An ‘Unfolding’ Latent Variable Model for Likert Attitude Data: Drawing Inferences Adjusted for Response Style

March 15, 2005

This paper has been submitted to the Journal of the American Statistical Association and should not be circulated or forwarded without the permission of the authors.
Abstract

Likert attitude data consist of responses to favorable and unfavorable statements about an entity, where responses fall into ordered categories ranging from disagreement to agreement. Social science and marketing researchers frequently use data of this type to measure attitudes toward an entity such as a policy or product. Here, an individual’s ‘attitude’ is defined as his or her location along an underlying, unobserved, continuum running from extremely unfavorable to extremely favorable evaluations of the entity. We introduce a Multidimensional Unfolding Model (MUM) to describe the relationship between the data and the attitudes underlying them. Unlike most existing models, the MUM allows the data to reflect not only attitudes but also ‘response style,’ defined as a consistent and content-independent pattern of response category selection, such as a tendency to agree with all statements. The model treats each response as a coarsened version of an unobserved continuous agreement variable; the coarsening thresholds are allowed to vary between individuals or groups to reflect response style differences. The unobserved agreement variable is modeled as a decreasing function of the distance between a parameter quantifying the statement’s favorability and a latent variable representing the individual’s attitude. The latent variable may easily be treated as multidimensional to allow multiple attitudes to underlie the data. This feature enables researchers to incorporate any additional Likert data into their analysis to increase information on response style when the data of primary interest are limited. The MUM can be used to estimate attitudes or to fit general linear models
for the effects of background variables on attitudes. We propose a maximum (marginal) likelihood approach to fitting, and also describe statistics for assessing local and global goodness of fit. MUM-based inferences about attitudes are adjusted for response style, and, thus, should be less biased. We illustrate the model with real data on American and British attitudes toward their respective nations (‘national pride’); data on the respondents’ immigration attitudes are included to increase the limited information on national differences in response style. We perform simulation experiments mimicking our real data analysis, which estimates the response-style-adjusted difference in two groups’ mean attitudes. The results strongly suggest that, unlike Likert’s popular scoring model, the MUM yields correct inferences even when there are unequal proportions of favorable and unfavorable statements.

KEY WORDS: Hierarchical model; Latent variable; Latent trait; Item response theory; Attitude measurement; Scale usage heterogeneity
1 INTRODUCTION

Researchers in the social sciences and marketing often measure individuals’ attitudes toward a particular entity. Here, attitude refers to a hypothesized underlying construct, not to the construct’s observable manifestations, which are called ‘attitudes’ in common parlance. More precisely, an attitude will be defined as ‘a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor’ (Eagly and Chaiken 1998, p. 269). The entity may be an object (e.g., a hybrid SUV), a group (e.g., doctors), an action (e.g., immigration), or a concept (e.g., democracy).

To collect data on individuals’ attitudes, researchers frequently use Likert’s (1932) attitude scale. A five-item ‘national pride’ scale is shown in Figure 1. Each Likert item presents a favorable or unfavorable statement about the entity (the respondent’s own country), followed by ordered response categories ranging from agreement to disagreement. In practice, there are usually between three and seven response categories, and it is especially common to use the five categories shown in Figure 1. Respondents are asked to select the category that best reflects their reaction to the statement. Responses to items of this type will be referred to as Likert attitude data, or Likert data for short.

Although researchers may be interested in measuring attitudes for their own sake, they are often more interested in how certain background variables affect the attitudes. For in-
stance, sociologists Smith and Jarkko (2001) used cross-national responses to the national pride scale to investigate how the attitude differs between nations.

All methods of analysis involve (albeit sometimes implicitly) a model for the relationship between Likert data and the attitudes underlying them. This model should fulfill six requirements, the sixth being our focus here:

1. **The model should appropriately represent the attitudes being measured.** Eagly and Chaiken’s definition implies they can be represented as locations along an *attitude continuum*, which is an unobserved and underlying continuum running from extremely unfavorable to extremely favorable evaluations of the entity. Without loss of generality, we assume that the continuum’s positive end corresponds to extreme favorability.

2. **The model should appropriately represent the statements.** Studies on the processing of attitude statements suggest that statements about controversial entities are perceived in relation to the two ends of the attitude continuum (see Bohner 2001, p. 244). This implies that a statement can be represented as a location along the attitude continuum. Of course, this location should be consistent with the statement’s content.

3. **The model should appropriately represent the relationship between an individual’s...**
attitude and his responses. This relationship can be classified as cumulative or unfolding (Coombs 1964), depending on whether disagreement increases as the individual’s attitude becomes more distant from the statement’s location (unfolding) or as it becomes more negatively displaced from the statement’s location (cumulative). See Figure 2 for an illustration. Since individuals disagree with a statement when it is more favorable than their attitude or when it is less favorable, unfolding models are better suited to Likert data from a theoretical standpoint. They are also preferable from a practical standpoint because of their greater generality: Cumulative models are a special case of unfolding models with very extreme statements.

4. The model should be able to incorporate covariates.

5. The model should be appropriate for ordinal responses.

6. The model should accommodate different response styles.

Following Johnson (2003), we define response style as a consistent pattern of response category selection, independent of statement content. Many types of patterns have been described, but we focus on wide response range, a tendency to select the outer categories, and acquiescence, a tendency to select more agreeable categories (see Baumgartner and Steenkamp 2001 for other types). An individual’s response style is influenced by his or her personality, category interpretation, and response environment (e.g., distraction level when responding). Response style appears to differ considerably between individuals from
different cultures (e.g., Clarke 2000; Gerardo, Gamba and Marin 1992). In addition, it appears to differ considerably between genders: An analysis that applied the model introduced below to items measuring abortion attitudes found that acquiescence and response range differed significantly between men and women (Author 2004).

Research suggests that ignoring response-style differences can result in misleading inference (Baumgartner and Steenkamp 2001; Clarke 2000; Heide and Gronhaug 1992). As one example, suppose national pride is estimated from responses to the five items in Figure 1. Acquiescence biases the estimates upwards because 80% of the statements are favorable. As a second example, suppose we are interested in the correlation between national pride and immigration attitudes, both measured by Likert scales. When response style differs between individuals but not between scales, the correlation’s magnitude is exaggerated.

Research also suggests that the bias is more severe the more unbalanced the scale, i.e., the farther from equal its proportions of favorable and unfavorable statements. For example, if all five national pride items were favorable, acquiescence would bias attitude estimates even further upwards. Unfortunately, many frequently-used scales are very unbalanced, despite recommendations to the contrary (Baumgartner and Steenkamp 2001).

