller (1986, Chapter 2), Oppenheim (1992), and Roberts (1995, Chapter 19) make recommendations for generating an item pool, describing how to choose appropriate statements.

[2] Mueller (1986), Spector (1990), and Roberts (1995) overview these methods.

then selected using a procedure either impicitly or explicitly based on (Pearson) correlations.

One popular method of item analysis[3] —implemented in SPSS's Reliability Analysis function—employs the total score to calculate two correlation-based statistics. One statistic is the *item-remainder* correlation, $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$, which is the Pearson correlation between (quantified) item j and the total score without item j. Items with the largest $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ values are included in the Likert scale, on the grounds that they best discriminate between people's attitudes as measured by the total score. The other statistic is *Cronbach's alpha* (1951), or more specifically Cronbach's alpha calculated from all items except j. Those items whose deletion leads to a larger value of $\alpha_{(-j)}$ are removed from the pool. In general, an effort is made to select items that not only meet this and the large $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ criteria, but that also balance pro- and anti-statements (so that acquiescence bias will effectively cancel out in responses to the resulting scale.)

The correlation-based methods discussed above were derived in the context of Classical Test Theory. They were not designed for data with the characteristics of Likert data, nor were they designed to reflect theories of attitudes and attitude formation. Further, they can only be used for items that are clearly against or in favour of the object.

As an alternative to the correlation-based methods, Roberts (1995, Chapter 19) describes an item-response-theory-based method of item analysis. In a first phase, principal components analysis is used to check dimensionality: In a heuristic approach based on the work of Davison (1977), those items that load highly on a third or higher component are removed. In a second phase, the GUM is fit to the remaining items and Wright and Masters' (1982) item-fit and person-fit statistics are calculated; the worst-fitting items and persons are removed in an iterative fashion. A scale is then formed by selecting remaining items whose estimated locations are evenly spaced along the evaluative continuum.

Roberts' procedure is reminiscent of Thurstone and Chave's (1929) method of equal-appearing intervals, which is used to develop another type of scale used for attitude measurement, the Thurstone scale. This type of scale contains items asking surveyees whether they 'Agree' or 'Disagree' with statements about the object; the median location of the agreeable statements for a surveyee is then used as an estimate

[3] Another correlation-based method uses either factor analysis or principal components analysis to perform item analysis, specifically to identify subscales of items within the initial pool. Spector (1990) discusses the use of explanatory and confirmatory factor analysis with Likert data.

of his attitude. Thurstone proposed three methods for creating Thurstone scales and estimating the locations of statements in them. All three collect data on the statements from a screening sample, but not by asking its members whether they agree with each statement. In the method of equal-appearing intervals, each member of the screening sample is asked to separate the statements into eleven ordered intervals based on their objective location along the evaluative continuum. It is assumed that the intervals are equal in width, so they are assigned integer scores (e.g., 1-11). For each statement, its median score is taken as an estimate of its location, and the IQR of its scores is taken as an estimate of the disagreement between judges over its location. Statements with the largest IQRs are removed from the pool, and then a scale is created by selecting items whose estimated locations span the evaluative continuum. These estimated locations are used when the resulting Thurstone scale is subsequently administered to surveyees.

In the following section, we introduce a new method of item analysis that shares the goals of the method of equal-appearing intervals: It seeks to select a group of statements that span the evaluative continuum and that do so in an order on which most people would agree. Like Thurstone and Chave's method, our method uses data (the screening sample data) on how the members of the screening sample rank the statements. However, in screening sample data, the rankings fan out from the person's own position on the evaluative continuum, rather than running from one end of the continuum to the other. Our method is similar to Roberts' approach, but involves the ULTMODR rather than the GUM and also refines some of his ideas.

5.2 A new method of item analysis

Our new method involves the simplest variant and the common threshold case of the ULTMODR; the model is fit to the data in the usual manner. The model is employed to calculate four statistics—(i) $\hat{\beta}_j$, (ii) $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$, (iii) $\hat{se}(\hat{\beta}_j)$, and (iv) χ^2_j, the *item j chi-square component* —that are then used to select items.

The $\hat{\beta}_j$ statistic reflects statement j's location along the evaluative continuum, making it analogous to the median in the method of equal-appearing intervals. More specifically, $\hat{\beta}_j$ is the estimated location for statement j when the model is fit to the entire screening sample. We will use the estimate to select items whose statements span the length of the evaluative continuum.

The $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$ statistic reflects the uncertainty in the estimated location for statement j, making it analogous to IQR in the method of equal-appearing intervals. However, the calculation of $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$ involves resampling. The model is repeatedly fit to a number of small (size n_s) sub-samples drawn (with replacement) from the screening sample data. In each sub-sample, the statements are ranked according to their estimated locations, and the statistic $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$ is then the standard deviation (across subsamples) of statement j's rank. A larger value of $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$ indicates that the statement's location differs depending on who is asked, and thus suggests that the item should be eliminated from consideration.

Like $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$, $\hat{se}(\hat{\beta}_j)$ reflects the uncertainty in the estimated location for statement j. The statistic is the square root of the estimated asymptotic variance of $\hat{\beta}_j$. The asymptotic variance can be estimated using the relevant diagonal element of the modified inverse of the observed information matrix (evaluated at the MLEs of the fixed-effects parameters). (The inverted matrix is modified in the manner described in Section 4.7.2; this modification is necessary because the likelihood is invariant with respect to changes of sign for all the β_js.) A larger value of $\hat{se}(\hat{\beta}_j)$ indicates a statement with a more uncertain location, suggesting that the item should possibly be eliminated from consideration.

The χ^2_j statistic can also be used to identify statements that should be eliminated from consideration. It is calculated using

$$\chi^2_j = \sum_k |\chi^2_{j(k)}|, \tag{5.1}$$

where $\chi^2_{j(k)}$ is the one-way signed Pearson residual discussed in Section 3.5. A large value of χ^2_j indicates that item j's data are not well explained by the model and should probably be eliminated.

The general strategy of the ULTMODR-based method is to select items with small $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$, $\hat{\sigma}^2_{rk(\hat{\beta}_j^{n_s})}$, and χ^2_j values and different $\hat{\beta}_j$s. However, after we perform item analysis on a simulated dataset in the following section, we will suggest several slight modifications to this strategy.

5.3 Application: Selecting items from a simulated dataset

The ULTMODR-based and Reliability Analysis methods were used to perform item analysis on a simulated dataset. The dataset was designed so that we could investigate

(i) how the item statistics used by the methods change with statement extremity and (ii) how the two methods perform at detecting confusing statements.

The dataset contains 140 persons and 46 statements, in order to make it comparable to the abortion attitude data analysed in the following section. The dataset was generated from a model that combined Case 1b of the ULTMODR's response structure with a latent structure similar to the simplest variant of the ULTMODR (section 3.2), except that some statements had two locations. Statements S-1 to S-36 were assumed to be interpreted identically by all 140 people and assigned evenly spaced locations between -4.6 and 3.6. However, statements S-37 to S-46 were assumed to be interpreted differently, depending on the person. More specifically, each statement has one location for (a random) 70 people, and a second location for the remaining 70 people. The locations for the statements are presented in Table 5.1. The values of the other parameters used to generate the simulated dataset were identical to those used in the simulation experiment (see Section 3.6).

Figure 5.1 presents two plots of the data, created using the existing and new MUA methods. In both plots, the straightforward statements (S-1 to S-36) are arranged around a horseshoe in an order that is roughly the same as their true order. Further, both plots make it apparent that there is something different about statements S-38, S-39, and S-42 to S-46 by locating them in the center of the straightforward statement horseshoe. (The existing method plot does the same for statement S-41). However, neither plot identifies statements S-37 or S-40 as different, not suprisingly since each is only slightly confusing.

For the ULTMODR-based method, the statistics $\hat{\beta}_j$, $\hat{se}(\hat{\beta}_j)$, $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ and χ^2_j were calculated using an R function written by the author. (The function is based on the $Q = 1$ version of the R function described in Section 4.7.2.) The left-hand side of Table 5.2 presents the values of the three statistics for each of the forty-six statements, ordered by $\hat{\beta}_j$. We note that $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ was calculated from twenty samples of size twenty.

Before the Reliability Analysis method could be used, the statements had to be divided into pro- and anti-object groups. We assigned the statements with a negative average location (i.e., S-1 to S-20 and S-40 to S-43) to the anti-object group, and the statements with a positive average location (i.e., S-21 to S-36 and S-37 to S-39) to the pro-object group. The statements with an average location of zero (i.e., S-42 to S-46) were excluded. We then applied SPSS's Reliability Analysis to the data, after reversing the scoring for the anti-object statements. Since statement S-20 had a

Table 5.1: Statement locations used to simulate dataset for item analysis

Straightforward			Confusing
β_1	=	-4.60	$\beta_{37} = \begin{cases} 0.00 \\ 1.00 \end{cases}$
β_2	=	-4.37	
β_3	=	-4.13	
β_4	=	-3.90	$\beta_{38} = \begin{cases} 0.00 \\ 2.00 \end{cases}$
β_5	=	-3.66	
β_6	=	-3.43	
β_7	=	-3.19	$\beta_{39} = \begin{cases} 0.00 \\ 3.00 \end{cases}$
β_8	=	-2.96	
β_9	=	-2.73	
β_{10}	=	-2.49	$\beta_{40} = \begin{cases} -1.00 \\ 0.00 \end{cases}$
β_{11}	=	-2.26	
β_{12}	=	-2.02	
β_{13}	=	-1.79	$\beta_{41} = \begin{cases} -2.00 \\ 0.00 \end{cases}$
β_{14}	=	-1.55	
β_{15}	=	-1.32	
β_{16}	=	-1.09	$\beta_{42} = \begin{cases} -3.00 \\ 0.00 \end{cases}$
β_{17}	=	-0.85	
β_{18}	=	-0.62	
β_{20}	=	-0.15	$\beta_{43} = \begin{cases} -4.00 \\ 0.00 \end{cases}$
β_{21}	=	0.09	
β_{22}	=	0.32	
β_{23}	=	0.55	$\beta_{44} = \begin{cases} -1.00 \\ 1.00 \end{cases}$
β_{24}	=	0.79	
β_{25}	=	1.02	
β_{26}	=	1.26	$\beta_{45} = \begin{cases} -2.00 \\ 2.00 \end{cases}$
β_{27}	=	1.49	
β_{28}	=	1.73	
β_{29}	=	1.96	$\beta_{46} = \begin{cases} -3.00 \\ 3.00 \end{cases}$
β_{30}	=	2.19	
β_{31}	=	2.43	
β_{32}	=	2.66	
β_{33}	=	2.90	
β_{34}	=	3.13	
β_{35}	=	3.37	
β_{36}	=	3.60	

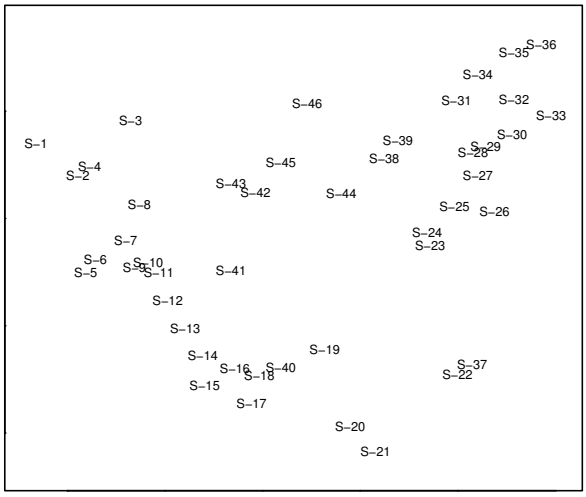


Figure 5.1: Plot of statement locations for the generated dataset. *The upper plot was produced using the SPSS implementation of ALSCAL. The lower plot was produced using R code written to implement the new MUA method introduced in Chapter 4.*

negative $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ value, it was reclassified as pro-object, and Reliability Analyis was performed a second time. The resulting $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ values are presented on the right-hand side of Table 5.2. The $\alpha_{(-j)}$ values are omitted because they are too uniform to be useful.

