Exploring monotonicity in polytomous item response data

Brian W. Junker
Department of Statistics
232 Baker Hall
Carnegie Mellon University
Pittsburgh PA 15213
(412) 268-8873
brian@stat.cmu.edu

In many psychological and educational settings it is natural or desirable to assume that the responses are ordered, but it is not always obvious what the correct order of responses is; examples from three psychological studies are given in the paper. General frameworks for ordered polytomous item response models have been studied recently by Holland and Rosenbaum (1986); Junker (1991); Junker and Ellis (1995); and Hemker, Sijtsma, Molenaar and Junker (1995a,b), among others; the principal condition of the models that constrains the order of the item responses is monotonicity (M) of the cumulative response functions $P[X_j > s|\theta]$. The goal of this study is to illustrate exploratory data analysis methods for (a) checking whether or not the labelling of item responses is compatible with M; and (b) discovering a labelling of item responses that is compatible with M.

Paper presented at the Annual Meeting of the National Council on Measurement in Education, April 10, 1996, New York. Research supported by the National Science Foundation, Grant DMS-94.04438.

1 Introduction

In many psychological and educational settings it is natural or desirable to assume that the responses are ordered, but it is not always obvious what the correct order of responses is. In this paper we consider three such examples: a test of transitive reasoning in children in which the order of the categories is naturally suggested by the logic of the items and the scoring rubric; a personality self-report assessment in which items are scored on a 5-point Likert scale in which the order indicated by the Likert scale on some items may disagree with the personality latent variable being measured; and a test measuring the effects of working memory (over-)load on errors in a series of cognitive tasks in which the item categories seem to be partially ordered at best. The general questions motivating this study can be phrased as follows: (1) can we confirm, nonparametrically, that the order of item responses in the transitive reasoning scale is compatible with a monotone unidimensional IRT model; and (2) can we select, nonparametrically, an order of item responses in the personality and working memory tests that make it possible to model the data with a monotone unidimensional IRT model? These are essentially questions about *exploratory data analysis* for montonocity in polytomous item response data.

General frameworks for ordered polytomous item response models have been studied recently by Holland and Rosenbaum (1986); Junker (1991); Junker and Ellis (1995); and Hemker, Sijtsma, Molenaar and Junker (1995a,b), among others. The principal condition of the models that constrains the order of the item responses is monotonicity (M), or stochastic ordering, of the item responses with respect to the latent variable in the model. In this paper we consider two ways of expressing monotonicity in IRT models for polytomous response variables X_j in terms of a unidimensional latent variable θ:

- the item step response functions (ISRF's) $P^*_{js}(\theta) = P[X_j > s|\theta]$ are increasing in θ;
- the item category response functions (ICRF's) $P_{js}(\theta) = P[X_j = s|\theta]$ satisfy the monotone likelihood ratio property that all ratios $P_{j(s+1)}(\theta)/P_{js}(\theta)$ are nondecreasing in θ.

Hemker et al. (1995a,b) call unidimensional IRT models satisfying local independence and the first condition here *nonparametric graded response models* (np-GRM's), since they nonparametrically generalize Samejima's (1969) parametric graded response model. They call unidimensional IRT models satisfying local independence and the second condition above *nonparametric partial credit models* (np-PCM's), since they nonparametrically generalize Masters' (1982) parametric partial credit model. It can be shown (Hemker, et al., 1995b) that the second monotonicity condition implies the first (every np-PCM is an np-GRM).

The first goal of this study is to illustrate the use of Guttman scaling and optimal scoring—and some variations on these classical techniques—to the problem of ordering or labelling the categories of polytomous items to try to achieve ISRF's that are compatible with the np-GRM. The second goal is to illustrate a new method of sorting or re-labelling item categories that is currently under development to try to achieve ICRF's that are compatible with the np-PCM.

In Section 2 we set up notation and a framework for discussing polytomous IRT models satisfying one or the other monotonicity condition above, and we briefly discuss some recent theoretical work on these models. In Section 3 we introduce the data sets that will be used for illustration in the paper. In Section 4 we sketch the theory of the methods we will consider, and in Section 5 we show how the methods can be used to check one or the other of the models above. Finally we discuss our results and suggest future work in Section 6.

2 Background

2.1 IRT model

Let X_1, X_2, ..., X_J be polytomous item response variables that can take m_1, m_2, ..., m_J discrete values, respectively. Our focus is on items with ordered categories, so without loss of generality we suppose the categories for X_j are labelled 1, 2, ..., m_j, or perhaps 0, 1, ..., $m_j - 1$; which labelling is being used will be clear from context.

The usual unidimensional item response theory (IRT) models for polytomous responses have the form

$$P[X_1 = x_1, \ldots, X_J = x_J] = \int P[X_1 = x_1, \ldots, X_J = x_J | \theta] \, dF(\theta) \qquad (1)$$

where we assume

Unidimensionality ($\boldsymbol{d} = 1$): θ takes values on the real line; and

Local Independence (LI):

$$P[X_1 = x_1, \ldots, X_J = x_J | \theta] = \prod_{j=1}^{J} P[X_j = x_j | \theta].$$

The item category response curves (ICRF's)

$$P_{js}(\theta) = P[X_j = s|\theta]$$

might assume nonparametric or parametric forms, including Samejima's (1969) Graded Response Model, Bock's (1972) Nominal Model, or Masters' (1982) Partial Credit Model.

When the responses can be ordered in some way it is also natural to assume

Monotonicity (M): The item step response function (ISRF)

$$P^*_{js}(\theta) = P[X_j > s|\theta] = \sum_{x=(s+1)}^{m_j} P_{jx}(\theta) \text{ is nondecreasing in } \theta \text{ for all } s.$$

For discrete polytomous responses it is equivalent to consider the monotonicity of $P[X_j > s|\theta]$ or $P[X_j \geq s|\theta]$ (the difference being whether category s is included in the probability or not). For mathematical derivation it is sometimes easier to work with the former function, and in computer programming it is often easier to work with the latter. In this paper we will work with the two functions interchangeably, the difference being clear from context and notation.