No existing model fulfills all of our six requirements, especially not the popular model proposed by Likert himself (1932). Likert’s model involves assigning ordered numerical scores to the response categories, and then estimating individuals’ attitudes by summing
their responses to all the items. As we discuss below, this model can only accommodate different response styles in an *ad hoc* manner, and it does not fulfill any of our other requirements. On the other hand, a variant of Johnson’s model (2003, pp. 576–578) does formally model response style and also fulfills most of our other requirements. However, the variant is cumulative rather than unfolding, and thus should not be used unless all statements are extreme.

In this paper, we introduce a new model that fulfills all six requirements. Most notably, our model allows the data to reflect both response style and attitudes. Our model can be used to estimate individuals’ attitudes and to fit general linear models for the effects of background variables on attitudes. Resulting inferences are adjusted for differences in response style.

Since response styles and attitudes are unobserved and may not be independent, we cannot entirely separate their effects on responses. In general, their separability depends on the data. Clearly, more items make separation easier: For example, attitude and response style would be hopelessly confounded with only one Likert item. In addition, balanced items make separation easier: For example, since acquiescence results in more agreeable responses to both favorable and unfavorable statements, it effectively cancels out when they are balanced. Finally, items measuring diverse and unrelated attitudes make separation easier because stylistic and substantive variance are less confounded (Baumgartner and Steenkamp 2001, p. 144). These guidelines suggest that researchers can improve sep-
arability by expanding their analysis to include items measuring attitudes not of primary interest. Conveniently, our model makes it simple to include additional items because it can easily allow multiple attitudes to underlie the data.

The rest of the paper is organized as follows. Section 2 introduces our model and explains how it can be combined with a general linear model for the effects of background variables on attitudes. Section 3 proposes a maximum (marginal) likelihood approach to fitting and also describes statistics for assessing global and local goodness of fit. Section 4 is an analysis of American and British responses to the national pride scale in Figure 1. Section 5 describes simulation experiments designed to compare inferences based on our model and Likert’s model. Finally, Section 6 discusses our model’s advantages and limitations.

2 A DESCRIPTION OF THE MODEL

2.1 Notation

It is necessary to introduce some notation before describing our model. We adopt some of Rossi et al.’s (2001) notation for the sake of consistency.

We observe a sample of \( n \) individuals, each with responses to \( J \) items. We let \( Y \) denote the \( n \times J \) data matrix of observed responses, \( Y_i \) denote the vector of \( J \) observed responses for person \( i \), and \( Y_{ij} \) denote person \( i \)’s observed response to item \( j \). \( Y_{ij} \) falls into one of \( K \)
ordered agreement categories, where \( K \) corresponds to the highest level of disagreement, without loss of generality. Note that all items are assumed to share the same response categories. For now, we also assume that all items measure the same single attitude.

### 2.2 The Basic Model

The Multidimensional Unfolding Model (MUM) combines a latent structure that represents one or more attitudes and a separate response structure that can accommodate differing response styles. The MUM’s formulation follows certain principles common to most latent variable models. First, it assumes independence for the individuals. In addition, it assumes ‘local independence.’ This means that the items are conditionally independent given the individual-specific parameters, which include the latent variable representing attitudes and possibly two individual-specific parameters describing response style.

The latent structure is a single dimension that represents the attitude continuum along which individuals and statements are located. The person location for individual \( i \) will be denoted \( \theta_i \), and the statement location for item \( j \) will be denoted \( \beta_j \). Note that the statement location does not depend on the individual, implying that the statement is perceived identically by all individuals.

The latent structure is connected to the response structure by assuming that \( Y_{ij} \) increases stochastically with \( d_{ij} \), the distance between \( \theta_i \) and \( \beta_j \). More formally, \( d_{ij} = |\theta_i - \beta_j| \).
The response structure is a mapping from the $d_{ij}$s to the joint distribution of the observed responses. This mapping occurs via the unobserved-variable-coarsened-by-thresholds approach, which is equivalent to the cumulative probability approach to modeling ordinal variables (see, for example, Agresti 2002, Section 7.2.3). Here, the unobserved response for individual $i$ on item $j$ is

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

where $\varepsilon_{ij}$ is assumed to have a logistic distribution with center 0 and scale equal to $1/1.7$ (alternatively, a normal distribution with mean 0 and variance 1); and $Y_{ij}^*$ represents an agreement continuum. This continuum is coarsened into ordered categories by a set of ordered thresholds:

$$-\infty = c_{i0}^i \leq c_{i1}^i \leq \ldots \leq c_{iK-1}^i \leq c_{iK}^i = \infty.$$  

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

The Item Category Curve (ICC) for category $k$ of item $j$ is then

$$P(Y_{ij} = k) = P(c_{k-1}^j \leq Y_{ij}^* \leq c_{k}^j) \quad (k = 1, \ldots, K)$$  

$$= P(c_{k-1}^j - |\theta_i - \beta_j| \leq \varepsilon_{ij} \leq c_{k}^j - |\theta_i - \beta_j|).$$
Finally, the likelihood for the data is

\[ L = \prod_{i=1}^{n} L(Y_i) = \prod_{i=1}^{n} \prod_{j=1}^{J} \prod_{k=1}^{K} P(Y_{ij} = k)^{I(Y_{ij}=k)}, \]  

(5)

with the first and second multiplications following from the assumed properties of individual independence and local independence, respectively.

2.3 Different Cases of the Response Structure

The thresholds in (2) can vary in numerous manners that we now describe. Our descriptions refer to exhaustive response style groups defined so that response style differs more between them than within them. The groups are indexed by \( g = 0, \ldots, G - 1 \), where \( G \) is the total number of groups and 0 is the reference group. As an example, we might use response style groups based on nationality, gender, or a combination.