Table 5.2: Item analysis statistics for simulated dataset

ULTMODR-based method					Reliability Analysis	
Item	$\hat{\beta}_j$	$\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$	$\hat{se}(\hat{\beta}_j)$	χ^2_j	$\rho(Y_{ij}, \hat{\theta}_i^{s-j})$	Item
S-1	-3.67	0.25	0.12	4.24	0.482	S-1
S-2	-3.66	0.57	0.12	8.21	0.549	S-2
S-3	-3.39	0.41	0.11	7.4	0.336	S-3
S-4	-3.15	0.26	0.11	6.63	0.494	S-4
S-5	-2.92	0.56	0.10	2.03	0.532	S-5
S-6	-2.77	0.89	0.10	14.48	0.590	S-6
S-7	-2.71	0.85	0.10	4.91	0.524	S-7
S-8	-2.58	0.75	0.10	2.6	0.509	S-8
†S-46	-2.4	266.37	0.10	9.21	NA[a]	S-46†
S-9	-2.4	0.54	0.10	9.92	0.554	S-9
S-10	-2.26	0.80	0.09	5.06	0.506	S-10
†S-43	-2.05	51.42	0.11	20.26	0.181	S-43†
S-11	-1.98	0.87	0.09	2.45	0.615	S-11
†S-45	-1.84	106.24	0.10	9.44	NA[a]	S-45†
S-12	-1.83	1.20	0.09	1.49	0.619	S-12
S-13	-1.55	1.06	0.09	7.15	0.622	S-13
†S-42	-1.51	28.24	0.13	1.92	0.215	S-42†
S-14	-1.33	2.06	0.10	9.4	0.453	S-14
S-15	-1.32	1.21	0.09	6.3	0.592	S-15
†S-41	-1.14	4.54	0.10	4.59	0.281	S-41†
S-16	-1.03	3.20	0.11	9.56	0.397	S-16
S-17	-0.9	3.42	0.09	11.43	0.422	S-17
S-18	-0.76	1.89	0.10	8.47	0.205	S-18
†S-40	-0.44	3.82	0.10	7.85	0.151	S-40†
S-19	-0.39	2.87	0.10	10.55	0.023	S-19
S-20	-0.2	2.45	0.09	12.82	0.080	S-20
S-21	-0.07	4.16	0.10	7.85	0.1439	S-21
S-22	0.2	2.88	0.08	9.15	0.445	S-22
S-23	0.35	2.77	0.10	4.62	0.418	S-23

continued on following page

a Item excluded from Reliability Analysis.

† Denotes confusing statement

Table 5.2: Item analysis statistics for simulated dataset (continued)

ULTMODR-based method					Reliability Analysis	
Item	$\hat{\beta}_j$	$\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$	$\hat{se}(\hat{\beta}_j)$	χ^2_j	$\rho(Y_{ij}, \hat{\theta}_i^{s-j})$	Item
continued from previous page						
†S-37	0.35	3.22	0.10	10.35	0.486	S-37†
S-24	0.43	2.87	0.10	5.49	0.423	S-24
†S-44	0.45	7.31	0.11	1.1	NA[a]	S-44†
S-25	0.77	1.25	0.10	3.58	0.584	S-25
†S-38	0.97	9.78	0.14	5.09	0.354	S-38†
S-26	1.00	0.78	0.10	3.74	0.632	S-26
S-27	1.11	2.20	0.10	3.18	0.648	S-27
S-28	1.37	1.78	0.10	1.26	0.605	S-28
S-29	1.38	1.15	0.11	3.72	0.550	S-29
†S-39	1.5	51.89	0.11	1.57	0.309	S-39†
S-30	1.7	0.68	0.10	2.1	0.596	S-30
S-31	1.9	1.00	0.10	11.54	0.508	S-31
S-32	2.06	0.96	0.11	4.05	0.534	S-32
S-33	2.17	0.45	0.10	13.3	0.622	S-33
S-34	2.36	0.41	0.11	5.96	0.464	S-34
S-35	2.55	0.26	0.11	13.6	0.527	S-35
S-36	2.71	0.13	0.11	6.74	0.511	S-36

a Item excluded from Reliability Analysis.
† Denotes confusing statement

We first comment on the $\hat{\beta}_j$s. These values order the straightforward statements (S-1 to S-36) in accordance with their true ordering, as we would expect from the simulation experiment described in Section 3.6. Also as we might expect, many of the confusing statements (i.e., statement S-37 to S-42) are located near the straightforward statements with true locations most similar to their average true locations. However, other confusing statements (i.e., S-43 to S-46, which have one anti-object and one pro-object interpretation and an average true location of zero) are not located near the straightforward statements with true locations of zero (e.g., S-20 and S-21). In fact, when fitting the simplest variant of the ULTMODR, we experienced problems with local maxima because of these items: The solution found depended on whether their starting values were positive or negative. (Table 5.2 presents the β_js from the solution with the largest $\ln(ML)$ value.) In none of these solutions were statements S-43 to S-46 located near S-20 or S-21; instead, they were always located either amidst the

anti-object statements or amidst the pro-object statements. Thus, it seems that the criterion used to assess the ULTMODR's fit penalizes grossly misfitting some people's data more than moderatedly misfitting everyone's data.

Turning to the other ULTMODR-based statistics, we see that $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ and $\hat{se}(\hat{\beta}_j)$, vary with statement extremity, but χ^2_j does not. The statistic $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ increases as statements become less extreme. This phenomenon occurs because central statements have more statements located within a certain distance of them than do extreme statements; thus, it is easier for a centrally located statement to change position in the ranking. On the other hand, $\hat{se}(\hat{\beta})$ increases slightly as statements become more extreme. As noted in Chapter 4, this phenomenon occurs because most people find extreme statements strongly disagreeable, meaning that their locations can be made more extreme without significantly altering the model's fit. These results suggest that $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ and $\hat{se}(\hat{\beta}_j)$ should be compared only between statements with similar locations, but that χ^2_j can be compared between any statements.

The Reliability Analysis statistic $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ also varies with statement extremity. More specifically, it has a larger value for extreme statements and its most optimal value for statements that are moderately extreme, peaking for statements with estimated locations of ± 1.5. The three anti-object statements with optimal $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ have adjacent estimated locations, as do two of the three pro-object statements with optimal $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$. These results suggest that the Reliability Analysis method will produce scales containing very similar pro-object statements and very similar anti-object statements, with both groups of statements moderately extreme.

The ULTMODR-based method and the Reliability Analysis method perform similarly well at detecting confusing statements. In the Reliability Analysis method, $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ is very small for each confusing statement except S-37, which is only slightly confusing. In fact, the confusing statements would be among the last selected for inclusion in a scale. In the ULTMODR-based method, the statistic $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ identifies all of the confusing statements. For each one, $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ suggests that a neigbouring (non-confusing) statement is preferable. Further, for the more confusing of the statements (i.e., S-38, S-49, S-42, S-44, S-45 and S-46), $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ is noticeably larger not just in relative terms, but also in absolute terms. As for the statistic $\hat{se}(\hat{\beta}_j)$, comparing its values locally identifies some but not all of the confusing statements as unusual. Similarly, global comparisons of the χ^2_j values identify a few of the confusing statements; For instance, the statistic takes its largest value for confusing statement S-43.

The results of this application suggest that the best way to identify confusing statements is by making local comparisons of $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$. In addition, global comparisons of the χ^2_j statistic can also be useful for identifying the most confusing statements.

5.4 Application: Selecting items for an abortion attitude scale

The ULTMODR-based method and the Reliability Analysis method were used to perform item analysis on the Likert items in the abortion attitude dataset. The goal was to to create a six item scale for measuring abortion attitudes.

As before (see Section 4.5)), we removed four items (14, 15, 19, and 21) that had more than 10% 'Don't Know/Can't Choose' responses and, for the remaining 46 items, recoded any 'Don't Know/Can't Choose' responses as 'Neither agree nor disagree.'

First, the ULTMODR-based statistics $\hat{\beta}_j$, $\hat{se}(\hat{\beta}_j)$, $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ and χ^2_j were calculated using the aforementioned R function written by the author. Here, $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ was calculated from ten samples of size twenty. The values for three of these ULTMODR-based statistics are presented in the left-hand side of Table 5.3 for each of the remaining 46 items, ordered by $\hat{\beta}_j$. The values of $\hat{se}(\hat{\beta}_j)$ are omitted because they are too uniform to be helpful. The left-hand side of the table also includes the proportion of 'Don't Know/Can't Choose' responses that were present before the data was recoded; obviously, we prefer items with fewer 'Don't Know/Can't Choose' responses.

Beginning with the $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ column, we see that the statements with the most uncertain rank are: S-44 ('I believe that abortion is generally wrong, but I think that it is necessary for it to be legal in today's society.'); S-30 ('I personally have not resolved how I feel about abortion.'); S-34 ('Abortion should generally be a woman's prerogative, but it should not be permitted in every case.'); S-33 ('I cannot wholeheartedly support either side of the abortion debate.'); S-39 ('Abortion should be legal under any circumstances.'); and S-41 ('Restrictions should never be placed on a woman's right to an abortion.'). These statements have $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ values that are much larger than other statements with similar locations. It is not hard to see how each might be interpreted differently by different people.

Turning to the χ^2_j column, we see that the worst-fitting statements are: S-25 ('Abortion, in general, should be legal, but should never be used as a conventional method of birth control.'); S-31 ('If abortion were not legal, (illegal) abortions would still be

performed.'); and S-29 ('I find myself agreeing with arguments both for and against abortion.'). We recall that S-25 and S-31 were tagged as unusual in the new MUA method plot (see Figure 4.8). We also note that statement S-25 contains two ideas, making it unsurprising that the statement is so poorly fit by the simplest ULTMODR variant. ("Double-barreled" statements like this one tend to confuse respondents because they aren't certain which idea to respond to.) In general, the statements expressing moderate views tend to have much larger χ_j^2 values. This is true, not surprisingly, for those statements that express neutrality through the use of two ideas (e.g., S-23, S-24, S-32, S-34, S-44). It is also true, however, for those statements that express neutrality directly (e.g., S-30).

We used the ULTMODR-based statistics to select six items for inclusion in a scale. Overall, we focused on selecting statements with varying $\hat{\beta}_j$ values and differing statement content. We used $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ to choose between statements with comparable locations, and also avoided statements with large χ_j^2 values. The resulting scale can be seen in Figure 5.4.

Second, SPSS's Reliability Analysis was run on the items that were clearly in favour of or against abortion. We identified these 31 (15 pro- and 16 anti-) statements using Figure 4.8, which we recall had a pro-abortion item cluster and an anti-abortion item cluster.[4] The $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ values from Reliability Analysis are presented on the right-hand side of Table 5.3; the $\alpha_{(-j)}$ values are omitted since they are too uniform to be helpful.

The $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ statistic suggests that statements S-47, S-12, and S-37 should be the first ones eliminated from consideration. Interestingly, in the existing MUA method plot of the abortion attitude items (see Figure 4.7), these three statements are pulled more towards the center of the statement horseshoe than the other 27 statements subjected to Reliability Analysis.

We formed a Reliability Analysis scale by selecting the three anti-abortion with the largest $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ values, and the three pro-abortion items with the largest $\rho(Y_{ij}, \hat{\theta}_i^{s-j})$ values. Note that the resulting scale (see Table 5.5) contains statements expressing two types of views: abortion as immoral, and abortion as a woman's right. The ULTMODR-based scale, on the other hand, is explicitly designed to include different types of pro- and anti-abortion views.

[4] We did not heed the warning of Spector (1990, p. 34), who discourages empirical determination of statement direction and encourages researchers to classify items a priori based on statement content.

Table 5.3: Item analysis statistics for abortion attitude items

	ULTMODR-based method					Reliability Analysis	
	Item[a]	$\hat{\beta}_j$	% 'DK/CC'	$\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$	χ^2_j	$\rho(Y_{ij}, \hat{\theta}_i^{s-j})$	Item[b]
Anti-	S-5	-4.62	0.03	0.00	20.7	0.50	S-5
	*S-2	-3.79	0.00	0.47	1.7	0.82	S-2
	S-1	-3.63	0.01	1.92	12.3	0.67	S-1
	S-11	-3.55	0.03	5.67	21.8	0.74	S-11
	S-20	-3.51	0.01	7.21	12.3	NA	S-20
	S-18	-3.49	0.01	5.15	11.4	0.74	S-18
	S-4	-3.46	0.06	3.92	7.3	0.79	S-4
	S-9	-3.38	0.04	4.43	7.1	0.68	S-9
	S-6	-3.29	0.08	4.03	3.6	0.88	S-6*
	*S-13	-3.26	0.04	2.67	4.9	0.71	S-13
	S-17	-3.16	0.04	2.68	2.1	0.75	S-17
	S-3	-3.16	0.07	3.08	2.4	0.83	S-3*
	S-8	-3.08	0.03	3.00	2.2	0.83	S-8
	S-12	-2.84	0.01	7.25	6.3	0.54	S-12
	*S-16	-2.84	0.03	2.01	1.0	0.84	S-16*
	S-10	-2.82	0.02	1.75	5.9	0.74	S-10
	S-7	-2.77	0.04	2.30	5.3	0.80	S-7
Mod.	S-30	-2.73	0.04	76.00	20.7	NA	S-30
	S-33	-2.44	0.06	35.63	19.2	NA	S-33
	S-28	-2.29	0.04	0.74	19.9	NA	S-28
	S-32	-2.22	0.02	1.12	24.4	NA	S-32
	S-44	-2.12	0.04	77.46	23.1	NA	S-44
	S-29	-2.01	0.05	0.68	35.0	NA	S-29
	S-23	-1.89	0.00	0.63	20.1	NA	S-23
	S-34	-1.64	0.01	37.82	20.9	NA	S-34
	S-27	-1.27	0.01	1.62	22.2	NA	S-27
	S-22	-1.24	0.01	4.06	8.0	NA	S-22
	S-26	-1.20	0.06	1.67	15.2	NA	S-26
	S-24	-0.91	0.01	1.63	27.6	NA	S-24
	S-25	0.05	0.06	1.29	94.0	NA	S-25
	S-31	0.20	0.04	1.67	47.0	NA	S-31
Pro-	S-43	0.54	0.05	1.27	27.2	0.54	S-43
	S-35	0.94	0.05	1.50	21.4	0.68	S-35
	*S-45	0.98	0.04	1.09	6.5	0.64	S-45

continued on following page

[a] Asterisk indicates that item was selected using the ULTMODR-based method.