The model described by $d = 1$, LI and M, has been studied under many names in the literature. Mokken (1971) studies binary versions of the model under the name monotone homogeneity; Holland and Rosenbaum (1986) call it the monotone unidimensional latent trait model; Junker (1991, 1993) calls it the strictly unidimensional model.

Hemker et al. (1995a,b) study this model under the name nonparametric graded response model (np-GRM), by analogy with Samejima's (1969) graded response model, and relate it to a nonparametric version of Masters' (1982) partial credit model. Their nonparametric partial credit model (np-PCM) replaces condition M with the assumption that

$$P[X_j = s+1|\theta, \text{ and } X_j \in \{s, s+1\}] = \frac{P_{j(s+1)}(\theta)}{P_{js}(\theta) + P_{j(s+1)}(\theta)} \text{ is nondecreasing in } \theta \text{ for all } s.$$

[In Masters' (1982) model this quantity is modeled like a binary Rasch item.] This monotonicity assumption is easily seen to be equivalent to the monotone likelihood ratio property

Monotone Item Likelihood Ratio (MILR):

$$\frac{P_{j(s+1)}(\theta)}{P_{js}(\theta)} \text{ is nondecreasing in } \theta \text{ for all } s.$$

2.2 Some current theoretical results

The general class of models satisfying $d = 1$, LI and M have been studied recently by Holland and Rosenbaum (1986); Junker (1991, 1993); Junker and Ellis (1995); Hemker, Sijtsma, Molenaar and Junker (1995a,b); and Sijtsma and Junker (1996); among others.

Holland and Rosenbaum (1986) studied the properties of the marginal distribution of the data—the "manifest distribution"—implied by the $d = 1$, LI, M model described above. A principal finding was that all such models lead to *conditional association*, which states that for any partitioning of the test items $X_1, \ldots, X_n$ into disjoint subsets $\boldsymbol{Y} = (Y_1, \ldots, Y_{J1})$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_{J2})$, and any nondecreasing summary functions $f(\cdot)$ and $g(\cdot)$,

$$\mathrm{Cov}\,(f(\boldsymbol{Y}), g(\boldsymbol{Y})|h(\boldsymbol{Z}) = c) \geq 0$$

no matter what conditioning function $h(\boldsymbol{Z})$ or fixed constant c is chosen.

Junker (1991) studied a similar polytomous model under the assumption that Stout's (1990) *essential independence* condition holds, rather than LI as in Section 2.1. Junker (1993) studied the binary IRT model under precisely the assumptions of Section 2.1 and obtained a partial characterization of such models, in terms of their manifest distributions.

Junker and Ellis (1995) obtain a complete characterization of sequences of binary, polytomous or continuous items satisfying $d = 1$, LI and M together with the condition that θ be consistently estimatable from the data. Junker and Ellis show that such sequences are characterized by two conditions on the manifest distributions: (1) conditional association; and (2) a *vanishing conditional dependence condition*, i.e. the conditional covariances

$$\mathrm{Cov}\,(f(\boldsymbol{Y}), g(\boldsymbol{Z})|X_{n+1}, \ldots, X_{n+m})$$

tend to zero as m grows, where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are defined as in the definition of conditional association above, for any fixed n and any functions f and g whatsoever. Roughly speaking, *conditional association* guarantees that (possibly polytomous) item response variables are strongly monotone functions of one another, and *vanishing conditional dependence* guarantees that an estimatable latent trait θ underlies the monotone relationship among the items.

There have been other attacks on the characterization question as well. Important examples include Ellis and van den Wollenberg (1993), who characterize a stronger version of the model in Section 2.1 for binary items, using only a condition like conditional association; and Scheiblechner (1995; see also Junker, 1996), who partially characterizes sets of items that satisfy certain latent comonotonicity conditions related to the condition M above. General results of the sorts treated by Junker (1993) and Junker and Ellis (1995) are attractive theoretically but do not easily translate into applicable methodology (however, see Chapter 5 of Ellis, 1994, for a proposed methodology to check conditional association).

Hemker et al. (1995b) show that propery MILR implies property M for each item separately; hence every nonparametric partial credit model is also a nonparameteric graded response model. They also develop a taxonomy of ordered-response polytomous IRT models that includes both parametric and nonparametric versions of the graded response and partial credit models. Within this taxonomy they study the inferential properties of models satisfying M or MILR—whether, for example, a higher response category always increases the probability that the subject has a higher θ value.

3 Some Interesting Examples

In this section we describe briefly three data sets that will be used to illustrate the methods. In addition we will also use an artificial data set of simulated responses from Masters' (1982) partial credit model.

3.1 Transitive reasoning

Sijtsma and Verweij (1992) consider a nine-item subset of a 12-item scale of transitivity tasks obtained from $n = 425$ pupils; these items measure the ability to reason about transitive relations with respect to length, weight and size. There are 12 items but items 7, 9 and 12 are not believed to scale well with the other items (see Sijtsma and Verweij, 1992, for details; and Sijtsma and Junker, 1996, for an alternative analysis). The items were scored polytomously, so that

$$X_j = \begin{cases} 0, & \text{if judgement of transitive relation incorrect;} \\ 1, & \text{if judgement of transitive relation correct;} \\ 2, & \text{if judgement correct, and pupil gives a correct verbal} \\ & \text{justification for the judgement.} \end{cases}$$

Here the scoring rule clearly indicates the correct order in which to label response categories. By this definition of scoring, if the items measure a unidimensional "transitive reasoning" construct, then they should at least satisfy the np-GRM ($d = 1$, LI and M in Section 2.1).

They may also satisfy the more constrained np-PCM (replacing M with MILR in Section 2.1). Thus we expect almost any method of determining the best ordering of responses to agree with the ordering suggested by these labels, up to random error.