Table 1 enumerates the response structure cases in terms of increasing generality. Case 1 assumes that the threshold set is the same for all individuals, implying that response style does not differ. Case 2 allows the set to vary between groups but assumes that it is the same for all individuals within a given group. Last, Case 3 allows the threshold set to vary between individuals. For reasons of computational convenience, this variation occurs in a restricted manner where thresholds vary only up to shifting and scaling parameters (Rossi et al. 2001; Wolfe and Firth 2002). More formally,

\[ c_k^i = (\sigma_i)^{-1} c_k + \tau_i \text{ for } k = 1, \ldots, K, \]  

(6)
Table 1. Summary of MUM response structure cases (and their parameters)

<table>
<thead>
<tr>
<th>Case</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>Thresholds are the same for all individuals.</td>
</tr>
<tr>
<td></td>
<td>{c_1, \ldots, c_{K-1}}</td>
</tr>
<tr>
<td>Case 2</td>
<td>Thresholds vary between groups in an unrestricted manner.</td>
</tr>
<tr>
<td></td>
<td>{c_{g1}^g, \ldots, c_{gK-1}^g : g = 0, \ldots, G - 1}</td>
</tr>
<tr>
<td>Case 3</td>
<td>Thresholds vary between individuals in a shifted and scaled manner.</td>
</tr>
<tr>
<td></td>
<td>{c_1, \ldots, c_{K-1}} and {\tau_i, \sigma_i : i = 1, \ldots, n}</td>
</tr>
</tbody>
</table>

- **a:** Common distribution of shifting and scaling parameters.  
  \{\tau_i, \ln(\sigma_i)\} = BVN(\varphi, \Lambda)

- **b:** Group-specific distributions of shifting and scaling parameters.  
  \{\tau_i, \ln(\sigma_i)\} \sim BVN(\varphi^g, \Lambda^g)  \text{ where } i \in g

where \{c_1, \ldots, c_{K-1}\} is the common threshold set; and \tau_i and \sigma_i are referred to as individual style parameters and are quantifications of acquiescence and wide response range, respectively.

### 2.4 A Multi-set Variant of the Latent Structure

We now consider situations where the \(J\) items do not all measure the same attitude. We assume that the items can be partitioned into \(S\) mutually exclusive and exhaustive item sets. Set \(I_s (s = 1, \ldots, S)\) contains all items measuring attitudes toward entity \(s\). (A scale is an item set with a small number of carefully chosen items.) For the sake of convenience, \(I_1\) will contain items \{1, \ldots, J_1\}, \(I_2\) will contain items \{\(J_1 + 1, \ldots, J_1 + J_2\}\), \ldots , and set \(S\) will contain items \{\(\sum_{s=1}^{S-1} J_s + 1, \ldots, J\}\). Further, \(S(j)\) will denote the index of the
set to which item \( j \) belongs, and the scalar \( \theta_{is} \) will refer to individual \( i \)'s attitude toward entity \( s \).

We can easily incorporate multiple sets into the MUM by allowing its latent space to be \( S \)-dimensional where dimension \( s \) represents the attitude continuum underlying \( I_s \). The space’s \( S \) axes will not necessarily be orthogonal since the \( S \) attitudes may be correlated. Individuals can be located anywhere in the space, but statements in \( I_s \) can only be located along the \( s^{th} \) axis. More formally, in this multi-set variant, \( \theta_i = [\theta_{i1} \theta_{i2} \ldots \theta_{iS}]^T \), and \( \beta_j = [0 \ldots \beta_{js} \ldots 0]^T \) where \( S(j) = s \). The (latent) distance, \( d_{ij} \), is the \( l_1 \)-distance between statement \( j \)'s location and the projection of individual \( i \)'s location onto dimension \( S(j) \). In other words, it is the absolute difference between the individual’s and statement’s \( S(j)^{th} \) elements:

\[
d_{ij} = |\theta_{iS(j)} - \beta_{jS(j)}|.
\]

Note that \( d_{ij} \) is not directly affected by individual \( i \)'s location along dimensions other than \( S(j) \). Note also that the multi-set variant reduces to the single-set variant presented in Section 2.2 when \( S = 1 \).

The multi-set variant enables us to simultaneously analyze multiple sets of primary interest. In addition, it allows us to incorporate ‘secondary’ sets into the analysis if the primary set contains limited information on response style. Effectively, the multi-set variant ‘borrows strength’ across sets in order to better measure response style. For example, Figure 3 depicts how a six-item immigration scale is used to augment the national pride
scale’s information on national differences in response style.

2.5 Combining the MUM with Regression Models

The MUM can be used to fit a general linear model for the effects of background variables on attitudes. We assume that the model takes the form

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

where $X_i$ is a length $P$ vector containing the background variables for individual $i$; $\gamma$ is a $P \times S$ matrix containing parameters that quantify how changes in the background variables affect the person location means; and $\delta_i$ is an error vector of length $S$. We assume that $\delta_i$ has a normal distribution with a mean vector that is 0 and a covariance matrix, $\Phi$, that is either equal to $I$ or stratified on the levels of any categorical background variables. Without loss of generality, the first row of $\gamma$, which corresponds to $\theta_{i1}$, will contain the parameters of interest.

The general linear model is embedded in the MUM, which amounts to setting $\theta_i$’s mean vector equal to $\gamma^T X_i$ and its covariance matrix to $\Phi$. Both models are fitted to the data simultaneously, rather than fitting the MUM in a first step and then using the resulting $\hat{\theta}_i$’s to fit the general linear model in a second step. We eschew the two-step approach because it results in more attenuated estimates (Zwinderman 1997, p. 245).

MUM-based inferences about $\gamma$ are adjusted for differences in response style, an im-
important adjustment since the background variables may affect Likert data not only through attitudes, but also through response styles.

2.6 Similar Existing Models

The MUM meets Luo’s (2001) definition of an ‘unfolding’ latent variable model for ordinal responses (see Appendix A). Other models in this class include the generalised hyperbolic cosine model (GHCM) (Andrich 1996; Rost and Luo 1997) and the generalized graded unfolding model (GGUM) (Roberts, Donoghue, and Laughlin 2000). However, these models cannot accommodate multiple attitudes or different response styles. Further, incorporating these extensions would be difficult because the models treat each response category as a combination of two unobserved categories (e.g., ‘Strongly disagree’ = ‘Strongly disagree b/c statement too pro-nation’ + ‘Strongly disagree b/c statement too anti-nation’).

The MUM’s latent structure resembles certain Multidimensional Scaling-like unfolding models for visualizing rank or ordinal data. (See Cox and Cox 2001, Chapter 8 or Borg and Groenen 1997, Chapter 14.) These non-probabilistic models seek to locate individuals and statements in a ‘latent’ space so that the distances between them best match the responses, which are viewed as dissimilarities between individuals and statements. Some models accommodate different response styles by comparing dissimilarities only within individuals; however, they are prone to degeneracy in practice (Borg and Groenen 1997,
Analogously, the MUM’s individual-specific response structure resembles certain cumulative latent variable models for ordinal responses (e.g., Johnson 2003, pp. 576–578; Shi and Lee 1998). These models employ continuous unobserved variables coarsened by individual-specific thresholds, but allow them to vary unrestrictedly.