[b] Asterisk indicates that item was selected using the Reliability Analysis method.

Table 5.3: Item analysis statistics for abortion attitude items

	ULTMODR-based method					Reliability Analysis	
	Item[a]	$\hat{\beta}_j$	% 'DK/CC'	$\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$	χ^2_j	$\rho(Y_{ij}, \hat{\theta}_i^{s-j})$	Item[b]
	continued from previous page						
Pro-	S-36	1.37	0.04	0.95	20.6	0.80	S-36*
	S-42	1.52	0.04	1.31	12.6	0.81	S-42*
	S-38	1.53	0.04	0.96	10.5	0.73	S-38
	S-49	1.77	0.07	1.88	14.7	0.82	S-49*
	S-46	1.77	0.05	1.57	21.0	0.78	S-46
	S-48	2.04	0.06	1.31	23.3	0.75	S-48
	S-40	2.07	0.03	1.40	6.9	0.69	S-40
	S-50	2.26	0.08	0.59	6.2	0.71	S-50
	*S-37	2.40	0.03	1.16	6.2	0.59	S-37
	S-41	2.70	0.04	20.77	12.5	0.68	S-41
	S-39	2.74	0.04	27.99	2.1	0.68	S-39
	*S-47	3.57	0.06	0.00	4.4	0.38	S-47
Excl.	S-14	NA	0.21	NA	NA	NA	S-14
	S-15	NA	0.11	NA	NA	NA	S-15
	S-19	NA	0.11	NA	NA	NA	S-19
	S-21	NA	0.44	NA	NA	NA	S-21

[a] Asterisk indicates that item was selected using the ULTMODR-based method.

[b] Asterisk indicates that item was selected using the Reliability Analysis method.

Table 5.4: Abortion attitude scale formed using ULTMODR-based method

Statement 1 (S-2):	'Abortion is a threat to our society.'
Statement 2 (S-13):	'Having an abortion is far worse than having an unwanted child.'
Statement 3 (S-16):	'Even if one believes that there may be some exceptions, abortion is still generally wrong.'
Statement 4 (S-45):	'If abortion became illegal, there would be negative consequences for society.'
Statement 5 (S-37):	'Only the woman who is pregnant can decide whether an abortion is warranted.'
Statement 6 (S-47):	'Abortion should be a socially acceptable method of birth control.'

Table 5.5: Abortion attitude scale formed using Reliability Analysis method

Statement 1 (S-6):	'Abortion is immoral.'
Statement 2 (S-3):	'Abortion is inhumane.'
Statement 3 (S-16):	'Even if one believes that there may be some exceptions, abortion is still generally wrong.'
Statement 4 (S-36):	'A woman should have control over what is happening to her own body by having the option to choose abortion.'
Statement 5 (S-42):	'Outlawing abortion violates a woman's civil rights.'
Statement 6 (S-49):	'Abortion is a reasonable alternative if a woman feels that having a baby might ruin her life.'

5.5 Conclusions

We have introduced a new ULTMODR-based method of item analysis. This method is an alternative to popular correlation-based methods of item analysis, such as SPSS's Reliability Analysis.

The results of the abortion attitude application suggest that the ULTMODR-based method, when used on an actual dataset, tends to remove most of the moderate statements *a posteriori*. The Reliability Analysis method, on the other hand, removes moderate statements from consideration *a priori.* However, even though both exclude moderate statements, the ULTMODR-based and Reliability Analysis method produce very different scales.

Judging from the results of our two applications, the Reliability Analysis method

produces scales with statements that cluster in two locations along the evaluative continuum, one moderately anti-object and the other moderately pro-object. The similarity of pro- and of anti- statements is particularly interesting when we note that Reliability Analysis scales are typically analysed using Likert's measurement model, which treats all statements with the same orientation identically.

In contrast, the ULTMODR-based method is explicitly designed to select statements whose locations differ. The resulting scale is therefore more in line with the aims of early attitudinal researchers such as Thurstone and Chave.

Chapter 6

Fitting structural models

Researchers, though occasionally interested in attitudes for their own sake, typically want to investigate how they are affected by certain background and behavioural *covariates*. As an example, we might use the NIS dataset to study how (British or American) nationality affects national pride. Here, we will use $\mathbf{X}$ to denote the covariates; further, we will refer to the model that specifies how $\mathbf{X}$ affects θ as the *structural model*.

Fitting a structural model to Likert data involves combining it with a measurement model, in a one-stage or two-stage procedure. In the former procedure, both models are fit to the data simultaneously. In the latter procedure, people's attitudes are estimated by fitting the measurement model to the data, and the estimated attitudes are then used to fit the structural model. For example, social scientists commonly use Likert's measurement model to estimate θ_i^s in a first stage, and then use the resulting $\hat{\theta}_i^s$s to fit a structural model in a second stage. (We will refer to this method as the *scoring method*.) There is some debate over which type of procedure is preferable,[1] but we prefer one-stage procedures since they model all possible effects on the data at once and result in less attenuated estimates.

Of course, the covariates may affect Likert responses not just through attitudes, but also through response category interpretation. In our example, differences in the American and British responses should not necessarily be attributed solely to differences in national pride. They may also stem from noted differences in the way that surveyees from the two nations interpret the response categories. Unfortunately, current structural-model-fitting methods (e.g., the scoring method) do not allow response

[1] Bartholomew and Knott (1999, Sections 8.12 and 8.13) discuss why a two-stage procedure might be preferable; Zwinderman (1997, p. 245) discusses why a one-stage procedure might be preferable.

category interpretation to differ. Adjusting for differing response category interpretation is a particular concern in our example because the national pride scale is unbalanced: Since four of the national pride scale's five statements are pro-country, more agreeable responses will result for a more acquiescent nation even if it does not feel more national pride.

In this chapter, we introduce a one-stage structural-model-fitting method that allows response category interpretation to differ. The method embeds the structural model in a *multi-set variant* of the ULTMODR. The multi-set variant has a latent structure that can incorporate items from multiple sets (if available); incorporating additional items can provide more information on how response category interpretation differs.

After introducing the new method, we conduct simulation experiments designed to assess the performance of the scoring method, including how it compares to our new method. Then, in an extended example, we apply the new method to the NIS dataset in order to investigate how nationality affects national pride, controlling for national differences in response category interpretation. Last, in a shorter example, we apply the new method to the abortion attitude items in order to investigate how nationality, gender, and religious status affect abortion attitudes, controlling for national and gender differences in response category interpretation

6.1 A new method of fitting structural models

We are primarily interested in how the covariates affect the attitude underlying one *primary* set of items. However, the variant of the ULTMODR employed in our structural-model-fitting method can incorporate items from additional *secondary* sets. We noted in Chapter 3 that the ULTMODR's ability to separate the effects of attitudes and response category interpretation depends on the data. For one, separating the effects is obviously more difficult when the number of items is small, as is the case in our national pride example. In addition, separating the effects is particularly difficult if the items are homogeneous. For example, in the national pride scale, because 80% of the statements are pro-country, it is nearly impossible to determine whether a person with more agreeable responses is more acquiescent, has more national pride, or both. Thankfully, the immigration items in the NIS dataset contain additional information

on how each person interprets the response categories. This is particularly true because the items express a wide variety of views and are balanced. Thus, incorporating the immigration items into our measurement model can help us separate the effects of acquiescence and national pride on responses to the national pride items. However, doing so requires making the assumption that response category interpretation does not differ for items belonging to different sets.

We now address the latent structure of the multi-set variant. It is an S-dimensional space where dimension s is the (one) dimension that underlies I_s and where the axes are not necessarily orthogonal to each other. Without loss of generality, the first dimension will be the one corresponding to the primary set. Both the persons and the statements are located in this space; however, each statement's location is restricted to the axis for the set to which it belongs. More formally, $\theta_i = [\ \theta_i^1 \quad \theta_i^2 \quad \dots \quad \theta_i^S\ ]^T$, and $\beta_j = [\ \beta_j^1 \quad \beta_j^2 \quad \dots \quad \beta_j^S\ ]^T$, where $\beta_j^s = 0$ if $S(j) \neq s$. The (latent) distance, d_{ij}, is the l_1-distance between statement j's location and the projection of person i's location onto dimension $S(j)$:

$$d_{ij} = \left|\theta_i^{S(j)} - \beta_j^{S(j)}\right|. \tag{6.1}$$

Note that d_{ij} is not *directly* affected by person i's location along dimensions other than $S(j)$. (Of course, the other θ_i^ss might affect d_{ij} indirectly if the S elements of θ_i are correlated.) Note also that, if no secondary sets are available (i.e., if $S = 1$), then the multi-set ULTMODR variant reduces to the simplest ULTMODR variant introduced in Chapter 3.

A structural model is embedded in the multi-set variant's latent structure. We assume that the structural model is a general linear model of the form

$$\theta_i = \gamma^T X_i + \delta_i \text{ for } i = 1, 2, \dots, n, \tag{6.2}$$

where X_i is a length P vector containing the covariates for person i; where γ is a $P \times S$ matrix containing parameters that quantify how changes in the covariates affect the mean of the person location(s); and where δ_i is an error vector of length S. We will assume that δ_i has a normal distribution with a mean vector that is $\mathbf{0}$ and a covariance matrix, $\mathbf{\Phi}$, that is either equal to $\mathbf{I}$ or modelled as a function of the covariates (in which case it is referred to as $\mathbf{\Phi}(X_i)$). Note that we are primarily interested in the elements of γ (and $\mathbf{\Phi}(X_i)$, if relevant) that correspond to θ_i^1 since these parameters quantify the relationship between the primary attitude and the covariates.

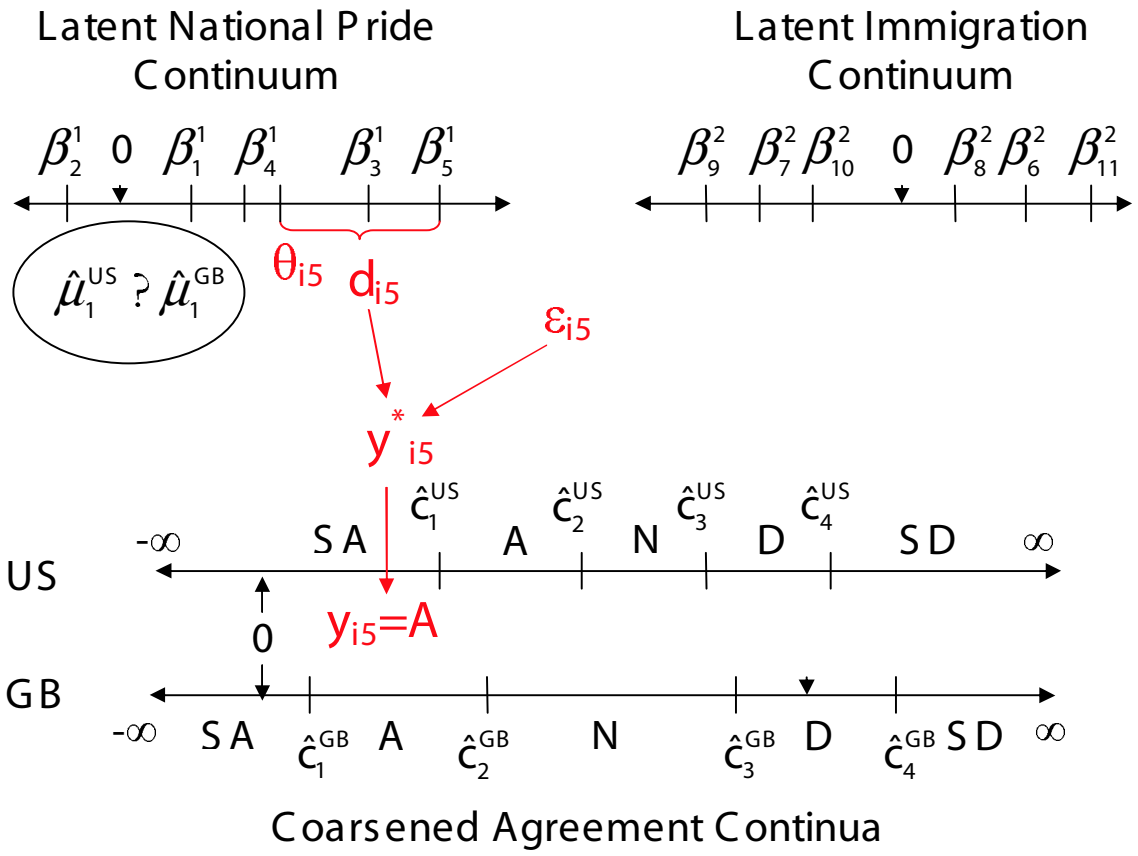


Figure 6.1: Case 2a and the multi-set ULTMODR variant applied to the NIS dataset. *Note that the response category thresholds differ by nation, but that they are the same for items from the national pride scale and the immigration scale. The red graphics illustrate how the model works for an American person responding to the fifth statement.*

Any case of the response structure can be used with this variant of the ULTMODR. As noted above, the same response structure will be assumed for all items, regardless of the set to which they belong. Figure 6.1 depicts how a group-specific response structure (Case 2a) might be combined with the multi-set ULTMODR variant for the NIS dataset.