3.2 Personality self-report

The Inventory of Interpersonal Problems (IIP) is a standard 127-item self-report form that is often administered as part of a psychiatric diagnosis. Each item is a statement about how one feels or how one perceives one's environment, accompanied by a five-point Likert scale (0 = most disagree, ..., 4 = most agree). The intention of the developers of the IIP is that the succesive points on the Likert scale indicate succesively more evidence of various personality disorders. Data from $n = 97$ depressed patients in a study of the relationship between personality disorder and clinical depression at Western Psychiatric Institute and Clinic in Pittsburgh PA will be considered.

Although the order of response categories seems clear from the Likert scale, there is at least anecdotal evidence that this natural order is not always most indicative of level of personality disorder. For example, compare the following two items[1] (intended for response on a 5-point (0–4) Likert scale):

1. *I cannot be happy unless most people I know admire me* .... ____
2. *If I don't set the highest standards for myself, I am likely to end up a second-rate person* .... ____

While it may be argued that strong disagreement (0) with the first item is healthy, and strong agreement (4) is unhealthy, the second item is a less clearcut. Indeed, it may be that a neutral response (2) is healthiest, whereas strong agreement (0) and strong disagreement (4) might both represent less successful personality types.

3.3 Working memory load in audio menus

Patz, Junker, Lerch and Huguenard (1994) consider polytomous response data collected from a small-scale experiment in cognitive psychology, in which $n = 87$ undergraduates were asked to navigate a tree-structured audio menu system designed for use by a college to provide

[1] These items are not actually on the IIP but they are similar to IIP items.

automated course registration and campus information services by telephone. The students were asked to perform 27 registration and information-gathering tasks of varying difficulty. The tasks were scored

$$X_j = \begin{cases} 5, & \text{if the subject completed the task efficiently with no reminders;} \\ 4, & \text{if the subject completed the task but did not take the most efficient path in the menu;} \\ 3, & \text{if the subject completed the task but had to be reminded of the task description;} \\ 2, & \text{if the subject completed the task but both errors 3 and 4 were made;} \\ 1, & \text{if the subject never completed the task.} \end{cases}$$

Here some response categories (1, 2, 5) have apparently clear order relations and others (3, 4) do not. We are interested in seeing whether our methods will be useful to confirm the order relations we think we see and to discover whether the data suggest an order relation for the remaining categories.

3.4 Simulated PCM data

To illustrate some of the features of our methods in a "pure" context, we also simulated item and person parameters, and item responses for a set of $J = 10$ three-category items satisfying Masters' (1982) partial credit model, and $n = 400$ subjects.

The trait θ was simulated from a $N(0, 1)$ distribution; and the item category parameters δ_{js} for each category s of each item X_j were simulated as follows:

$$\delta_{js} = \beta_j + \epsilon_{js}$$

where β_j was drawn from a $N(0, 1)$ distribution and ϵ_{js} was drawn from a $N(0, (0.1)^2)$ distribution, and all draws were independent of one another.

The item responses for each simulated subject were then drawn using the ICRF's

$$P_{js}(\theta) = P[X_j = s|\theta] = \frac{\exp[\sum_{x=0}^{s-1}(\theta - \delta_{jx})]}{\sum_{t=0}^{m_j-1} \exp[\sum_{x=0}^{t-1}(\theta - \delta_{jx})]}$$

(where, for the simulated data set, $m_j = 3$ for all items, and a sum whose upper index of summation is less than the lower index of summation evaluates to zero). Finally the category labels 0, 1, 2 were randomly permuted for each item separately.

4 Theory and Methods

4.1 Stochastic ordering and monotone likelihood ratio properties

It is easy to see that if the items are dichotomized at arbitrary levels of response, a non-parametric graded response model for polytomous data becomes a monotone unidimensional IRT model for the resulting dichotomous data—that is, if $d = 1$, LI and M hold for the polytomous items, the same three assumptions will hold for the dichotomized items. Using this idea, one can deduce a "manifest monotonicity" property for polytomous items that is analogous to property MM of Junker (1993), namely

$$P[X_i \geq r | S = s] \text{ is nondecreasing in } s \text{ for each } r, \tag{2}$$

where S is a score calculated from all the dichotomized items not including X_j, i.e. S is a "rest score" (in practice, as shown by Ramsay, 1991, for example, it is usually not necessary to omit X_j if the test is long; however the computation of rest scores is not difficult and the storage of rest scores requires roughly the same amount of memory as storage of the original subjects-by-items raw item response matrix). This is a stochastic ordering property, which can be checked for various permutations of the response labels of the items. If one is found for which the inequalities hold, this can be considered an ordering of responses consistent with M.

The nonparametric partial credit model is characterized by $d = 1$, LI and MILR. Using an argument similar to that of Junker (1993, Lemma 4.1), one can show that these assumptions imply the MLR-like property

$$\frac{P[X_i = r+1 | X_j = s]}{P[X_i = r | X_j = s]} \leq \frac{P[X_i = r+1 | X_j = s+1]}{P[X_i = r | X_j = s+1]} \text{ for all } r \text{ and } s, \tag{3}$$

for pairs of items. This property is easy to check in the $m_i \times m_j$ contingency table cross-classifying responses to X_i and X_j: it is equivalent to finding positive association in the 2×2 subtable

		X_j	
		s	$s+1$
X_i	r	n_{rs}	$n_{r(s+1)}$
	$r+1$	$n_{(r+1)s}$	$n_{(r+1)(s+1)}$

If the association is negative, then either the rows or the columns may be permuted to achieve (3). If for some labelling of response categories, all such subtables (for all i, j, r and s) exhibit positive association, then this labelling is consistent with MILR.