3 Model Fitting

3.1 Identifiability, Estimation, and Inference

To fit the MUM, we suggest a maximum likelihood approach that treats any individual-level parameters as random effects, and all other parameters (and hyperparameters) as fixed effects. To simplify the model, the person locations and individual style parameters are assumed to be independently distributed. To reduce computational complexity, both distributions are assumed to be normal.

For the person locations, the density \( g_1(\theta_i) \) is (possibly multivariate) normal with mean \( \mu \) and variance \( \Sigma \), both of which may be fixed or estimated. Since Equation (4) shows that the person locations are additively confounded with the statement locations, identifiability constraints must be imposed. One option is to set the mean of \( g_1(\theta_i) \) to 0 for at least some individuals, which identifies the statement and person locations up to a change of sign. (The variance of \( \theta_i \) should also be set to a pre-specified value to eliminate any potential
confounding between the latent and response structures.) A second option is to set the statement locations to pre-specified values based on informed judgment or empirical investigation. This option must be chosen when $M$ is embedded in the MUM since $g_1(\theta_i)$’s hyperparameters are determined by equation (8) and estimated. Fixing the locations of the statements in $I_s$ identifies column $s$ of $\gamma$ in terms of magnitude and sign.

For Case 3 of the response structure, the density $g_2(\tau_i, \ln(\sigma_i))$ is assumed to be bivariate normal with mean $\varphi$ and variance $\Lambda$, both of which are estimated. These hyperparameters can be shared by all individuals (Case 3a) or can depend on the individual’s response style group (Case 3b), in which case they are referred to as $\varphi^g$ and $\Lambda^g$. Equation (6) reveals that the thresholds are additively confounded with the $\tau_i$s and multiplicatively confounded with the $\sigma_i$s. These problems are resolved by following Rossi et al. (2001) and 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 3b, the constraints $\varphi_1^g = 0$ and $\varphi_2^g = -\lambda_{2,2}^g$ are used only for group $g = 0$.

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. For the single-set variant, the marginal likelihood takes the form:
\[ ML = \prod_{i=1}^{n} \left\{ \int g_1(\theta_i) \cdot \prod_{j=1}^{J} \prod_{k=1}^{K} P(Y_{ij} = k)^I(Y_{ij} = k) \, d\theta_i \right\}, \quad (9) \]

or, for Cases 3a and 3b,

\[ ML = \prod_{i=1}^{n} \int \cdots \int \left\{ g_1(\theta_i) \cdot g_2(\tau_i, \ln(\sigma_i)) \cdot \prod_{j=1}^{J} \prod_{k=1}^{K} P(Y_{ij} = k)^I(Y_{ij} = k) \right\} \, d(\theta_i) \, d(\tau_i) \, d(\ln(\sigma_i)). \quad (10) \]

(Marginal) MLEs will be used even for any variance parameters that are estimated, although one might consider using an alternative and more complex method, like REML, that avoids under-estimation.

The inverted observed information matrix evaluated at the MLEs is used to estimate asymptotic standard errors for the MUM’s parameters and hyperparameters. In addition, the likelihood ratio test (LRT) is used to compare competing hypotheses about parameters of interest.

Finally, the person locations are 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.
3.2 Goodness of Fit

Assessing goodness of fit is difficult because the number of cells in the \( J \)-way contingency table is typically large, making sparsity a problem.

The maximum \( \ln(ML) \) value is used to assess overall goodness of fit. Since the standard asymptotic theory is not appropriate for ungrouped data, we interpret \( \ln(ML) \) as a logarithmic score (Good 1983), i.e., the sum of the negative log probabilities of the events that occurred. Thus, \( \exp \{ \ln(ML)/\left(n \times J\right) \} \) can be thought of as ‘the average predicted probability of the item-category that the individual actually selected.’

Signed Pearson residuals are used to assess how well the probabilities predicted by the MUM match the observed probabilities for the univariate and bivariate margins. The signed univariate Pearson residual for category \( k \) of item \( j \) is

\[
\chi^2_{j(k)} = \text{sign} \left( f_{j(k)} - \hat{p}_{j(k)} \right) \cdot \frac{n \left( f_{j(k)} - \hat{p}_{j(k)} \right)^2}{\hat{p}_{j(k)}},
\]

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

\[
\chi^2_{j(k)l(m)} = \text{sign} \left( f_{j(k)l(m)} - \hat{p}_{j(k)l(m)} \right) \cdot \frac{n \left( f_{j(k)l(m)} - \hat{p}_{j(k)l(m)} \right)^2}{\hat{p}_{j(k)l(m)}},
\]

where \( f_{j(k)l(m)} \) and \( \hat{p}_{j(k)l(m)} \) are the bivariate analogues of \( p_{j(k)} \) and \( \hat{p}_{j(k)} \). For the single-set variant with common or group-specific response structures (Cases 1 or 2), the expected
univariate and bivariate probabilities are calculated using

\[ \hat{p}_{j(k)} = \int P(\hat{c}_{k-1}^i - |\theta_i - \hat{\beta}_j| \leq \varepsilon_{ij} \leq \hat{c}_k^i - |\theta_i - \hat{\beta}_j|) g_1(\theta_i) d\theta_i \]

and

\[ \hat{p}_{j(k,l(m)} = \int P(\hat{c}_{k-1}^i - |\theta_i - \hat{\beta}_j| \leq \varepsilon_{ij} \leq \hat{c}_k^i - |\theta_i - \hat{\beta}_j|) \\
\int P(\hat{c}_{m-1}^i - |\theta_i - \hat{\beta}_l| \leq \varepsilon_{il} \leq \hat{c}_m^i - |\theta_i - \hat{\beta}_l|) g_1(\theta_i) d\theta_i, \]

respectively, where \( \hat{\beta}_j \) and \( \hat{c}_k \) are the MLEs for those parameters. (For Cases 3a and 3b, analogous expressions are used.) Since the model was not fitted using either the univariate or bivariate marginal frequencies, the residuals in (11) and (12) cannot be used to formally assess goodness of fit. Instead, we will use them to informally investigate how well the MUM predicts the observed frequencies for each item.

### 3.3 A Small Simulation Experiment

We performed a small simulation experiment designed to answer two questions about the MUM:

i. Should we model response style if we are primarily interested in statement order?

ii. What range of \( \ln(ML) \) values indicates a well-fitting model?

First, we used the single-set variant with an individual-specific response structure (Case 3a) to generate \( N = 100 \) datasets, each containing \( n = 140 \) people and \( J = 6 \)
items comprising a scale. Because scales often contain a small number of statements ranging from very to extremely unfavorable and the same number of statements ranging from very to extremely favorable, the statement locations were set to $\beta_1 = -4$, $\beta_2 = -3$, $\beta_3 = -2$, $\beta_4 = 2$, $\beta_5 = 3$, and $\beta_6 = 4$. For each person, $\theta_i$ was generated from a $N(0, 1)$ distribution. The response structure parameters were chosen based on experience with real data.