The multi-set variant of the ULTMODR and the embedded structural model are fitted simultaneously using the approach outlined in Section 3.4. The hyperparameters of $g_1(\theta_i)$ are now determined by Equation (6.2). More specifically, its mean vector equals $\gamma^T \mathbf{X}_i$, and its covariance matrix is either $\mathbf{I}$ or $\mathbf{\Phi}(X_i)$. If we want to estimate these hyperparameters, we will need to fix the statement locations to pre-specified

values. Fixing the β_js for set s establishes the scale and orientation of the metric measured by θ_i^s, and therefore identifies column s of γ in terms of magnitude and sign.

Since the combined models are fit by maximising an appropriate likelihood, estimates of asymptotic standard errors for their parameters can be obtained by inverting the observed information matrix evaluated at the MLEs. In addition, we can use Likelihood Ratio Tests to compare competing hypotheses about the models' parameters.

6.2 Simulation experiments

We conducted two sets of experiments designed to investigate the performance of the oft-used scoring method and to see how it compared to our new method.

In these experiments, datasets were generated from the simplest ULTMODR variant with a person-specific response structure (Case 1). We assumed that the persons belonged to two groups, which we refer to as 1 and 2; these groups can be thought of as G.B. and U.S., say. We were interested in testing the hypothesis $\mu^1 = \mu^2$, where $\mu = \mathrm{E}(\theta_i^1)$. In particular, we wanted to see whether various approaches to testing this hypothesis were robust to other types of group differences (aside from differences in mean attitude).

In every experiment, most of the hyperparameters of $g_2(\tau^i, \ln(\sigma^i))$ and $g_1(\theta_i^1)$ were identical for the two groups. In variation A, all hyperparameters were the same; in variation B, only the variance of θ_i^1 differed by group; in variation C, only the mean of τ^i differed by group; in variation D, only the mean of $\ln(\sigma^i)$ differed by group; and, in variation E, only the covariance matrix of τ^i and $\ln(\sigma^i)$ differed by group. (Note that the mean of θ_i^1 was, of course, always the same for both groups.) The values used for the hyperparameters (see Table 6.1) were based on the results of the NIS application in Section 6.3, as were the values used for the thresholds ($c_1 = 1.39$, $c_2 = 2.89$, $c_3 = 3.88$, and $c_4 = 5.54$).

The first set of simulation experiments were designed to test the performance of the scoring method and to see how its performance differed for three different scales. The first scale was created to resemble the five-statement national pride scale: It used the same statement locations as the national pride scale (see the top half of Figure 6.3), one of which was negative and four of which were positive. The second scale was created to resemble the six statement immigration scale: It used the same statement locations as the immigration scale (see the bottom half of Figure 6.3), three of which were

Table 6.1: Hyperparameter values used in the simulation experiments

Parameter		Variation A		Variation B		Variation C		Variation D		Variation E	
		Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2	Grp 1	Grp 2
$g_1(\mu, \mathbf{\Phi})$	μ_1	–0.13		–0.13		–0.13		–0.13		–0.13	
hyperparameters	ϕ_1	0.56		0.66	0.47	0.56		0.56		0.56	
$g_2(\varphi, \mathbf{\Lambda})$ hyperparameters	φ_1	0.37		0.37		0.00	0.74	0.37		0.37	
	φ_2	0.005		0.005		0.005		–0.09	0.10	0.005	
	$\lambda_{1,1}$	1.07		1.07		1.07		1.07		1.33	0.81
	$\lambda_{1,2}$	0.29		0.29		0.29		0.29		0.34	0.25
	$\lambda_{2,2}$	0.09		0.09		0.09		0.09		0.09	0.08

Table 6.2: Sorted p-values from three hypothesis testing approaches

Variation	Method	P-values									
	Scoring	0.03	0.17	0.17	0.23	0.44	0.53	0.77	0.81	0.84	0.85
1	Case 3	0.13	0.31	0.42	0.45	0.63	0.67	0.81	0.83	0.88	0.94
	Case 2a	0.13	0.4	0.43	0.55	0.66	0.75	0.76	0.84	0.85	0.91
	Scoring	0.05	0.21	0.25	0.29	0.63	0.66	0.7	0.71	0.89	0.96
2	Case 3	0.16	0.36	0.49	0.55	0.69	0.75	0.75	0.89	0.92	0.98
	Case 2a	0.19	0.43	0.51	0.56	0.59	0.73	0.75	0.82	0.93	0.97
	Scoring	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.06	0.07	0.37
3	Case 3	0.00	0.01	0.01	0.02	0.03	0.05	0.06	0.11	0.18	0.47
	Case 2a	0.17	0.41	0.53	0.62	0.79	0.83	0.84	0.88	0.91	0.95
	Scoring	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.03	0.04	0.04
4	Case 3	0.00	0.00	0.02	0.04	0.05	0.07	0.08	0.12	0.13	0.15
	Case 2a	0.22	0.43	0.52	0.57	0.61	0.75	0.89	0.91	0.92	0.94
	Scoring	0.01	0.15	0.22	0.25	0.52	0.61	0.74	0.76	0.77	0.91
5	Case 3	0.08	0.36	0.37	0.49	0.65	0.67	0.75	0.81	0.86	0.94
	Case 2a	0.11	0.39	0.47	0.56	0.64	0.64	0.73	0.83	0.85	0.94

negative and three of which were positive. The third scale contained six statements: Three had locations equal to -5 and three had locations equal to 5. This scale was created, especially when the primary set is small and unbalanced. Of course, in doing so, we will be making the assumption that response category interpretation is the same for these additional sets.

6.3 Application: Investigating national pride

We were interested in using the NIS dataset to investigate whether Americans or the British exhibit more national pride, while adjusting for national differences in response category interpretation.

First, we used the scoring method to make some preliminary comparisons of the American and British NIS data, though we were mindful of the method's limitations.[2] The national pride total score, which we denote $\widehat{\theta}_i^1$, was calculated using category

[2]Smith and Jarkko (2001) also use the scoring method to compare mean levels of general national pride for respondents from the different nations included in the NIS survey. Their comparison produced a ranking (in terms of decreasing national pride) with the U.S. in second place behind Austria, and Great Britain in fourteenth place.

scores 'AS'$= -2$, 'A'$= -1$, 'N'$= 0$, 'D'$= +1$, 'DS'$= +2$ for anti-country item 2, and category scores 'AS'$= +2$, 'A'$= +1$, 'N'$= 0$, 'D'$= -1$, 'DS'$= -2$ for pro-country items S-1, S-3, S-4, and S-5. Therefore, $\widehat{\theta}_i^1$ can range between -10 (flag burning) and 10 (flag waving). The distribution, by nation, of $\widehat{\theta}_i^1$ can be seen in Figure 6.2, which also contains frequency plots of the immigration total scores, or $\widehat{\theta}_i^2$s.[3] The means of the $\widehat{\theta}_i^1$ are 2.08 and 0.34 for the American and British surveyees, respectively, suggesting that Americans have more national pride. (The p-value for the equality of means is less than $2 \cdot 10^{-16}$ according to two-sample t-tests with equal and unequal variances and the Wilcoxon rank sum test.) The sample variances are 9.38 and 10.60 for the American and British surveyees, respectively, suggesting that Americans have more uniform national pride attitudes. In addition, we calculated Pearson correlations between $\widehat{\theta}_i^1$ and $\widehat{\theta}_i^2$, which were 0.33 and 0.41 for the American and British surveyees, respectively. It makes sense that, for individuals, a higher level of national pride would accompany a more negative attitude towards immigration.[4]

Second, we applied our new structural-model-fitting method to the NIS data. We recall that the method requires fixing the eleven statement locations to pre-specified values. We chose to estimate these values and did so using the multi-set variant of the ULTMODR on its own (with no structural model). In order to resolve the additive confouding issue in the latent space, we set the mean and variance of θ_i to $\mathbf{0}$ and $\mathbf{1}$, respectively. We first fit the common threshold case (Case 3) to the British data and American data separately; we did so to check that the order of the $\hat{\beta}_j$s was the same for both groups of surveyees.[5] It was, so we then fit the group-specific threshold case (Case 2a) to the pooled data.Figure 6.3 presents the resulting statement location estimates; we adopt these as the pre-specified values for the statement locations. Note that, reassuringly, the estimates imply an ordering of the statements that is consistent with their content.

[3]The immigration total score, $\widehat{\theta}_i^2$, was calculated using category scores 'AS'$= -2$, 'A'$= -1$, 'N'$= 0$, 'D'$= +1$, 'DS'$= +2$ for pro- immigration statements and category scores 'AS'$= +2$, 'A'$= +1$, 'N'$= 0$, 'D'$= -1$, 'DS'$= -2$ for anti-immigration statements. Thus, $\widehat{\theta}_i^2$ can range between -12 (pro-immigration) and 12 (anti-immigration).

[4]Note that this relationship does not hold at the national level: Americans exhibit more national pride and more pro-immmigration attitudes than the British. Though this might seem contradictory *prima facie*, we recall that "unlike most nation states which were built up around a primordial tribe, the US is based on a set of shared ideals . . . [which] allows American pride to be not only particularistic, but also universal." (Smith and Jarkko, 2001).

[5]Recall the simulation experiment performed in Section 3.6, which demonstrated that the true item ordering could be recovered without modelling response category interpretation at a person-specific level.