4.2 Nonparametric regression

A simple way to examine the condition M is by using nonparametric regression methods such as those proposed by Ramsay (1991) and Abrahamowicz and Ramsay (1992). Briefly, an estimate $\hat{\theta}$ of the latent trait is obtained (usually a monotone transformation of total score; this is ordinally consistent for θ under a wide variety of conditions; see Stout, 1990 or Junker, 1991) and then $P^*_{js}(\theta)$ is estimated by kernel-based or spline-based logistic regression of the observed frequencies of scoring s or better on item j onto the estimates $\hat{\theta}$. For purposes of illustration we use S-PLUS's `gam` function to fit cubic B-splines as nonparametric regression estimators of $\log[P^*_{js}(\hat{\theta})/(1 - P^*_{js}(\hat{\theta}))]$ as functions of $\hat{\theta}$, for each j and s. This method can be used to visually examine the criterion (2) for example.

4.3 Mokken's scaling coeffients

A classical scaling method that continues to receive much modern attention is based on Loevinger's scaling coefficients, as developed by Mokken and his co-workers and students (e.g. Mokken, 1971; Mokken and Lewis, 1982; or Sijtsma and Verweij, 1992). These coefficients have been recently extended to the case of polytomous item response data by Molenaar (1991). The coefficients H_{ij} (for scaling a pair of items), H_i (for fit of an item i to an existing scale of items) and H (for overall scaling of a set of items) are defined as follows:

$$\begin{aligned} H_{ij} &= \frac{\text{Cov}\,(X_i, X_j)}{\text{Cov}_{\,max}(X_i, X_j)}\ ; \\ H_i &= \frac{\sum_{j\neq i} \text{Cov}\,(X_i, X_j)}{\sum_{j\neq i} \text{Cov}_{\,max}(X_i, X_j)}\ ;\ \text{and} \\ H &= \frac{\sum\sum_{i\neq j} \text{Cov}\,(X_i, X_j)}{\sum\sum_{i\neq j} \text{Cov}_{\,max}(X_i, X_j)}\ ; \end{aligned}$$

where in each case $\text{Cov}_{\,max}(X_i, X_j)$ is the maximum possible covariance between items with individual item category response frequencies fixed to be the same as those observed for X_i and X_j.

When H is large ($H > 0.3$ is a typical criterion), the items are thought to fit the non-parametric graded response model of Section 2.1 well. Maximizing H maximizes total score variance with individual item response category frequencies fixed. H_i is a scaled item-total covariance, used as a kind of item misfit statistic. Maximizing H_{ij} maximizes linear relationship between X_i and X_j with individual item response category frequencies fixed.

[Mokken's scaling coefficients provide an easy-to-compute measure of fit to the np-GRM. In the future it will also be useful to consider other measures such as Stout's T statistic (Stout, 1990; Li and Stout, 1995), which in contrast to H explicitly accounts for the role of the latent trait as well as nuisance dimensions in assessing model fit.]

4.4 Optimal scoring

The goal of optimal scoring is to find category scores for two items that maximizes the correlation between the items (Basilevsky, 1994, pp. 540 ff.). Letting $\boldsymbol{N}$ represent the two-way contingency table arising from cross-classifying the categories $s = 0, \ldots, m_i - 1$ of item X_i with the categories $s = 0, \ldots, m_j - 1$ of item X_j, and letting $\boldsymbol{N}_i$ and $\boldsymbol{N}_j$ be diagonal matrices whose diagonal entries are the observed frequencies of response in each category for items i and j, it can be shown that it is enough to maximize (in the vectors $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$)

$$\rho_{ij} = \boldsymbol{S}_i^T \boldsymbol{N} \boldsymbol{S}_j$$

subject to the constraints

$$\begin{aligned} \boldsymbol{S}_i^T \boldsymbol{N}_i \boldsymbol{S}_i &= n \text{ , and} \\ \boldsymbol{S}_j^T \boldsymbol{N}_j \boldsymbol{S}_j &= n \end{aligned}$$

This is a Lagrange multipliers problem, and the solution comes from the second eigenvalue/eigenvector pairs of two appropriately-formulated principal components decompositions (the first eigenvalue/eigenvector pair corresponds to centering so that ρ_{ij} becomes a true correlation). In this case, $\boldsymbol{S}_i$ is the second eigenvector of the square matrix $\boldsymbol{N}_i^{-1} \boldsymbol{N} \boldsymbol{N}_j^{-1} \boldsymbol{N}^T$, and $\boldsymbol{S}_j$ is the second eigenvector of the square matrix $\boldsymbol{N}_j^{-1} \boldsymbol{N}^T \boldsymbol{N}_i^{-1} \boldsymbol{N}$, scaled to unit length, say. The optimal scores for the categories of X_i are listed as successive elements of $\boldsymbol{S}_i$, and similarly for X_j and $\boldsymbol{S}_j$.

Increasing the correlation between X_i and X_j by scoring item categories with the optimal scores should increase the degree to which the re-scored X_i and X_j are both monotonically driven by a single latent variable θ. In our context we do not want to use the optimal scores directly, but rather sort the item categories according to the optimal scores, and then relabel the sorted categories with integers $r = 0, \ldots, m_i - 1$ and $s = 0, \ldots, m_j - 1$ again.

To see that this re-ordering of integer categories should also increase the degree to which the re-ordered items X_i and X_j are driven by a single latent variable, consider replacing the raw counts n_{rs} in the original two-way table $\boldsymbol{N}$ with the entries

$$a_{ij} = \begin{cases} 1, & \text{if the standardized residual } (n_{ij} - n_{i\cdot}n_{\cdot j}/n)/\sqrt{n_{i\cdot}n_{\cdot j}/n} \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The resulting matrix $\boldsymbol{A}$ is a kind of adjacency matrix for the categories of X_i and X_j. Positive association in the table $\boldsymbol{N}$ corresponds to 1's (positive standardized residuals) lining up on or near the main diagonal of $\boldsymbol{A}$ and 0's (negative standardized residuals) filling out the off diagonal elements of $\boldsymbol{A}$.