The single-set variant with a common response structure (Case 1) was fitted to each of the simulated datasets. The $\hat{\beta}_j$s ordered the statements correctly every time, suggesting that the MUM can recover the true statement order without modeling response style. For each dataset, we calculated $\exp\left\{\ln(ML)/(140 \cdot 6)\right\}$ and found that the average probability (that an individual selected the response category that he/she did) ranged between 0.25 and 0.30. Since Case 1 ignores response style, we would expect these values to be on the lower side. For this reason, we also fit the single-set variant with an individual-specific response structure (Case 3a) and found that the average predicted probabilities ranged between 0.30 and 0.34 for the datasets.

4 NATIONAL PRIDE EXAMPLE

To illustrate the MUM, we used the national pride scale from the 1995 National Identity Survey (NIS) to address the question, do Americans or the British exhibit more national pride after adjusting for differences in response style? Nationality (U.S. or G.B.) was used
to define the response style groups in the MUM’s response structure. Since the national pride items contained limited information on response style, the analysis also included six NIS items measuring immigration attitudes. The sample was limited to the 807 British and 998 American respondents who had no missing or ‘Don’t Know / Can’t Choose’ responses for the national pride and immigration items. Additional information on the NIS can be obtained at http://www.gesis.org/en/data_service/issp.

We began our investigation with Likert’s popular scoring model (1932). As described above, the model requires that each item in $I_s$ first be quantified by assigning its categories ordered numerical scores that run in opposite directions for favorable and unfavorable items. Then, $\hat{\theta}_i$ is the sum of the $J_s$ quantified responses for individual $i$; we refer to this sum as the total score. The scoring model is very easy to implement but fulfills none of our requirements. Some researchers do attempt to accommodate response style differences by normalizing each individual’s quantified responses before summing them. However, the mean and standard deviation estimates used to normalize the data do not have the desired statistical properties when $J_s$ is small. Further, this ad hoc approach assumes that the data are continuous and from an elliptically symmetric distribution (Rossi et al. 2001).

Nonetheless, we used the scoring model to make some preliminary comparisons of the American and British data. Figure 4 depicts the marginal distributions of $\hat{\theta}_{i1}$ (national pride) and $\hat{\theta}_{i2}$ (immigration attitude), by nation. The American and British national pride means were 2.08 and 0.34; a $t$-test assuming unequal variances yielded a p-value smaller
than 0.0001, suggesting that Americans have more national pride. The Pearson correlations between $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ were $-0.33$ and $-0.41$ for the American and British respondents, respectively.

Before using the MUM to formally examine the difference in British and American means, the statement locations had to be set to pre-specified values. We estimated them using the multi-set MUM variant on its own (with no embedded model), with the mean and variance of $\theta_i$ set to 0 and 1, respectively. First, we fit the model with a common response structure (Case 1) to the American data and British data separately to check that the statement locations had the same order for both groups. Second, we fitted the model with a group-specific response structure (Case 2) to the combined data. (Recall the simulation experiment in Section 3.3, where the true statement order was recovered without modeling response style at an individual-specific level.) Figure 5 presents the estimated statement locations. Reassuringly, they seem consistent with the statements’ content.

We embedded the following general linear model in the multi-set variant:

$$\theta_i = \gamma_0 + \gamma_1 X_i + \delta_i, \quad (13)$$

where $\theta_i$ contains individual $i$’s locations along the national pride and immigration attitude continua; $X_i$ is an indicator for being American; $\gamma_0$ contains the mean locations for the British; and $\gamma_1$ is the difference in mean locations for Americans and the British. Note that $\gamma_{1,1}$ is the parameter of primary interest. Last, $\delta_i$ is an error vector that comes from
a bivariate normal distribution with mean vector $\mathbf{0}$ and a covariance matrix, $\Phi$, that is allowed to differ for U.S. and G.B.

To find out whether the data were better explained by differences in attitudes alone or by differences in response style alone, we compared the fit of nation-specific $g_1(\theta_i)$ distributions combined with a common $g_2(\tau_i, \ln(\sigma_i))$ distribution to the fit of a common $g_1(\theta_i)$ distribution combined with nation-specific $g_2(\tau_i, \ln(\sigma_i))$ distributions. Both models have the same number of parameters, but their respective $-2\ln(ML)$ values were 49,284 and 49,412. This suggests that differences in American and British responses are better explained by national differences in national pride; however, the true explanation is probably a combination.

Next, we fitted nation-specific $g_1(\theta_i)$ distributions combined with each response structure case. Since Case 3b fit the best according to AIC, it will be our focus from now on. Estimates of its parameters and hyperparameters, as well as some estimated standard errors, can be seen in Table 2.

Comparing $\hat{\mu}_1^{US}$ to $\hat{\mu}_1^{GB}$ suggests that Americans exhibit more national pride. We used an LRT to test the hypothesis that $\gamma_{1,1} = \mu_1^{US} - \mu_1^{GB} = 0$. The value of $-2\ln(ML)$ was 49,400 when $\gamma_{1,1} = 0$, compared to 49,284 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 style.
### Table 2. Parameter estimates and errors from fitting Case 3b to the NIS data

<table>
<thead>
<tr>
<th></th>
<th>Estimate (Standard error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1(\theta_i) = BV N(\mu, \Sigma)$</td>
<td>$g_2(\tau_i, \ln(\sigma_i)) = BV N(\varphi, \Lambda)$</td>
</tr>
<tr>
<td>1 = National Pride</td>
<td>1 = Acquiescence</td>
</tr>
<tr>
<td>2 = Immigration</td>
<td>2 = Response Range</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
</tr>
<tr>
<td>G.B.</td>
<td></td>
</tr>
<tr>
<td>-0.38</td>
<td>-0.25</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>U.S.</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>-0.08</td>
</tr>
<tr>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

**NOTE:** $\hat{c}_1 = 1.39$, $\hat{c}_2 = 2.89$, $\hat{c}_3 = 3.88$, $\hat{c}_4 = 5.54$

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

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

Comparing some of $g_1(\theta_i)$'s other hyperparameters suggests that Americans have more uniform national pride, as well as more favorable and more uniform immigration attitudes. The estimated correlations suggest that favorable immigration attitudes accompany lower levels of national pride for individuals in each nation. These correlations were smaller than the (unreported) correlations estimated using a common response structure (Case 1). This result is consistent with our previous assertion that ignoring response style results in exaggerated correlations. We might then wonder why the correlations between total scores, which ignore response style differences, were only $-0.33$ and $-0.41$. The answer is that...
these correlations were calculated in the second step of a two-step approach, which caused attenuation that outweighed the exaggeration due to ignoring response style (Zwinderman 1997, p. 245).