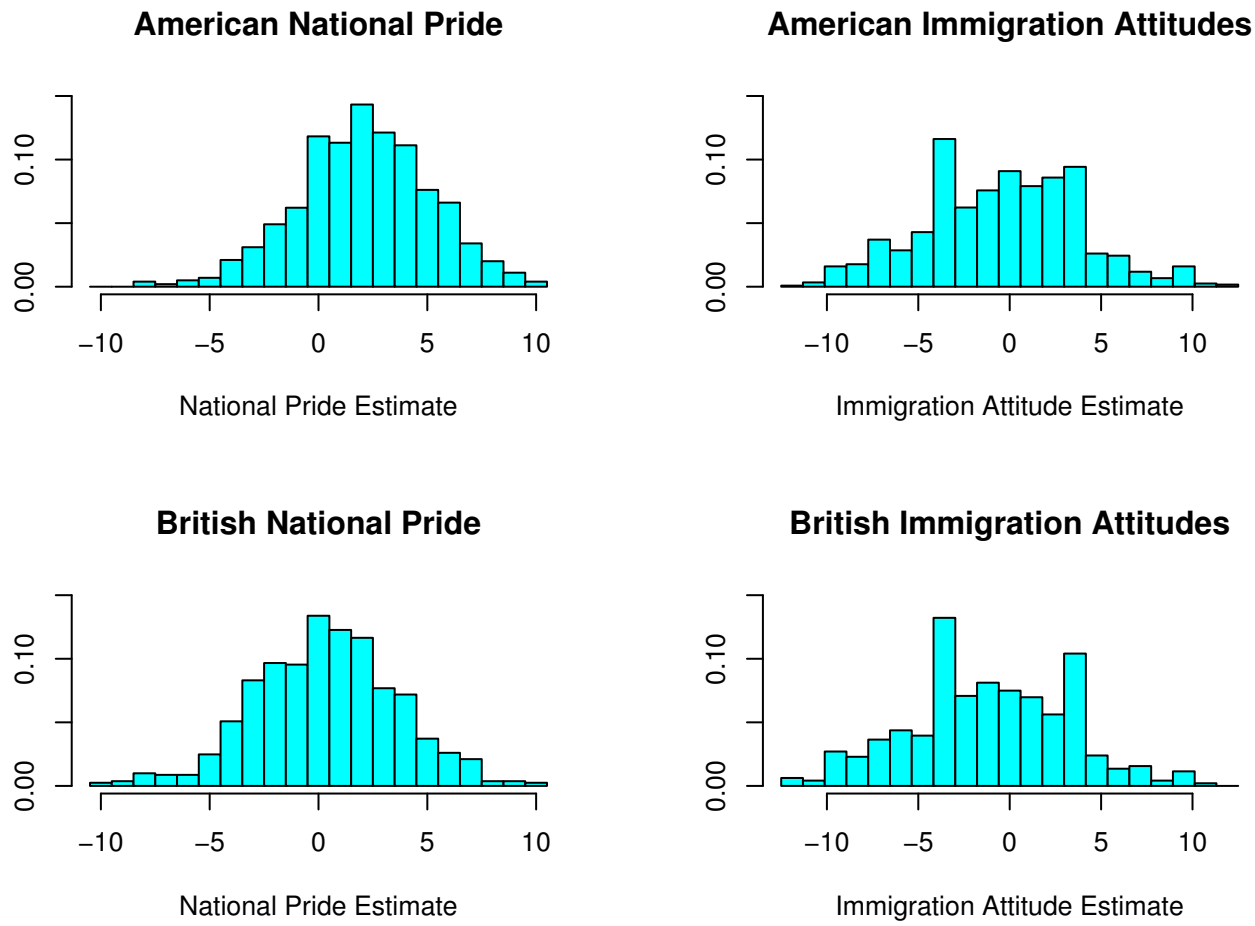


Figure 6.2: Total Scores, by Nation, for National Pride and Immigration Sets. *The left-hand frequency plots reveal that the distribution of national pride total scores has roughly the same spread and a normal shape for both nations, but appears to be shifted more to the left (towards flag-burning) for the British. Similarly, the right-hand frequency plots reveal that the distribution of immmigration total scores has roughly the same spread and quasi-normal shape for both nations, but also seems to be shifted more to the right (towards anti-immigration) for the British.*

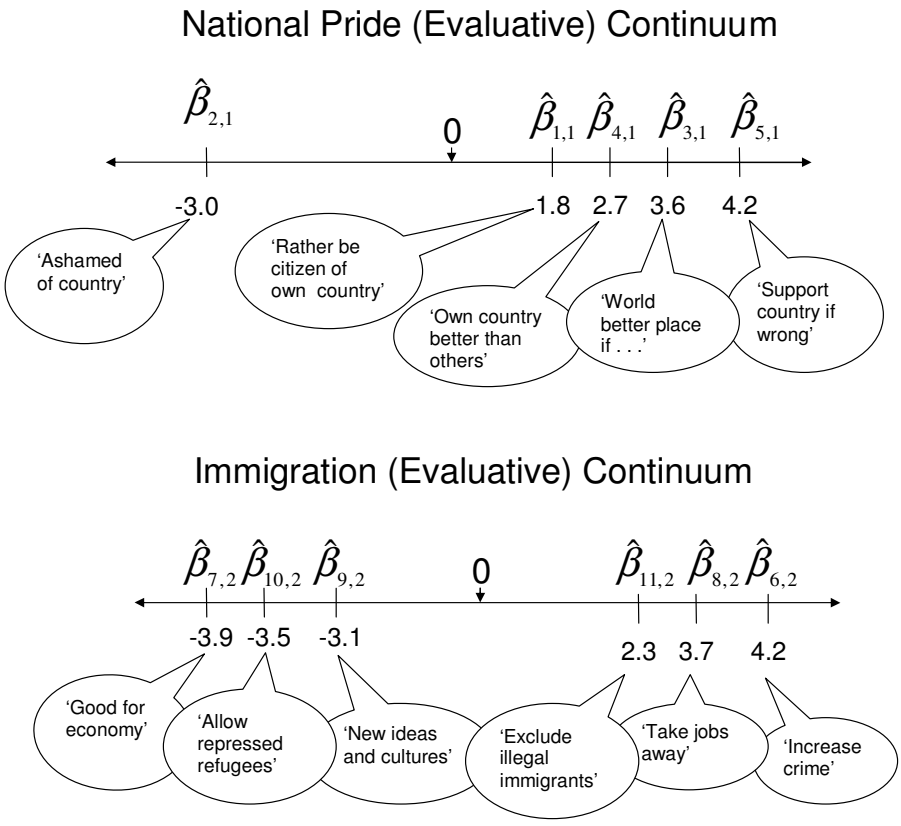


Figure 6.3: Statement locations for the national pride and immigration sets. *The statements locations were estimated by fitting Case 2a of the ULTMODR to the combined British and American data, with the same mean vector (*$= \mathbf{0}$*) and variance vector (*$= \mathbf{1}$*) used in* $g_1(\theta_i)$ *for both nations.*

We embedded the following structural model in the multi-set variant of the ULTMODR:

$$\theta_i = \gamma_0 + \gamma_1 X_i + \delta_i, \tag{6.3}$$

where θ_i is a vector containing person i's locations along the national pride continuum and the immigration continuum; where X_i indicates whether person i is American (e.g., $X_i = 0$ if person i is British and $X_i = 1$ if person i is American); where γ_0 is a vector containing the mean national pride attitude and the mean immigration attitude for the British; and where γ_1 is a vector containing the difference in mean national pride and the difference in mean immigration attitude between Americans and British people. Lastly, δ_i is an error vector that comes from a bivariate normal distribution

with mean vector $\mathbf{0}$ and covariance matrix, $\mathbf{\Phi}(X_i)$, that is itself a function of X_i:

$$\mathbf{\Phi}(X_i) = \begin{cases} \mathbf{\Phi}^{(US)} \text{ if } X_i = 1 \\ \mathbf{\Phi}^{(GB)} \text{ if } X_i = 0 \end{cases} . \tag{6.4}$$

Note that $\gamma_{1,1}$ is the parameter of primary interest.

All five cases of the model were fit to the data using the R function described in the last section of this chapter. The $-\ln(ML)$ values for these five models can be seen in the first five rows of Table 6.4; comparing them reveals that the less we constrain response category interpretation, the better the model fits the data. According to Likelihood Ratio Tests, Case 1a fits the best. Thus, from now on, we focus exclusively on this model. Estimates of Case 1a's parameters and hyperparameters, as well as some estimated standard errors, can be seen in Table 6.3.

In the fitted model, differences in American and British responses stem from both national differences in attitudes and national differences in response category interpretation. We were curious whether the NIS data could be better explained by the former differences alone or by the latter differences alone. To find out, we compared the fit of Case 1b (which has a common $g_2(\tau^i, \ln(\sigma^i))$ distribution but nation-specific $g_1(\theta_i)$ distributions) to the fit of an ULTMODR model with nation-specific $g_2(\tau^i, \ln(\sigma^i))$ distributions but a common $g_1(\theta_i)$ distribution for both nations. The models' $\ln(ML)$ values were -24642 and -24706, respectively. Model 1b fit substantially better than the additional model, despite having the same number of parameters. This suggests that differences in American and British responses are better explained by national differences in national pride rather than by national differences in response category interpretation. Obviously, however, the true explanation is probably a combination of these two phenomena, as is true for Case 1a.

In Case 1a, the parameter of primary interest suggests the same conclusion as the scoring method. Comparing $\widehat{\mu}_1^{(US)}$ to $\widehat{\mu}_1^{(GB)}$ suggests that Americans exhibit more national pride. We tested the hypothesis that $\gamma_{1,1} = \mu_1^{(US)} - \mu_1^{(GB)} = 0$ using a Likelihood Ratio Test. The value of $\ln(ML)$ was -24700 when $\gamma_{1,1} = 0$, compared to -24642 when $\gamma_{1,1}$ was estimated. Thus, we rejected the constrained model and concluded that Americans have a higher level of national pride, even after allowing for national differences in response category interpretation.

The remaining hyperparameters of $g_1(\theta_i)$ can be used to investigate other national differences in attitudes. Comparing $\widehat{\phi}_1^{(US)}$ to $\widehat{\phi}_1^{(GB)}$ suggests that Americans have more

Table 6.3: Parameter estimates and errors for Case 1a fit to the NIS data

Parameter		Estimate (standard error)	
		British ($n = 807$)	American ($n = 998$)
$g_1(\mu, \mathbf{\Phi})$ hyperparameters[a]	μ_1^{a1}	–0.38 (0.01)	0.12 (0.02)
	μ_2^{a2}	0.25 (0.01)	0.08 (0.02)
	$\phi_{1,1}$	0.66	0.47
	$\rho_{1,2}^{\dagger}$	0.47	0.45
	$\phi_{2,2}$	0.93	0.60
Threshold parameters	c_1	1.39	
	c_2	2.89	
	c_3	3.88	
	c_4	5.54	
$g_2(\varphi, \mathbf{\Lambda})$ hyperparameters[b]	φ_1^{b1}	$0.00^{\triangledown}$	0.74 (0.05)
	φ_2^{b2}	$-0.09^{\triangledown}$	0.10 (0.03)
	$\lambda_{1,1}$	1.33	0.81
	$\lambda_{1,2}$	0.34	0.25
	$\lambda_{2,2}$	0.09	0.08

a: Elements with subscripts 1 and 2 pertain to the national pride and immigration scales, respectively.

a1: More positive values imply greater national pride.

a2: More positive values imply more pro-immigration attitudes.

†: $\rho_{1,2} = \phi_{\mathbf{1},\mathbf{2}}/\sqrt{\phi_{\mathbf{1},\mathbf{1}}\phi_{\mathbf{2},\mathbf{2}}}$

b: Elements with subscripts 1 and 2 pertain to acquiescence and extremity, respectively.

b1: More positive values imply greater acquiescence.

b2: More positive values imply greater extremity.

$\triangledown$: Indicates that the value is pre-specified rather than estimated.

uniform national pride attitudes. Comparing $\widehat{\mu}_2^{(US)}$ to $\widehat{\mu}_2^{(GB)}$ and $\widehat{\phi}_2^{(US)}$ to $\widehat{\phi}_2^{(GB)}$ suggests that Americans exhibit immigration attitudes that are more pro-immigration and more uniform, respectively. Further, the estimates $\widehat{\rho}_{1,2}^{(US)}$ and $\widehat{\rho}_{1,2}^{(GB)}$ suggest that anti-immigration attitudes accompany pro-country attitudes for individuals from both nations. These estimated correlations are larger and more nearly equal than the corresponding Pearson correlations calculated using the scoring method. The lesser magnitude of the Likert score correlations is not surprising since they were calculated in two-stages, which results in attentuated estimates due to measurement error (Zwinderman, 1997, p. 245).

The hyperparameters of $g_2(\tau^i, \ln(\sigma^i))$ can be used to investigate national differences in response category interpretation. As we would have hypothesised *a priori*, Americans are significantly more acquiescent; this can be seen by comparing $\widehat{\varphi}_1^{(US)}$ and $\widehat{\varphi}_1^{(GB)}$, the estimated U.S. and G.B. means for τ^i. Also unsurprisingly, Americans are significantly more extreme, as revealed by a comparison of $\widehat{\varphi}_2^{(US)}$ and $\widehat{\varphi}_2^{(GB)}$, the estimated U.S. and G.B. means for $\ln(\sigma^i)$. A comparison of the diagonal elements in the covariance matrices suggests that Americans are more uniform in terms of acquiescence than the British, but equally uniform in terms of extremity. Also, the relationship between greater acquiescence and greater extremity is positive for both nations, but stronger for the British.

We assessed the overall goodness-of-fit of Case 1a to the entire NIS dataset. The maximum $\ln(ML)$ value from fitting the model was -24642, which translates to an average predicted probability of $exp\{-24642/(1805 \cdot 11)\} = 0.29$. This probability is very reasonable especially when we recall that, in Section 3.6, fitting the ULTMODR to data simulated from the ULTMODR produced probabilities that were not much larger. Further, the $-\ln(ML)$ value for Case 1a compared favourably to various proportional odds models for the NIS dataset. The logarithmic scores for these models can be seen in the bottom half of Table 6.4, which reveals that none fits as well as Case 1a. Note that two of the models (2 and 3) are product-multinomial models and have item-specific category cut-offs: The first assumes no differences between persons, and the second assumes no differences between persons from the same nation. All of the other models have the same category cut-offs for all items. These other models may have item-specific slopes and/or either nation- or person-specific slopes.[6] Note that, in

[6]All parameters, even the person-specific slopes, were treated as fixed-effects when fitting the proportional odds models.

some models with nation-specific or person-specific slopes, the sign of those slopes is reversed for the anti-country and pro-immigration items (i.e., 2, 7, 9, and 10).