Indeed, the categories of X_i and X_j can also be re-ordered by sorting on the optimal scores computed from the adjacency matrix $\boldsymbol{A}$. This corresponds to the use of optimal scoring for seriation problems in archaeology (and other fields; see Basilevsky, 1994, pp. 570ff.), in which time-order (which plays the role of latent variable) of archaeological finds is deduced from the optimal scores of an incidence matrix relating objects found to their locations or depths in the dig. Sorting on the optimal scores tends to bring 1's to the main diagonal and 0's away from the main diagonal (increasing correlation between the optimal scores for the incidence matrix).

Optimal scoring inherently deals with two items at a time, and thus it is possible to arrive at different orderings for the categories of X_1, say, depending on whether it is cross-classified with X_2 or X_3. Because of this deficiency, optimal scoring will not be directly considered in the examples below. Guttman scaling, which is described next, treats all items at once and alleviates this problem.

4.5 Guttman scaling

The goal of Guttman scaling is to find category scores for each category of all items simultaneously to maximize the total-score variance on the exam. This problem can be formulated (see Basilevsky, 1994, pp. 534 ff.) as wishing to maximize (in the vector $\boldsymbol{u}$)

$$\nu \equiv \boldsymbol{u}^T \boldsymbol{Y}^T \boldsymbol{Y} \boldsymbol{u}$$

subject to the constraint

$$\boldsymbol{u}^T \boldsymbol{D} \boldsymbol{u} = 1$$

where $\boldsymbol{Y}$ is the usual raw response data matrix with persons in rows and item responses coded as dummy variables in columns, and $\boldsymbol{D}$ is an inner-product matrix; $\boldsymbol{D}$ is usually taken to be the diagonal matrix whose main diagonal is the same as $\boldsymbol{Y}^T\boldsymbol{Y}$. This maximization can be carried out with Lagrange multipliers, and is easily seen to be a principal components problem for $\boldsymbol{D}^{-1/2}\boldsymbol{Y}^T\boldsymbol{Y}\boldsymbol{D}^{-1/2}$; if the data in Y is not centered then the first eigenvalue will be 1 (corresponding to centering) and the second eigenvalue is the maximum value ν, with associated eigenvector $\boldsymbol{u}$, scaled to unit length. The Guttman scores for each category of each item will be listed as successive elements of the vector $\boldsymbol{u}$.

Maximizing the total score variance will approximately maximize the covariance between items also. Again we do not wish to use the the Guttman category scores $\boldsymbol{u}$ directly, but rather we will use their values to sort the categories within items, and then relabel the categories with integer labels, to obtain approximately maximum covariances between pairs of items. This sorting and relabeling should increase the extent to which the items can be modelled as montonically driven by a single latent variable.

To see a potential defect of Guttman scaling for this purpose, consider more carefully the matrix $\boldsymbol{Y}^T\boldsymbol{Y}$. Let $\boldsymbol{N}_{ij}$ be the two-way table of counts cross-classifying (original) categories of X_i with those of X_j, and let $\boldsymbol{N}_i$ be the diagonal matrix with frequencies of response to each category of X_i in the main diagonal. Then it is easy to see that

$$\boldsymbol{Y}^T\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{N}_1 & \boldsymbol{N}_{12} & \cdots & \boldsymbol{N}_{1J} \\ \boldsymbol{N}_{12}^T & \boldsymbol{N}_2 & \cdots & \boldsymbol{N}_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{N}_{1J}^T & \boldsymbol{N}_{2J}^T & \cdots & \boldsymbol{N}_J \end{bmatrix}$$

The maximum total score variance is achieved by selecting category scores that simultaneously maximize covariance between items (corresponding to the submatrices $\boldsymbol{N}_{ij}$) and maximize variance within items (corresponding to the blocks $\boldsymbol{N}_i$ on the main diagonal). For our purposes, it would be bad if the maximum total score variance is achieved by maximizing within-item variance at the expense of between-item covariances.

Here too one can proceed with standardized residual adjacencies instead of the raw data. First, replace the submatrices $\boldsymbol{N}_{ij}$ with adjacency submatrices $\boldsymbol{A}_{ij}$ constructed as in Section 4.4, and we replace all the main diagonal counts with zeros. Call the resulting matrix

$\boldsymbol{A}$. Now compute "pseudo-" Guttman scores for the categories of each item by performing a Guttman analysis of the matrix $\boldsymbol{A}$, instead of $\boldsymbol{Y}^T\boldsymbol{Y}$, and replacing $\boldsymbol{D}$ with the identity matrix, since the diagonal elements of $\boldsymbol{A}$ are now all zeros.

The result of Guttman analysis of $\boldsymbol{A}$ will be category scores that approximately maximize association in each pairwise table $\boldsymbol{N}_{ij}$, following the logic of Section 4.4; and since the block matrices on the diagonal of $\boldsymbol{A}$ are all zeros, the maximum must be achieved by maximizing just association in the two-way tables, not by maximizing variance within items. Both Guttman analysis of $\boldsymbol{Y}^T\boldsymbol{Y}$ and Guttman analysis of A are considered in Section 5.1.

4.6 Sorting to achieve MLR

The methods described so far address positive association and comonotonicity between items, as expressed by assumption M of Section 2 or (2) of Section 4.1. It is also possible to consider sorting item categories to achieve the stronger monotone likelihood ratio (MLR) properties of MILR and (3). We have begun to explore a combinatorial, rather than principal components approach to this problem.

The method is best described by example. Consider a two-way table of counts cross-classifying responses from a three-level item X_1 with those of a three-level item X_2:

```
         X2
       0  1  2
   0  11  0  3
X1 1 238 11 57
   2  83  1 21
```

Associations for each two-by-two subtable can be calculated by a variety of methods such as

- odds ratios or log odds ratios;
- determinants;
- signed chi-squared (this is the usual chi-squared test statistic for independence but multiplied by ± 1 so that its sign agrees with the log odds ratio).