Comparisons of \( g_2(\tau_i, \ln(\sigma_i)) \)'s hyperparameters suggest that Americans are significantly more acquiescent and have a significantly wider response range. Further, Americans are more uniform in terms of acquiescence than the British, but both are equally uniform in terms of wide response range.

The maximum \( \ln(ML) \) value for Case 3b was \(-24642\). This translates to an average predicted probability of \( \exp\left\{-24642/(1805 \cdot 11)\right\} = 0.29 \), which seems very reasonable since fitting the MUM to data simulated from the MUM produced probabilities not much larger.

We calculated univariate and bivariate residuals, by nation. Their values are omitted for the sake of brevity, but can be obtained from the author. According to both types of residuals, the model fits both nations’ S-8 margins the best. For the British, the model fits 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 fits 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.

Finally, we checked whether Case 3b produced national pride estimates that sensibly incorporated information on response style. Specifically, we compared two imaginary
respondents with responses \( Y_1 = [ A \ D \ A \ A \ A \ SA \ SD \ SD \ SA \ SA \ SD ] \) and \( Y_2 = [ A \ D \ A \ A \ A \ A \ D \ D \ A \ A \ D ] \). Although both individuals responded identically to the five national pride items, the estimate of \( \theta_i \) was 0.08 for the first respondent and 0.50 for the second respondent. Since the second respondent’s immigration responses reveal an aversion to using the outermost categories, his or her national pride responses seem a stronger indication of national pride. Without the immigration data and a model that accommodates different response styles, the estimated \( \theta_i \)'s would mislead us since they would be identical to each other.

5 Simulation Experiments

We conducted two sets of simulation experiments designed to mimic our analysis of the NIS data. In particular, we wanted to investigate the MUM’s performance when drawing inference about the difference in two groups’ mean locations (\( \gamma_{1,1} \) in Equation (13)).

We simulated data from the single-set MUM variant with an individual-specific response structure (Case 3). The values of the random-effects hyperparameters, which can be seen in Table 3, were set equal to the estimates obtained in Section 4. Five different situations (A-E) were investigated: In each, \( \gamma_{1,1} \) was assumed to be zero. In A, there were no differences in the two groups’ hyperparameters, but in each of B through E, there was
one type of difference in their hyperparameters. We were interested in how each situation would affect the results of testing $H_o : \gamma_{1,1} = 0$.

### 5.1 Set 1

The first set of experiments was designed to test the performance of Likert’s scoring model for three different scales. The first scale resembled the national pride scale: It used the five estimated statement locations (one negative and four positive) from the top half of Figure 5. The second scale resembled the immigration scale and used the six estimated statement locations (three positive, three negative) from the bottom half of Figure 5. The third scale also contained six statements, three with locations equal to -5 and three with locations
Table 4. Empirical $\alpha$ levels for an $\alpha = 0.05$ two group $t$-test based on total scores

<table>
<thead>
<tr>
<th>Situation</th>
<th>Unbalanced Scale</th>
<th>Balanced Scale</th>
<th>Extreme Balanced Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.045</td>
<td>0.056</td>
<td>0.055</td>
</tr>
<tr>
<td>B</td>
<td>0.047</td>
<td>0.049</td>
<td>0.053</td>
</tr>
<tr>
<td>C</td>
<td>1.000</td>
<td>0.068</td>
<td>0.056</td>
</tr>
<tr>
<td>D</td>
<td>1.000</td>
<td>0.065</td>
<td>0.056</td>
</tr>
<tr>
<td>E</td>
<td>0.080</td>
<td>0.054</td>
<td>0.051</td>
</tr>
</tbody>
</table>

equal to 5. This scale was designed to be ideal for use with Likert’s scoring model: Not only was it balanced, but it contained only extreme statements.

For each of the three scales, $N = 1000$ datasets were generated using each of the five hyperparameter sets. The resulting datasets contained 807 group 0 people and 998 group 1 people, to make them comparable to the real dataset in Section 4. For every simulated dataset, total scores were calculated; a $t$-test with unequal variances was used to compare the means of group 0 and group 1. Table 4 shows the percent of the 1000 $p$-values below 0.05 in each experiment.

These results suggest that Likert’s scoring model produces valid inferences for balanced scales. In fact, the model appears robust to all the group differences for the second scale and even more so, albeit not significantly, for the third scale. However, the model did not produce valid inferences for the unbalanced scale when there were group differences in acquiescence and wide response range. This suggests that Likert’s scoring model should not be used to analyze data from unbalanced scales in situations where response style is
expected to differ.

5.2 Set 2

The second set of experiments was designed to compare inferences based on Likert’s scoring model to those based on the MUM. Only the unbalanced national-pride-like scale was used since Likert’s scoring model seemed to perform well for balanced scales. $N = 10$ datasets with 81 group 0 people and 100 group 1 people were generated using each of the five hyperparameter sets. For each dataset, we employed three methods to test $H_o : \gamma_{1,1} = 0$. The first method was the $t$-test used in the first set of experiments. The second and third methods involved fitting the single-set MUM variant with (i) the mean of $g_1(\theta_i)$ constrained to be equal for both groups and (ii) the mean allowed to differ for the two groups; if (ii) fit better than (i) according to an LRT, then $H_o$ was rejected. Method two used a common response structure (Case 1), and method three used a group-specific response structure (Case 2). The two methods were implemented using an R function described in Appendix B.