Since we were primarily interested in the national pride items, we also examined the overall goodness-of-fit of Case 1a to those items only. The national pride component of $\ln(ML)$ is -11274. This value translates into an average predicted probability of 0.29, which is encouraging. Further, Case 1a fits the national pride data better than various proportional odds models, whose logarithmic scores can be seen in Table 6.5. These models were similar to the ones described in the previous paragraph. Note that the sixth model resembles the Rasch model (Rasch, 1960/1980) in formulation. In addition, note that the seventh model distinguishes between items only by indicating whether they are pro- or anti-country in the score for each person; thus, this model can be viewed as a probabilistic version of Likert's measurement model. Although two proportional odds models do fit better than Case 1a, they have a very large number of parameters (1818) because their person-specific slopes are treated as fixed-effects.

We examined the fit of Case 1a to the univariate margins for each nation. Figure 6.4 presents the British expected and observed frequencies for each margin, along with the corresponding unsigned Pearson residual; Figure 6.5 presents the same statistics for the American responses. For both nations, the model fit the S-8 margins the best. For the British, the model fit the S-7 margins the worst, mostly because it underpredicts the number of British respondents who are neutral on whether immigrants are good for the economy. For the Americans, the model fit the S-9 margins the worst, mostly because it underpredicts the number of Americans who agree that immigrants make the country open to new ideas and cultures.

We also examined the fit of Case 1a to the bivariate margins for each nation. For the British and American responses, we calculated the observed and expected frequencies, and the signed Pearson residuals, for every pair of statements. For example, Table 6.6 presents these statistics for the American S-1 and S-2 margins, which are among the most poorly fit. (The other $2 \cdot 55 - 1$ tables are omitted for the sake of brevity.) The table reveals that the model severely underpredicts the frequency for the 'SA' category of S-1 and the 'A' category of S-2. In general, the bivariate statistics highlight the same problems with fit as the univariate statistics. For the British, the model fits all two-way margins involving S-7 most poorly, because it underpredicts the frequencies involving the 'N' category. For the Americans, the model fits all two-way margins involving S-9 most poorly, because it underpredicts of the frequencies involving the 'A' category.

Table 6.4: Logarithmic scores for models fit to the national pride and immigration data

Model	Log. Score† (num. param.)
ULTMODR models	
Case 1b	24642 (33)
Case 1a	24706 (28)
Case 2b	25218 (29)
Case 2a	25239 (27)
Case 1	25325 (28)
Case 1b, with common $g_1(\theta_i)$	24742 (28)
Proportional odds models	
$N(i)$: indicates the nation (GB or US) to which person i belongs $R(j)$: indicates if item j is 'reversed' $S(j)$: indicates the set (NP or I) to which item j belongs	
$\text{logit}P(Y_{ij} \leq k) = \beta_k + \beta_j$	27158 (14)
$\text{logit}P(Y_{ij} \leq k) = \beta_{jk}$	26754 (44)
$\text{logit}P(Y_{ij} \leq k) = \beta_{jkN(i)}$	26452 (88)
$\text{logit}P(Y_{ij} \leq k) = \beta_k + \beta_j + \beta_{N(i)}$	27081 (15)
$\text{logit}P(Y_{ij} \leq k) = \beta_k + \beta_j + (1 - I(R_j))\beta_{N(i)} - I(R_j)\beta_{N(i)}$	27039 (15)
$\text{logit}P(Y_{ij} \leq k) = \beta_k + \beta_j + \beta_{S(j)N(i)}$	27057 (16)
$\text{logit}P(Y_{ij} \leq k) = \beta_k + \beta_j + (1 - \text{I}(R_j))\beta_{S(j)N(i)} + \text{I}(R_j)\beta_{S(j)N(i)}$	26993 (16)
$\text{logit}P(Y_{ij} \leq k) = \beta_k + \beta_j + \theta_i$	25532 (1818)

†: Log. Score = $\sum_{i=1}^{1805} \sum_{j=1}^{11} \sum_{k=1}^{5} -I\left(Y_{ij} = k\right) \ln P\left(Y_{ij} = k\right)$

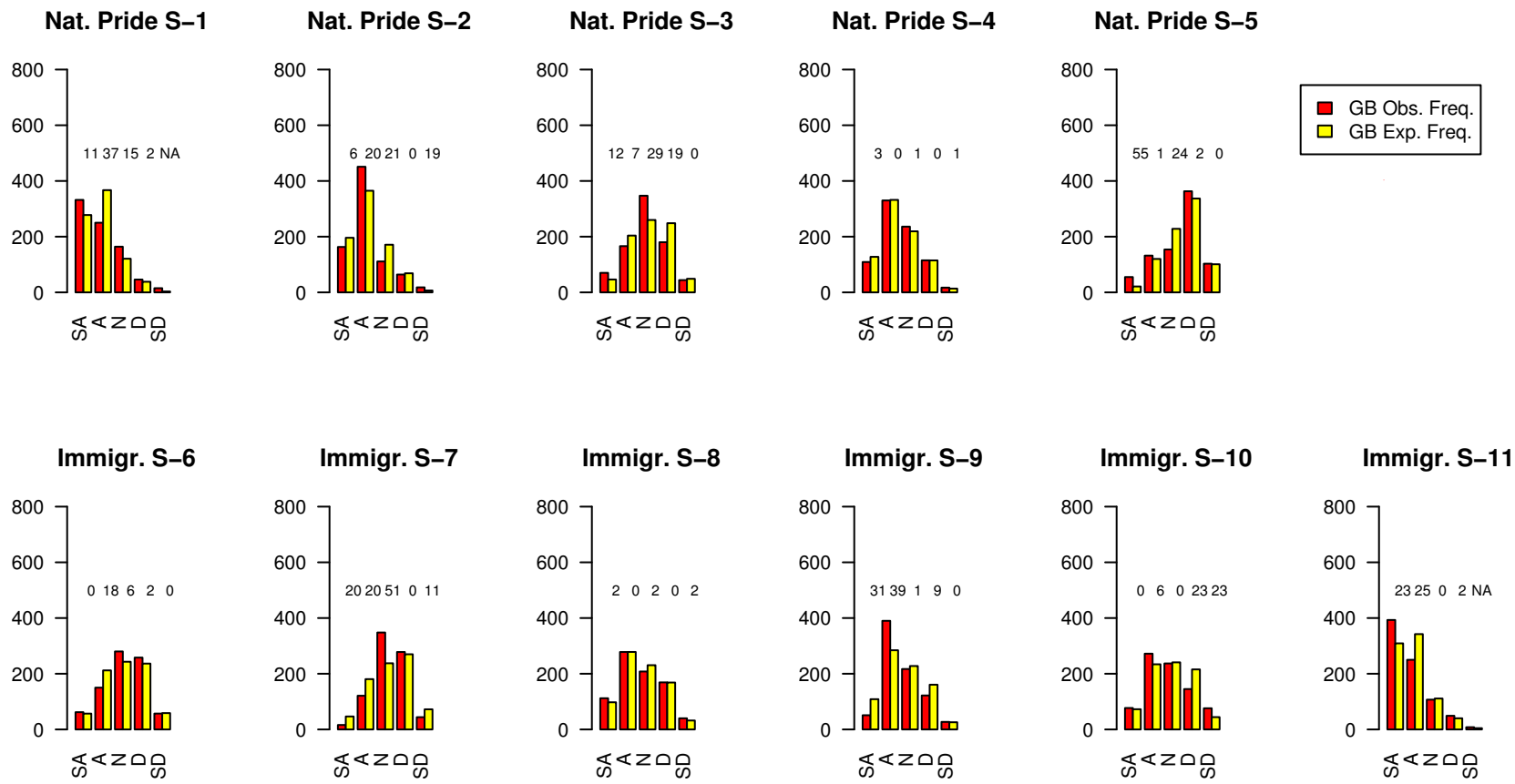


Figure 6.4: Case 1a univariate goodness-of-fit statistics for British responses in the NIS dataset. *Each plot pertains to an item in the NIS dataset. In each plot, the five pairs of bars depict observed and expected frequencies for the categories. The expected frequencies were calculated using an equation similar to (3.5) with the British* $g_1(\theta_i)$ *and* $g_2(\tau^i, \ln(\sigma^i))$ *distributions and each integration approximated using Monte Carlo immigration. The number written above each pair of bars is the corresponding unsigned Pearson residual. This value is omitted when the expected frequency is smaller than five.*

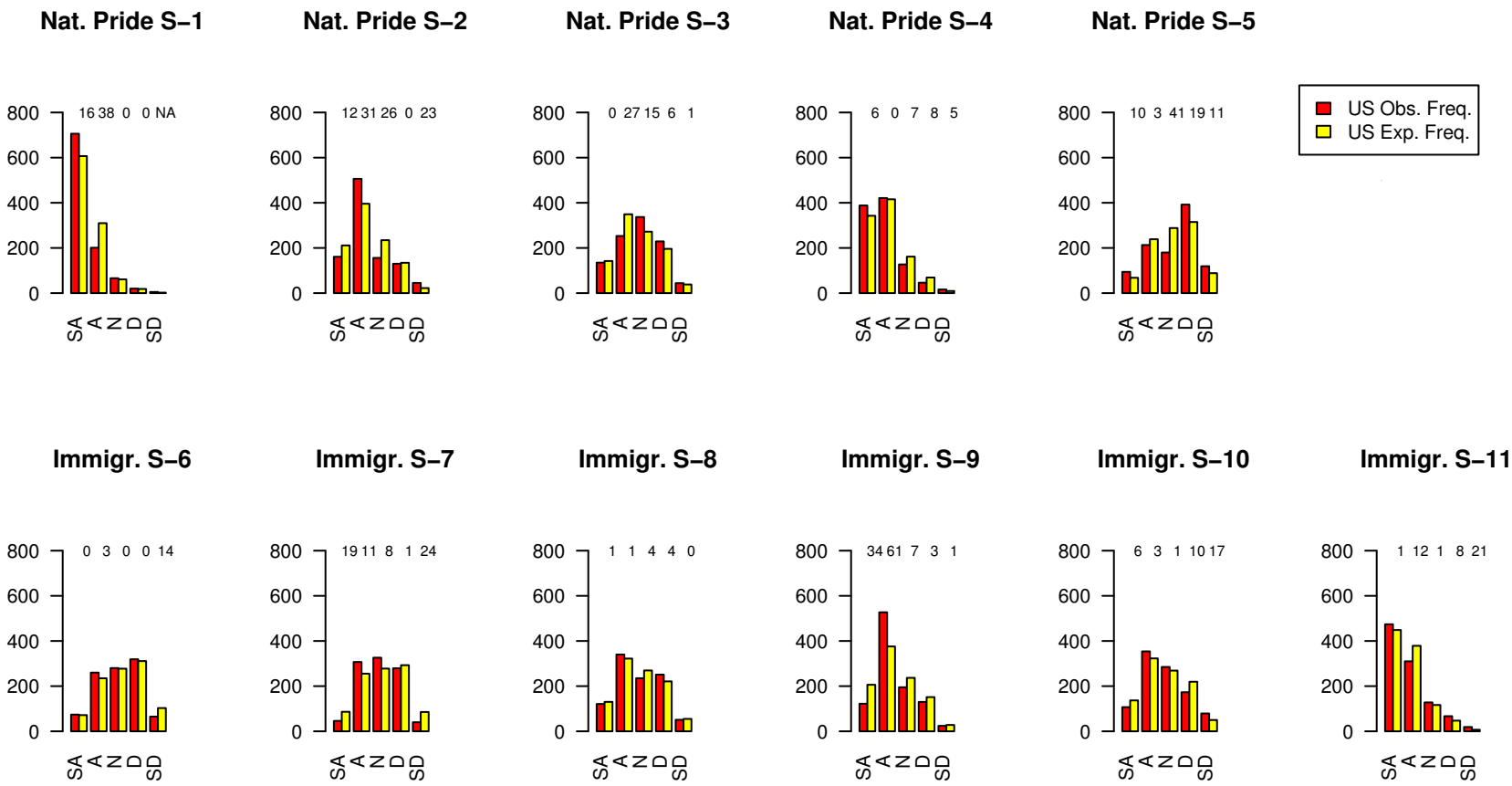


Figure 6.5: Case 1a univariate goodness-of-fit statistics for American responses in the NIS dataset. *Each plot pertains to an item in the NIS dataset. In each plot, the five pairs of bars depict observed and expected frequencies for the categories. The expected frequencies were calculated using an equation similar to (3.5) with the American* $g_1(\theta_i)$ *and* $g_2(\tau^i, \ln(\sigma^i))$ *distributions and each integration approximated using Monte Carlo immigration. The number written above each pair of bars is the corresponding unsigned Pearson residual. This value is omitted when the expected frequency is smaller than five.*