As an illustration, the matrix of 2×2 associations for the table above, using determinants to measure association, is

	0,1	0,2	1,2
0,1	121	-87	-33
0,2	11	-18	-3
1,2	-675	267	174

[For example, the determinant of the 2×2 subtable relating response categories 1 and 2 of X_1 with categories 1 and 2 of X_2 is $(11)(21) - (57)(1) = 174$, as shown by the element in row (1,2) and column (1,2) in the matrix of associations.]

It is easy to see that if a pair of rows (or columns) of the original table cross-classifying X_1 with X_2 are transposed, the entries of the matrix of associations will change position and possibly sign, but not magnitude. For example, if columns 0 and 1 of the original two-way table were swapped, the resulting matrix of associations would be

	1,0	1,2	0,2
0,1	-121	-33	-87
0,2	-11	-3	-18
1,2	675	174	267

For items with three response categories it is easy to write an algorithm that will produce an ordering of the item categories that is consisitent with (3), if one exists. The algorithm repeatedly sweeps through the matrix of associations, and swaps rows or columns original two-way table if either of the following conditions holds:

Fully reversed labels If any row or column of the matrix of 2×2 associations has all-negative associations, swap the corresponding pair of rows or columns of the two-way table;

Look-ahead If any swap of adjacent rows or columns of the two-way table leads to fully-reversed labels on the next sweep, perform that swap.

A "greedy" version of this algorithm—that is to say, on each sweep the first swap that meets one of these criteria (with fully reversed labels taking precedence over look-ahead) is performed—will be illustrated in Section 5.2.

Since only the signs of the measures of 2×2 association matter, it is possible that this algorithm will produce the wrong order due to capitalization on chance: small, nonzero associations may have the "wrong" sign by chance, and this will lead the algorithm to the wrong ordering of item categories. An alternative is to use a measure of association that has an "indifference" region near zero, to reduce the chances of such misdirection. A simple, interpretable measure of this sort is to compute the signed chi-squared statistic ξ for each 2×2 subtable, and then replace it with a decision function

$$\delta_c = \begin{cases} 1, & \text{if } \xi > c; \\ 0, & \text{if } -c \leq \xi \leq c; \\ -1, & \text{if } \xi < -c, \end{cases}$$

where c is a fixed cutoff value. For example, $c = 3.84$ corresponds to an approximate level 0.05 test of association in each 2×2 table, and $c = 1.00$ corresponds to an approximate level 0.32 test of association. The larger c is, the more conservative (less re-labelling) the algorithm will be. We have found in preliminary work that $c = 1.00$ works somewhat better than $c = 3.84$ in recovering the correct labels from randomly mislabelled simulation data.

5 Exploring Monotonicity in the Example Data Sets

In this section we illustrate exploratory data analysis for monotonicity as expressed by conditons M and MILR in Section 2.1, using the tools described in Section 4. In Section 5.1 we consider M and condition (2) of Section 4.1. In Section 5.2 we consider MILR and condition (3) of Section 4.1.

Before beginning it may be useful to consider the illustration in Figure 1. The graph on the left in Figure 1 shows the ICRF's for a five-category parametric partial credit item. The graph on the right shows the corresponding ISRF's. The essential feature here is that the ISRF's for an item satisfying M are always nondecreasing, like the ICC's of a monotone IRT model for binary data. It is harder to see what MILR means in terms of a graph of the ISRF's, but roughly speaking we expect items satisfying MILR to have approximately "parallel" ISRF's. [Note also that the ISRF for the lowest category is always identically equal to 1; therefore in further plots this ISRF will always be omitted.]

5.1 Condition M and the np-GRM

5.1.1 M and transitive reasoning

Recall from section Section 3.1 that the item categories for the transitive reasoning items are labelled 0, 1 and 2, and that we expect either M or MILR to hold for all nine transitive

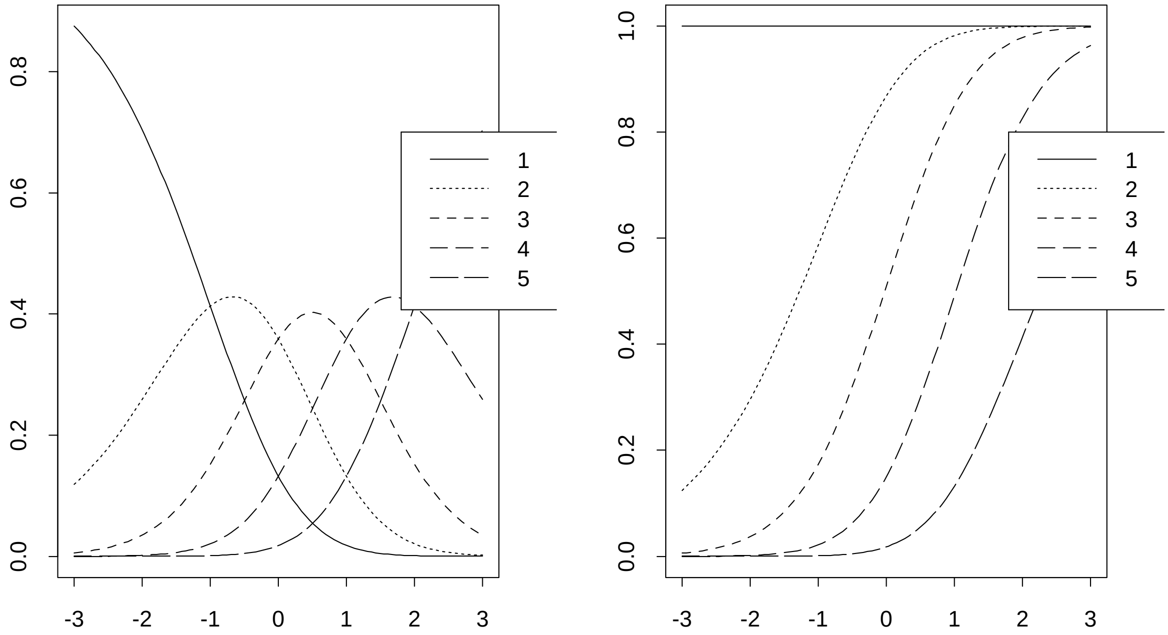

Figure 1: ICRF's (left plot) and ISRF's (right plot) of a Masters' partial credit item with five response categories.

reasoning items. To illustrate both the use of nonparametric regression plots to check M, as well as the effect of mislabelling item categories on M, consider Figure 2.