Table 5 presents the ten (sorted) $p$-values for each of the three methods in each situation. As observed in the first set of experiments, Likert’s scoring model was not robust to group differences in acquiescence and wide response range. The same was true for method two, which also ignored response style. On the other hand, method three was robust to all the group differences tested, even acquiescence and wide response range, because it mod-
Table 5. Sorted $p$-values from three hypothesis-testing approaches

<table>
<thead>
<tr>
<th>Situation</th>
<th>Method</th>
<th>$p$-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.03 0.17 0.17 0.23 0.44 0.53 0.77 0.81 0.84 0.85</td>
</tr>
<tr>
<td>A</td>
<td>Case 1</td>
<td>0.13 0.31 0.42 0.45 0.63 0.67 0.81 0.83 0.88 0.94</td>
</tr>
<tr>
<td></td>
<td>Case 2</td>
<td>0.13 0.4 0.43 0.55 0.66 0.75 0.76 0.84 0.85 0.91</td>
</tr>
<tr>
<td></td>
<td>Scoring</td>
<td>0.05 0.21 0.25 0.29 0.63 0.66 0.7 0.71 0.89 0.96</td>
</tr>
<tr>
<td>B</td>
<td>Case 1</td>
<td>0.16 0.36 0.49 0.55 0.69 0.75 0.75 0.89 0.92 0.98</td>
</tr>
<tr>
<td></td>
<td>Case 2</td>
<td>0.19 0.43 0.51 0.56 0.59 0.73 0.75 0.82 0.93 0.97</td>
</tr>
<tr>
<td></td>
<td>Scoring</td>
<td>0.00 0.0 0.0 0.0 0.00 0.01 0.01 0.06 0.07 0.37</td>
</tr>
<tr>
<td>C</td>
<td>Case 1</td>
<td>0.00 0.01 0.01 0.02 0.03 0.05 0.06 0.11 0.18 0.47</td>
</tr>
<tr>
<td></td>
<td>Case 2</td>
<td>0.17 0.41 0.53 0.62 0.79 0.83 0.84 0.88 0.91 0.95</td>
</tr>
<tr>
<td></td>
<td>Scoring</td>
<td>0.00 0.0 0.0 0.0 0.00 0.01 0.01 0.03 0.04 0.04</td>
</tr>
<tr>
<td>D</td>
<td>Case 1</td>
<td>0.00 0.0 0.02 0.04 0.05 0.07 0.08 0.12 0.13 0.15</td>
</tr>
<tr>
<td></td>
<td>Case 2</td>
<td>0.22 0.43 0.52 0.57 0.61 0.75 0.89 0.91 0.92 0.94</td>
</tr>
<tr>
<td></td>
<td>Scoring</td>
<td>0.01 0.15 0.22 0.25 0.52 0.61 0.74 0.76 0.77 0.91</td>
</tr>
<tr>
<td>E</td>
<td>Case 1</td>
<td>0.08 0.36 0.37 0.49 0.65 0.67 0.75 0.81 0.86 0.94</td>
</tr>
<tr>
<td></td>
<td>Case 2</td>
<td>0.11 0.39 0.47 0.56 0.64 0.64 0.73 0.83 0.85 0.94</td>
</tr>
</tbody>
</table>

eled response style. These results suggest that, even with a highly restricted model for response style differences, the MUM yields valid inferences about mean attitudes.

6 CONCLUSIONS

The national pride example and simulation experiments indicate that the MUM performs extremely well when used to analyze Likert data. Like all models, it has shortcomings that could cause concern in some situations.

One concern is that Likert data often contain large numbers of ‘Don’t Know / Can’t
Choose’ responses. With maximum likelihood approaches, these responses can be treated as missing data. Although we chose to use the simple case deletion approach in our real data example, a more sophisticated approach to handling missing data could be adopted.

A second concern is the use of a restricted model for threshold variation between individuals in Case 3. Rossi et al. (2001) note that this restricted model cannot produce certain response styles (for example, one where ‘Agree’ and ‘Disagree’ are preferentially selected). Fortunately, there is evidence that modeling acquiescence and wide response range is sufficient to characterize response style (Wolfe and Firth 2002, p. 251). If we did want to allow unrestricted threshold variation between individuals, fitting would be easier using a Bayesian approach based on Markov Chain Monte Carlo (MCMC) methods. This would involve developing a Metropolis-within-Gibbs sampling algorithm (Gilks, Richardson, and Spiegelhalter 1996, p. 85) similar to the MCMC algorithms of Johnson (2003, pp. 568-569 and pp. 579-581), Wolfe and Firth (2002), Rossi et al. (2001) and Shi and Lee (1998).

However, even without modification, the MUM represents a significant improvement over existing models for Likert data. Sections 4 and 5 reveal that using group-specific or individual-specific response structures results in more correct inferences because they are adjusted for differences in response style. More specifically, estimates of differences in mean attitudes are not contaminated by differences in response style, and estimations of correlations between attitudes are not exaggerated. Further, using a formal probability
model like the MUM allows us to assess goodness of fit, giving us additional insight into the data and the individual items.

One might wonder whether it is always necessary to use group-specific or individual-specific response structures when analyzing Likert data. If we are interested in the statements, then there is some empirical evidence that it is acceptable to ignore response style. However, if we are primarily interested in individuals’ attitudes or how background variables affect them, then we should not ignore response style, especially when the items are not balanced.

The results in Sections 4 and 5 have implications for the design, collection, and analysis of Likert data. Ideally, researchers should balance the items measuring the attitude of interest. However, if this is not possible, they should also administer scales measuring other attitudes, preferably ones unrelated to the attitude of interest. Finally, researchers analyzing data from unbalanced scales should use a model that, like ours, accommodates response style differences. To better measure those differences, they should include the additional scales in the analysis.
A UNFOLDING LATENT VARIABLE MODELS FOR ORDINAL DATA

Luo (2001) developed a general framework for a class of probabilistic, unfolding, unidimensional latent variable models for ordered data. The models in this class must exhibit the following characteristic (Luo 2001, Theorem 1): The expected response function for item $j$ is a unimodal function of the latent variable, $\theta_i$, with the maximum at the statement location, $\beta_j$. We define the expected response function for item $j$ as

$$E(Y_{ij} | \theta_i) = \sum_{k=1}^{K} (K - k) \cdot P(Y_{ij} = k).$$  \hspace{1cm} (14)

**Theorem 1** In the single-set MUM variant, the expected response for item $j$ is a unimodal function of $\theta_i$, with the maximum occurring at $\theta_i = \beta_j$.

**Proof.** Using Equation (1),

$$E(Y_{ij} | \theta_i) = \sum_{k=1}^{K} (K - k) \cdot P(c_{i,k-1} \leq |\theta_i - \beta_j| \leq c_i^k - |\theta_i - \beta_j|).$$

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

$$= \sum_{k=1}^{K} P\left(\varepsilon_{ij} \leq c_i^k - |\theta_i - \beta_j|\right).$$
Each of the $K - 1$ terms decreases with the expression $c^i_k - |\theta_i - \beta_j|$. Since this expression is an upside-down-V-shaped function of $\theta_i$ with its maximum at $\beta_j$, each term in the summation is a unimodal function of $\theta_i$ with its maximum located at $\beta_j$. Summing them to obtain $E(Y_{ij} | \theta_i)$ thus results in a function that is unimodal in $\theta_i$ with the maximum located at $\beta_j$.