Table 6.5: Logarithmic scores for models fit to the national pride data

Model	Log. Score† (num. param.)
ULTMODR model[a]	
Case 1b	11274 (22)
Proportional odds models[b]	
$N(i)$: indicates the nation (GB or US) to which person i belongs	
$R(j)$: indicates if item j is 'reversed'	
$S(j)$: indicates the set (NP or I) to which item j belongs	
$\text{logit}(P(Y_{ij} \leq k) = \beta_k + \beta_j$	11927 (8)
$\text{logit}(P(Y_{ij} \leq k) = \beta_{jk}$	11694 (20)
$\text{logit}(P(Y_{ij} \leq k) = \beta_{jkN(i)}$	12129 (40)
$\text{logit}(P(Y_{ij} \leq k) = \beta_k + \beta_j + \beta_{N(i)}$	12042 (9)
$\text{logit}(P(Y_{ij} \leq k) = \beta_k + \beta_j + (1 - \text{I}(R_j))\beta_{N(i)} - \text{I}(R_j)\beta_{N(i)}$	11984 (9)
$\text{logit}(P(Y_{ij} \leq k) = \beta_k + \beta_j + \theta_i$	9981 (1812)
$\text{logit}(P(Y_{ij} \leq k) = \beta_k + (1 - \text{I}(R_j))\theta_i - \text{I}(R_j)\theta_i$	11590 (1808)
$\text{logit}(P(Y_{ij} \leq k) = \beta_k + \beta_j + (1 - \text{I}(R_j))\theta_i - \text{I}(R_j)\theta_i$	9625 (1812)

†: Log. Score = $\sum_{i=1}^{1805} \sum_{j=1}^{11} \sum_{k=1}^{5} -I\left(Y_{ij} = k\right) \ln P\left(Y_{ij} = k\right)$

[a] This model was fit to the National Pride and Immigration items

[b] These models were fit to the National Pride items only

Last, we checked that Case 1a produced attitude estimates that sensibly incorporated differing response category interpretation. Specifically, we calculated EAP estimates for two imaginary British surveyees: The first surveyee had data

$$Y_i = [\ A \quad D \quad A \quad A \quad A \quad SA \quad SD \quad SD \quad SA \quad SA \quad SD\ ]^T, \tag{6.5}$$

and the second surveyee had data

$$Y_i = [\ A \quad D \quad A \quad A \quad A \quad A \quad D \quad D \quad A \quad A \quad D\ ]^T. \tag{6.6}$$

Although both surveyees responded identically to the five national pride items, the estimate of θ_i^1 is 0.08 for the first respondent and 0.50 for the second respondent (more positive values correspond to being more pro-country.) This difference reflects the surveyees' differing responses to the immigration items: Since the second surveyee's responses to these items reveal an aversion to using the outermost categories, her responses to the national pride items seem a stronger indication of national pride. With-

Table 6.6: Goodness-of-fit statistics for American S-1 and S-2 two-way margins

Entry(k, m)
=
$n_{US} \cdot p_{1k2m}$ $(n_{US} \cdot \hat{p}_{1k2m})$
χ^2_{1k2m}

Categories		S-2 'SA'	'A'	'N'	'D'	'SD'
S-1	'SA'	117 (120) χ^2 = 0	356 (216) χ^2 = 91	94 (152) χ^2 = -22	96 (100) χ^2 = 0	43 (18) χ^2 = 33
	'A'	25 (65) $\chi^2_{2,1}$ = -25	106 (144) $\chi^2_{2,2}$ = -10	43 (69) $\chi^2_{2,3}$ = -10	27 (29) $\chi^2_{2,4}$ = 0	0 (3) $\chi^2_{2,5}$ = NA[a]
	'N'	13 (18) χ^2 = -2	33 (28) χ^2 = 1	16 (10) χ^2 = 3	3 (4) χ^2 = NA[a]	1 (0) χ^2 = NA[a]
	'D'	6 (7) $\chi^2_{4,1}$ = 0	8 (7) χ^2 = 0	2 (3) χ^2 = NA[a]	4 (1) χ^2 = NA[a]	0 (0) χ^2 = NA[a]
	'SD'	0 (1) χ^2 = 0	3 (1) χ^2 = 0	1 (0) χ^2 = NA[a]	0 (0) chi^2 = NA[a]	1 (0) χ^2 = NA[a]

[a] χ^2 statistic omitted because $n_{US} \cdot \hat{p}_{jklm} < 5$

out the immigration items (and a model that incorporates differing response category interpretation), the estimates of θ_i^1 would be the same for both surveyees.

In conclusion, both the scoring method and our new method suggest that Americans exhibit more national pride than the British. However, we have more confidence in the new method's conclusions because they are adjusted for national differences in response category interpretation.

6.4 Application: Investigating abortion attitudes

Finally, we used the new method to analyse a subset of the abortion attitude items. In particular, we wanted to investigate how nationality (GB or US), gender, and religious status (Christian or none) affect abortion attitudes, while controlling for national and gender differences in response category interpretation. Note that, unfortunately, the dataset does not include any secondary attitude items that could provide additional information on response category interpretation. (If the author had been more prescient, she might have expanded the web-based survey described in Section 1.3.1 to include some statements about non-abortion objects.) However, the presence of additional items is not as crucial in this application because the primary set contains a large number of items expressing heterogeneous and relatively balanced views.

Before performing any analysis, we discarded those abortion attitude items that were problematic according to our ULTMODR-based method of item analysis in Section 4.5. We removed four items with outlying 'DK/CC' proportions (i.e., S-14, S-15, S-19, and S-21), three items with outlying $\hat{\sigma}^2_{rk(\hat{\beta}_j^{ns})}$ values (S-25, S-29, and S-31), and six items with outlying χ_j^2 values (S-30, S-33, S-34, S-39, S-41, and S-44). After doing so, 37 Likert items remained, and their 'DK/CC' responses were recoded as 'Neither agree nor disagree.' Note that the remaining items do not completely balance anti-abortion and pro-abortion statements: There are more of the former than the latter. Thus, according to the results of the simulation experiments in Section 6.2, we should not use the scoring method to draw formal conclusions about group differences in abortion attitudes. In fact, we cannot even use the scoring method to perform preliminary investigations because some of the 37 statements cannot be classified as pro- or anti- abortion.

We also discarded the eleven persons who currently practice religions other than Christianity. We did so because the American sample included several Jewish persons

whereas the British sample did not, and we wanted to make sure that the religious samples were reasonably similar for the two nations. After discarding the eleven persons (most of them American), 129 persons remained. The resulting frequencies for each of the eight nationality×gender×religion groups can be seen in Table 6.7.

Before we could use the new-structural-model-fitting method, we had to fix the statement locations to prespecified values. To do so, we fit Case 2a of the simplest ULTMODR variant to the remaining 129 persons and 37 items. (This model was fit using the R function described in Section 4.7.2.) In the model, the thresholds were assumed to differ by nationality×gender. However, the distribution of θ_i^1 was assumed to be the same for persons from both nations; the common mean was fixed to zero, and the common variance to one. The $\hat{\beta}_j$s that resulted from fitting this model were consistent with the statements' content and ranged from -4.90 for S-5 to -3.65 for S-47. Thus, in the following analysis, more positive θ_i^1 values correspond to more liberal attitudes towards abortion.

We then used these statement locations to implement our structural-model-fitting method. Specifically, we fit Case 1a of the multi-set ULTMODR variant (with $S = 1$) to the remaining data. The hyperparameters of the $g_2(\tau^i, \ln(\sigma^i))$ distribution were allowed to differ by nationality×gender, and the hyperparameters of the $g_1(\theta_i^1)$ distribution were allowed to differ by nationality×gender×religion. The model was fit using the R function described in the last section of this chapter. Estimates (and some estimated standard errors) for the model's hyperparameters can be seen Table 6.7. The maximum $\ln(ML)$ value for the model is -5133. This corresponds to an average predicted probability of $exp\{-5133/(129 \cdot 37)\} = 0.34$. This value is even higher than the probabilities observed in the simulation experiment in Section 3.6, which indicates that the model explains the data very well.

We used the eight sets of estimated $g_1(\theta_i^1)$ hyperparameters to investigate group differences in attitudes. Beginning with the mean of θ_i^1, we see that British Christian males have the most conservative attitudes, and that American and British non-religious females have the most liberal attitudes. As we might expect, female students have significantly more liberal attitudes than male students in every nationality× religion group. Similarly, non-religious student have significantly more liberal abortion attitudes than Christian students in every nationality×gender group. Interestingly, American Christian students are more liberal than British Christian students, significantly so for females. Non-religious British males have more liberal attitudes than their

American counterparts, whereas non-religious British females have attitudes identical to their American counterparts. Turning to the variance of θ_i^1, we see that male and then female American Christians have the most heterogeneous views, and that British and then American non-religious females have the most uniform views.

We used the four sets of estimated $g_2(\tau^i, \ln(\sigma^i))$ hyperparameters to interpret group differences in response category interpretation. They suggest that each of the four nationality$\times$ gender groups has significantly different response category interpretation. We see that American males are the least acquiescent and least extreme, and that American females are the most acquiescent and most extreme. British females fall somewhere in between, and British males are very similar to American females. British students have more uniform response category interpretation than American students, and male students have more uniform response category interpretation than female students.

Last, we tested whether response category interpretation does differ by nationality$\times$ gender. To do so, we fit Case 1b of the model. (Recall that the hyperparameters of $g_2(\tau^i, \ln(\sigma^i))$ are the same for all persons in Case 1b.) The $\ln(ML)$ value for this model is -5166, and the Likelihood Ratio Test comparing it to Case 1a has a p-value less than 0.01. Thus, it seems that response category interpretation does indeed differ by gender and nation.

6.5 Conclusions

We have introduced a new method of fitting (structural) models for the effect of covariates on attitudes. The method embeds the structural model in the ULTMODR measurement model. Because of this embedding, both models can be fit simultaneously so that the resulting estimates of the structural model's parameters are less attenuated. Further, these estimates can be adjusted for differing response category interpretation by modelling it in the ULTMODR's response structure. This adjustment can be made even in a situation where there are only a few, unbalanced items measuring the attitude of primary interest. If other items (measuring secondary attitudes) are available, they can be incorporated into a multi-set variant of the ULTMODR; doing so provides additional information on how response category intepretation differs. If secondary sets are incorporated, the multi-set variant allows us to obtain less attenuated estimates of the correlation between the primary and secondary attitudes, an additional bonus.

Table 6.7: Parameter estimates and errors for Case 1a fit to 37 abortion attitude items

Parameter		Estimate (standard error)							
		British				American			
		Male		Female		Male		Female	
		None ($n = 28$)	Chrstn ($n = 9$)	None ($n = 22$)	Chrstn ($n = 8$)	None ($n = 19$)	Chrstn ($n = 14$)	None ($n = 20$)	Chrstn ($n = 9$)
$g_1(\mu, \mathbf{\Phi})$	μ_1^{a1}	0.30 (0.12)	–1.39 (0.13)	0.61 (0.09)	–0.87 (0.14)	0.03 (0.17)	-0.94 (0.18)	0.61 (0.07)	0.17 (0.13)
hyperparameters	$\phi_{1,1}$	0.64	1.22	0.19	0.26	0.48	2.94	0.33	1.84
$g_2(\varphi, \mathbf{\Lambda})$ hyperparameters^b	φ_1^{b1}	0.75 (0.06)		0.48 (0.11)		$0.00^{\triangledown}$		0.81 (0.14)	
	φ_2^{b2}	0.27 (0.04)		0.08 (0.06)		$-0.15^{\triangledown}$		0.28 (0.09)	
	$\lambda_{1,1}$	0.13		0.44		0.31		0.40	
	$\lambda_{1,2}$	0.07		0.25		0.18		0.28	
	$\lambda_{2,2}$	0.07		0.15		0.15		0.21	

$a1$: More positive values imply more liberal attitudes towards abortion.

b: Elements with subscripts 1 and 2 pertain to acquiescence and extremity, respectively.

b1: More positive values imply greater acquiescence.

b2: More positive values imply greater extremity.

$\triangledown$: Indicates that the value is pre-specified rather than estimated.

Simulation experiments indicate that inferences drawn using the new method are robust to group differences in response category interpretation, even for an unbalanced scale. On the other hand, the scoring method based on Likert's measurement model does not appear to perform well for unbalanced scales.

In our applications, we find that the new method does make it possible to separate the effects of response category interpretation from the effects of attitudes. In the national pride application, this separation is accomplished by borrowing strength across scales.

6.6 Details of the R function

We describe the R function written by the author to fit the multi-set variant of the ULTMODR with all five response structure cases.