The top two plots in Figure 2 show nonparametric regression estimates of ISRF's for item 1 of the transitive reasoning items, calculated according to formula (2) of Section 4.1 using the nonparametric regression technique of Section 4.2. The θ estimate for the top left figure is g(S), where S is the rest score (sum of all items except item 1) calculated after dichotomizing the items at category 1, and g is a monotone transformation for which $\theta = g(S)$ will have a standard normal distribution. The θ estimate for the top right figure is the same, except that S is the rest score calculated after dichotomizing the items at category 2. If the $d = 1$, LI, M model holds then, aside from random error, the ISRF's plotted in these top two figures should all be monotone increasing. This appears to be the case.[2]

[2]Approximate pointwise confidence envelopes for the ISRF's were calculated but not plotted since they clutter the figure too much. These envelopes are sufficiently wide near the nonmonotonicities in the top two plots of Figure 2 that these nonmonotonicities may be attributed to random error.

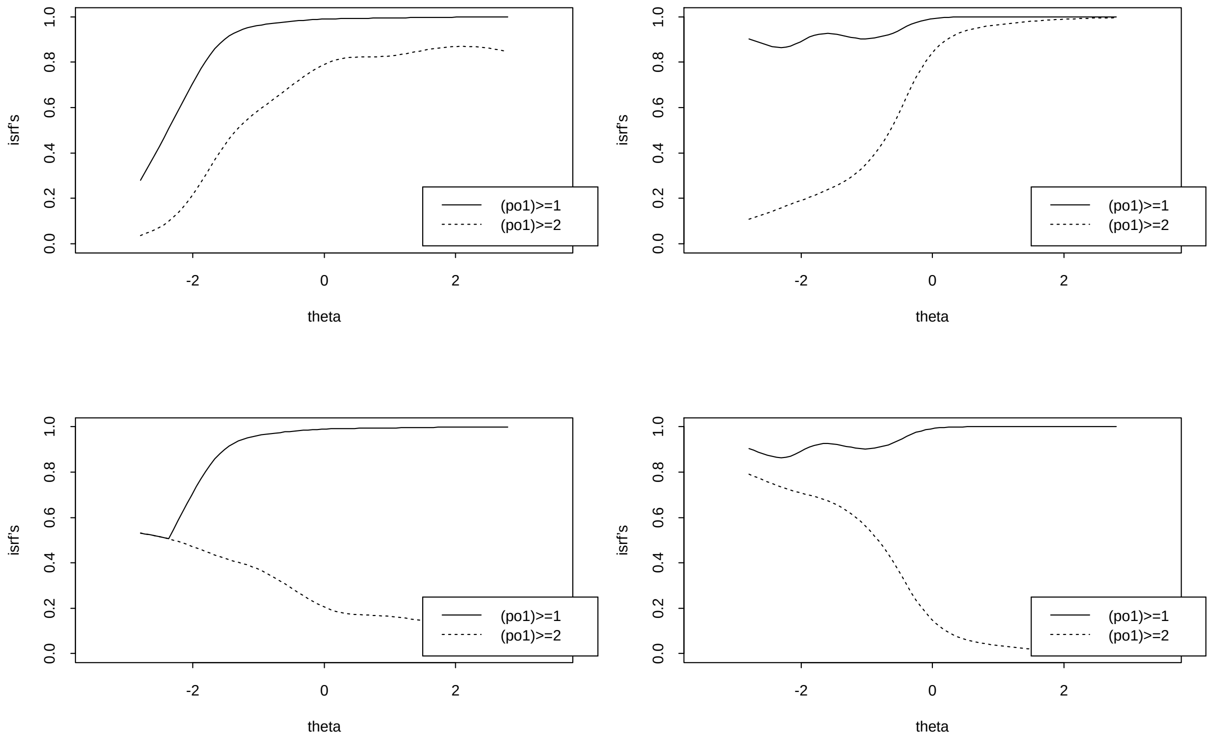

Figure 2: Effect of mislabelling categories on item 1 of the Transitive Reasoning scale. *Top row:* nonparametric-regression ISRF's for correctly labelled item categories, using rest score with items dichotomized at level 1 (left) and rest score with items dichotomized at level 2 (right). *Bottom row:* nonparametric-regression ISRF's for mislabelled categories, using rest score with items dichotomized at level 1 (left) and rest score with items dichotomized at level 2 (right).

The bottom row shows the same regressions after relabelling the item categories from 0, 1, 2 to 0, 2, 1. The effect of the relabelling is clear: the estimated ISRF for the new category 2 is no longer monotone. In some cases (not shown here) relabelling does not actually produce a violation of monotonicity in the estimated ISRF's but it greatly reduces the discrimination (slope) of each ISRF. Thus—as might be expected—even when monotonicity is not violated, mislabelling tends to reduce the apparent information available to us in estimating a single latent variable from the item responses.

To explore and confirm the labelling of item categories described in Section 3.1, both forms of Guttman category scores (using the matrix of counts $\boldsymbol{Y}^T\boldsymbol{Y}$ and the matrix of standardized residual adjacencies $\boldsymbol{A}$) from Section 4.5 were calculated and used to re-sort the items. In addition, the scaling coefficient H from Section 4.3 was calculated.

For the original data, $H = 0.4094$. This value suggests that the items do indeed form a np-GRM scale.[3] When Guttman scaling is applied using the matrix of counts $\boldsymbol{Y}^T\boldsymbol{Y}$ and the item categories are sorted according to the Guttman category scores, no changes in item category order occur. This tends to confirm the notion that the item categories are indeed correctly labelled. Since there is no change in labelling, the scaling coefficient H is again 0.4094.