**B DETAILS OF FITTING THE MUM**

This appendix contains descriptions of the R code used to simultaneously fit a general linear model and the multi-set MUM variant with any of the response structure cases. To make our description more concrete, we refer to the general linear model and multi-set variant used in the national pride example.

The fixed-effects parameters and hyperparameters were estimated by maximizing the relevant marginal likelihood. For Cases 1 and 2, the marginal likelihood was

$$ML = \prod_{i=1}^{n} \left\{ \int \int g_1(\theta_i) \cdot \prod_{j=1}^{J} \prod_{k=1}^{K} I(Y_{ij} = k) P(Y_{ij} = k) d(\theta_{i1})d(\theta_{i2}) \right\}. \quad (15)$$

In our national pride application, for instance,

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

For Cases 3a and 3b, the Marginal Likelihood was
\[ ML = \prod_{i=1}^{n} \int \int \int \left\{ \frac{g_1(\theta_i) \cdot g_2(\tau_i, \ln(\sigma_i))}{\prod_{j=1}^{J} \prod_{k=1}^{K} I(Y_{ij} = k) P(Y_{ij} = k)} \right\} d(\theta_{i1}) d(\theta_{i2}) d(\tau_i) d(\ln(\sigma_i)). \] 

(16)

where, for Case 3a,

\[ g_2(\tau_i, \ln(\sigma_i)) = BV N(\varphi, \Lambda) \text{ with } \varphi_1 = 0 \text{ and } \varphi_2 = -\lambda_{2,2}, \]

and, for Case 3b,

\[ g_2(\tau_i, \ln(\sigma_i)) = \begin{cases} 
N(\varphi^{US}, \Lambda^{US}) & \text{if } g = 1 \\
N(\varphi^{GB}, \Lambda^{GB}) \text{ where } \varphi_{1GB} = 0 \text{ and } \varphi_{2GB} = -\lambda_{2,2} & \text{if } g = 0 
\end{cases}. \]

The \( n \) integrals in (15) or (16) were approximated using Monte Carlo integration. More specifically, with (15), the change-of-variables technique was used to transform the \( i^{th} \) integral into an integral with respect to \( \theta_i' \), where \( \theta_i' \) comes from a standard bivariate normal distribution. The transformed integral was then approximated using Monte Carlo integration implemented with \( N = 600 \) points generated from the standard bivariate normal distribution. For Cases 3a and 3b, we adopted an analogous approach to approximate each integral in (16), but employed \( N = 1000 \) points generated from a standard four-dimensional normal distribution. In general, we chose the largest value of \( N \) for which the code still ran reasonably quickly.

The logarithm of the relevant \( ML \) was maximized 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 $\Phi$ matrices (and, in Case 3a and 3b, the $\Lambda$ matrices) remained symmetric and positive definite throughout optimization, $\ln(ML)$ was maximized with respect to the elements of the matrices’ Choleski decompositions. Also, to ensure that the threshold set(s) remained ordered, $\ln(ML)$ was maximized with respect to the differences in the thresholds, subject to the constraints that these differences were positive.

To begin the optimization process, sensible starting values were used for the fixed-effects parameters. The starting values for any response structure parameters were based on experience with the model; the starting values for the statement locations (when estimated) were based on statement content; and the starting values for the mean vector and covariance matrix of $g_1(\theta_i)$ were assumed to be $\mathbf{0}$ and $\mathbf{I}$, respectively, for both nations. Since $\ln(ML)$ can have multiple modes, the optimization process was 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 were retained as estimates.

The final $\ln(ML)$ value was calculated by evaluating either (15) or (16) 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 $\hat{\phi}$s were calculated by taking the square root of the relevant diagonal elements of the inverse of the observed information matrix, evaluated at the MLEs of the fixed-effects parameters. 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 constraints on the thresholds were not active in any of the solutions found.

Last, the person locations were estimated with the thresholds and statement locations fixed at their MLEs. Thus, the conditional distribution of the person locations (given the data) could be factored into separate distributions for each person. Each person location was then estimated by the mean of its posterior distribution, with the integral approximated in the manner described above.

References


Figure 1. National Pride Scale. *This scale was used to collect data on respondents’ ‘national pride,’ i.e., their attitude toward their own country. The scale was included in the National Identity Survey administered in 1995 as part of the International Social Survey Programme.*
Figure 2. Unfolding Models Versus Cumulative Models. *The top and bottom rows depict characteristics of unfolding models and cumulative models, respectively. The two left-hand plots show the probability of selecting each response category as a function of attitude. The two right-hand plots show the expected response as a function of attitude, where expected response is defined as a weighted sum of the response category curves using weights $AS = 4, A = 3, N = 2, D = 1, DS = 0$. Note that disagreement is bidirectional in unfolding models, but unidirectional in cumulative models. Also note that cumulative models are a special case of unfolding models with extreme statements. To see this, cover the right side of each unfolding plot to remove individuals with attitudes more favorable than the statement; the result is identical to the corresponding cumulative plots.*
Figure 3. Graphical Illustration of Case 2 of the Multi-set MUM Variant. We are interested in national pride, but include the immigration scale to increase the information on response style. Note that the thresholds differ by nation, but are the same for all eleven items. The arrows illustrate how the model works for a very proud Briton responding to statement 5. The individual’s location is $\theta_i$, and the statement’s location is $\beta_5$. The distance between them, $d_{i5}$, is combined with the error $\epsilon_{i5}$ to produce the continuous unobserved response, $Y^*_i$. Since $Y^*_i$ falls between $c^1_{GB}$ and $c^2_{GB}$, the observed response is $Y_{i5} = \text{‘Agree’}$. The observed response would be ‘Agree strongly’ for an otherwise identical American.
Figure 4. Total Scores for the National Pride and Immigration Scales, by Nation. The total scores were calculated using the category scores $K, \ldots, 1$ for favorable items and $1, \ldots, K$ for unfavorable items. The left-hand plots reveal that the distribution of national pride scores has roughly the same spread and a normal shape for both nations, but appears to be shifted more to the left for the British. Similarly, the right-hand plots reveal that the distribution of immigration scores has roughly the same spread and roughly normal shape for both nations, but also seems to be shifted more to the left for the British.
Figure 5. Statement Locations for the National Pride and Immigration Items. The statement locations were estimated by fitting Case 2 of the single-set MUM to the combined British and American data, with the same mean vector \((\mu = 0)\) and variance vector \((\sigma^2 = 1)\) used in \(g_1(\theta_i)\) for both nations.