The fixed-effects parameters and hyperparameters are estimated by maximising the relevant Marginal Likelihood. For Cases 2a, 2b and 3, the Marginal Likelihood is

$$ML = \prod_{i=1}^{n} \left\{ \iint g_1(\theta_i) \cdot \prod_{j=1}^{J} \prod_{k=1}^{K} I\left(Y_{ij} = k\right) P\left(Y_{ij} = k\right) d(\theta_i^1) d(\theta_i^2) \right\}. \tag{6.7}$$

In our national pride application, for instance,

$$g_1(\theta_i) = \begin{cases} BVN(\mu^{(US)}, \mathbf{\Phi}^{(US)}) \text{ where } \mu^{(US)} = \gamma_0 + \gamma_1 & \text{if } X_i = 1 \\ BVN(\mu^{(GB)}, \mathbf{\Phi}^{(GB)}) \text{ where } \mu^{(GB)} = \gamma_0 & \text{if } X_i = 0 \end{cases}.$$

For Cases 1a and 1b, the Marginal Likelihood is

$$\begin{aligned} ML \;=\; & \prod_{i=1}^{n} \iiiint \left\{ g_1(\theta_i) \cdot g_2(\tau^i, \ln(\sigma^i)) \cdot \right. \\ & \left. \prod_{j=1}^{J} \prod_{k=1}^{K} I\left(Y_{ij} = k\right) P\left(Y_{ij} = k\right) \right\} d(\theta_i^1) d(\theta_i^2) d(\tau^i) d(\ln(\sigma^i)). \end{aligned} \tag{6.8}$$

where, for Case 1b,

$$g_2(\tau^i, \ln(\sigma^i)) = BVN(\varphi, \Lambda) \text{ with } \varphi_1 = 0 \text{ and } \varphi_2 = -\lambda_{2,2},$$

and, for Case 1a,

$$g_2(\tau^i, \ln(\sigma^i)) = \begin{cases} BVN(\varphi^{(g)}, \Lambda^{(g)}) & \text{if } g > 1 \\ BVN(\varphi^{(g)}, \Lambda^{(g)}) \text{ where } \varphi_1^{(g)} = 0 \text{ and } \varphi_2^{(g)} = -\lambda_{2,2}^{(g)} & \text{if } g = 1 \end{cases}.$$

The n integrals in (6.7) or (6.8) are approximated using Monte Carlo integration. More specifically, with (6.7), the change-of-variables technique was used to transform the i^{th} integral into an integral with respect to θ_i', where θ_i' comes from a standard bivariate normal distribution. The transformed integral is then approximated using Monte Carlo integration implemented with $N = 600$ points generated from the standard bivariate normal distribution. (For Cases 1a and 1b, we adopt an analogous approach to approximate each integral in (6.8), but employ $N = 1000$ points generated from a standard four-dimensional normal distribution.)

The logarithm of the relevant ML is maximised using *R*'s optim() function with method="L-BFGS-B", which implements the optimization technique of Byrd et al. (1995), a quasi-Newton method that allows box constraints. To ensure that the $\mathbf{\Phi}$ matrices (and, in Case 1a or 1b, the Λ matrices) remain symmetric and positive definite throughout optimisation, $\ln(ML)$ is maximised with respect to the elements of the matrices' Choleski decompositions. Similarly, in order to ensure that the threshold set(s) remain ordered, $\ln(ML)$ is maximised with respect to the differences in the thresholds, subject to the constraint that these differences are positive. Lastly, to ensure that $\sigma^{(g)}$ in Case 2b remains positive, $\ln(ML)$ is maximised with respect to $\ln(\sigma^{(g)})$.

To begin the optimisation process, sensible starting values suggested by a combination of theory and experience are used for the fixed-effects parameters. However, since $\log(ML)$ can have multiple modes, the optimisation process is then repeated using alternative starting values arrived at by jittering the initial starting values. The values of the fixed-effects parameters corresponding to the largest maxima are retained as estimates.

The final $ln(ML)$ value is calculated by evaluating either (6.7) or (6.8) at the estimated values of the fixed-effects parameters, using $N = 10000$ points in the Monte Carlo approximations for the n integrals.

Estimated standard errors for the $\hat{\mu}$s and the $\hat{\varphi}$s are calculated by taking the square root of the relevant diagonal elements of the inverse of the observed information matrix, evalulated at the MLEs of the fixed-effects parameters. The observed information matrix is calculated by optim(), which returns a numerical approximation of the Hessian matrix at the solution found. Although optim() actually returns the Hessian matrix of the unconstrained problem, the box constraints are not active in the solutions found for any of the models.

Last, the person-effects estimates are estimated in the manner described in Section 3.4.

Chapter 7

A few final comments

This thesis has introduced the ULTMODR—a new measurement model for Likert data—and demonstrated how its different variants can be used to analyse Likert data in a variety of ways. The ULTMODR-based methods of analysis generally perform better than comparable existing methods. This is particularly true for our method of fitting structural models.

However, our ULTMODR-based methods do suffer certain limitations. For one, they are implemented using a normal prior distribution for the person locations. This may not be appropriate if the (population) distribution of attitudes is multimodal, which is entirely possible. Bartholomew and Knott (1999) argue that, in latent variable modelling, the prior distribution has little effect, meaning that we should not concern ourselves too much with its form. This may be true in situations where we want to draw conclusions about the items (e.g., when visualising the relationships between them or when determining which ones confuse people). However, the appropriateness of the prior distribution is a greater cause for concern when we want to draw conclusions about the persons' attitudes (e.g., when visualising them or when formally comparing them between groups). In these instances, we might try to fit the ULTMODR using an alternative prior distribution, such as a mixture of two normals.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. Hoboken, NJ: John Wiley.

Andrich, D. (1978). "A rating formulation for ordered response categories." *Psychometrika*, 43, 561–573.

— (1996). "A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies." *British Journal of Mathematical and Statistical Psychology*, 49, 347–365.

Bartholomew, D. and Knott, M. (1999). *Latent Variable Models and Factor Analysis*. 2nd ed. London: Arnold.

Bartholomew, D., Steele, F., Moustaki, I., and J.I., G. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Boca Raton, FL: Chapman and Hall / CRC.

Bohner, G. (2001a). "Attitudes." In *Introduction to Social Psychology*, eds. M. Hewstone and W. Stroebe, 3rd ed., 239–382. Oxford: Blackwell.

— (2001b). "Attitudes." In *Introduction to Social Psychology*, eds. M. Hewstone and W. Stroebe, 3rd ed., 239–382. Oxford: Blackwell.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley.

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*. New York, NY: Spring-Verlage.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). "A limited memory algorithm for bound constrained optimization." *SIAM Journal of Scientific Computing*, 16, 1190–1208.

Churchill Jr., G. (1999). *Marketing Research: Methodological Foundations*. Seventh ed. Orlando, FL: Dryden Press.

Coombs, C. (1950). "Psychological scaling without a unit of measurement." *Psychological Review*, 57, 145–158.

— (1964). *A Theory of Data*. New York, NY: Wiley.

Cox, T. and Cox, M. (2001). *Multidimensional Scaling*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Cronbach, L. J. (1951). "Coefficient alpha and the internal structure of tests." *Psychometrika*, 16, 297–334.

Davison, M. (1977). "On a metric, unidimensional unfolding model for attitudinal and developmental data." *Psychometrika*, 42, 523–548.

Eagly, A. and Chaiken, S. (1998). "Attitude structure and function." In *The Handbook of Social Psychology*, eds. D. Gilbert, S. Fiske, and G. Lindzey, 4th ed., 269–322. New York, NY: McGraw-Hill.

Erwin, P. (2001). *Attitudes and Persuasion*. Hove, East Sussex: Psychology Press.

Fishbein, M. and Ajzen, I. (1972). "Attitudes and opinions." *Annual Review of Psychology*, 23, 487–544.

— (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis, MN: University of Minnesota Press.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Thousand Oaks, CA: Sage.

Johnson, M. (2001). "Parametric and Non-Parametric Extensions of Unfolding Response Models." Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Jöreskog, K. and Moustaki, I. (2001). "Factor analysis of ordinal variables: A comparison of three approaches." *Multivariate Behavioral Research*, 36, 347–387.

Judd, C. and Kulik, J. (1980). "Schematic effects of social attitudes on information processing and recall." *Journal of Personality and Social Psychology*, 38, 569–578.

Kaiser, H. (1958). "The varimax criterion for analytic rotation in factor analysis." *Psychometrika*, 23, 187–200.

Kendall, D. (1971). "Seriation from abundance matrices." In *Mathematics in the Archaeological and Historical Sciences*, eds. F. R. Hodson, D. G. Kendall, and P. Tatu, 215–252. Edinburgh: Edinburgh University Press.

Likert, R. (1932). "A technique for the measurement of attitudes." *Archives of Psychology*, No. 140.

Lingoes, J. (1973). *The Guttman-LINGOES Nonmetric Program Series*. Mathesis Press, Ann Arbor, Michigan.

Luo, G. (2001). "A class of probabilistic unfolding models for polytomous responses." *Journal of Mathematical Psychology*, 45, 224–248.

Marden, J. (1995). *Analyzing and Modeling Rank Data*. London: Chapman & Hall.

Masters, G. (1982). "A Rasch model for partial credit scoring." *Psychometrika*, 47, 149–174.

Masters, G. and Wright, B. (1984). "The essential process in a family of measurement models." *Psychometrika*, 49, 529–544.

Moustaki, I. (2000). "Structural Equation Modeling: Present and Future." In *A review of exploratory factor analysis for ordinal categorical data.*, eds. R. Cudeck, S. Du Toit, and D. Sörbom. Scientific Software International.

Mueller, D. J. (1986). *Measuring Social Attitudes: A Handbook for Researchers and Practitioners*. New York: Teachers College Press.

Muraki, E. (1992). "A generalized partial credit model: Application of an EM algorithm." *Applied Psychological Measurement*, 16, 159–176.

Oppenheim, A. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*. New ed. London: Continuum.

Pratkanis, A., Breckler, S., and Greenwald, A. (1989). "The cognitive representation of attitudes." In *Attitude structure and function*, ed. A. Pratkanis, 71–98. Hillsdale, NJ: Erlbaum.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. expanded ed. Chicago: The University of Chicago Press. Reprint of the original 1960 publication by the Danish Institute for Eductional Research.

Roberts, J. (1995). "Item Response Theory Approaches to Attitude Measurement." Ph.D. thesis, University of South Carolina, Columbia, SC.

Roberts, J., Donoghue, J., and Laughlin, J. (2000). "A general item response theory model for unfolding undimensional polytomous responses." *Applied Psychological Measurement*, 24, 3–32.

Roberts, J. and Laughlin, J. (1996). "A unidimensional item response model for unfolding responses from graded disagree-agree response scale." *Applied Psychological Measurement*, 20, 231–255.

Roskam, E. (1979). "A Survey of the Michigan-Israel-Netherlands-Integrated Series." In *Geometric Representations of Relational Data.*, eds. J. Lingoes, E. Roskam, and I. Bor, 289–312. Ann Arbor, MI: Mathesis Press.

Rossi, P., Gilula, Z., and Allenby, G. (2001). "Overcoming scale usage heterogeneity: A Bayesian hierarchical approach." *Journal of the American Statistical Association*, 96, 20–31.

Rost, J. and Luo, G. (1997). "An application of a Rasch based model to a questionnaire on adolescent centrism." In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, ed. G. Bohner. Münster: Waxmann Verlag GMBH.

Samejima, F. (1969). "Estimation of latent ability using a response pattern of graded scores." *Psychometrika Special Monograph*, Monograph Supplement No. 17.

Shepard, R. (1974). "Representation of structure in similarity data: Problems and prospects." *Psychometrika*, 39, 373–421.

Shi, J. and Lee, S. (1998). "Bayesian sampling-based approach for factor analysis models with continuous and polytomous data." *British Journal of Mathematical and Statistical Psychology*, 51, 233–252.

Smith, T. and Jarkko, L. (2001). *National Pride in Cross-National Perspective*. National Opinion Research Center / University of Chicago.

Spector, P. (1990). *Summated Rating Scale Construction: An Introduction*. Newbury Park, CA: Sage.

Thissen, D. and Steinberg, L. (1986). "A taxonomy of item response models." *Psychometrika*, 51, 567–577.

Thurstone, L. and Chave, E. (1929). *The Measurement of Attitude: A Psychophysical Method and Some Experiments with a Scale for Measuring Attitude Towards Church*. Chicago, IL: University of Chicago Press.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wright, B. and Masters, G. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Young, F. and Lewyckyj, R. (1979). *ALSCAL-4 user's guide*. Psychometric Laboratory, University of North Carolina at Chapel Hill. Implemented in the software package SPSS.

Zwinderman, A. (1997). "Response models with manifest predictors." In *Handbook of Modern Item Response Theory*, eds. W. van der Linden and R. Hambleton, 245–256. New York, NY: Springer-Verlag.