When Guttman scaling is applied using the residual adjacency matrix $\boldsymbol{A}$ described in Section 4.5, a few relabellings occur: items 1, 2 and 3 have categories 0 and 1 reversed, and item 10 has categories 1 and 2 reversed. To explore the effects of the reversal we consider items 1 and 8 only, in Figure 3. The top row of Figure 3 shows the effect of relabelling on item 1: The relabelling increases the discrimination of the ISRF's of the new categories 1 and 2, near the center of the estimated θ distribution (i.e. near 0 on the θ scale). In item 10, the discrimination for the new category 1 ISRF is greatly increased—though not particularly near the center of the estimated θ distribution—and the increase is bought at the expense of introducing a non-monotonicity in the category 2 ISRF.

The value of H for the transitive reasoning items, after relabelling with the second Guttman analysis, is 0.3754. The reduction in H from 0.4094 to 0.3754 is perhaps not surprising; Guttman scaling with $\boldsymbol{Y}^T\boldsymbol{Y}$ works to increase actual inter-item covariances (as well as within-item variances), and this is what the numerator of the H coefficient detects. On the other hand Guttman scaling with the residual adjacency matrix $\boldsymbol{A}$ tends to increase association but does not directly manipulate inter-item covariances. In the next example, we will see that scaling with $\boldsymbol{A}$ tends to produce more montonicity but does not maximize H.

Whether the relabelling produced by Guttman analysis with $\boldsymbol{A}$—or any automatic technique—is appropriate must be decided on the basis of subject matter expertise. In the case of these transitive reasoning items, the original scoring categories are essentially forced by the partial-credit nature of the items; relabelling probably does not make sense here. The fact that some items are relabelled may suggest some degree of misfit to a single unidimensional model for items 1, 2, 3 and 10. In this case the H_i item misfit coefficients for the nine items remaining after items 7, 9 and 12 were omitted (see Section 3.1):

[3]Sijtsma and Junker, 1996, analyze a dichotomized version of the same data and conclude that items 2 and 4 must be omitted to obtain an item set satisfying Mokken's more restrictive model of double monotonicity.

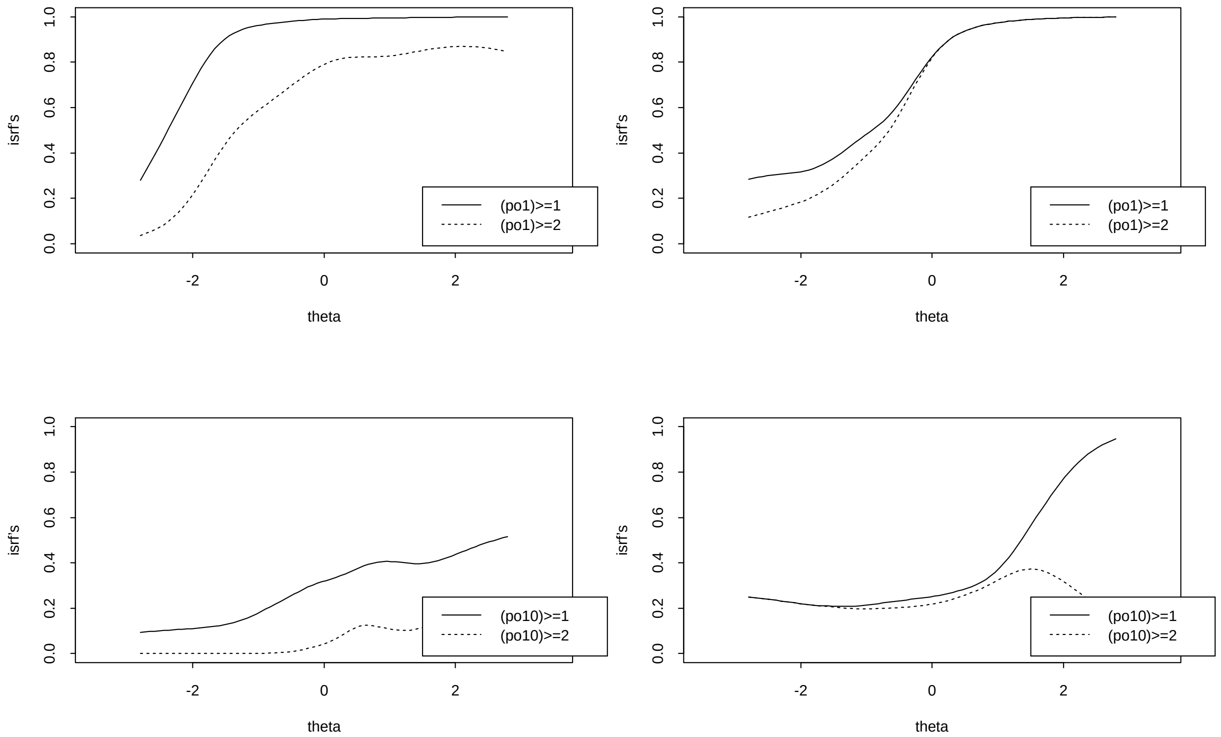

Figure 3: Effect of Guttman scaling of residual adjacencies on items 1 and 10 of the Transitive Reasoning scale. *Left column:* the items before re-labelling. *Right column:* the items after re-labelling. Nonparametric-regression ISRF's were calculated using rest scores with items dichotomized at level 2 in all plots.

Item (i):	1	2	3	4	5	6	8	10	11
H_i:	0.49	0.13	0.52	0.19	0.49	0.53	0.52	0.34	0.35

These do not suggest any particular misfit for the for relabelled items; indeeed only items 2 and 4 have relatively low H_i values. Alternatively we might look at the relabelling algorithm itself: Since no effort has been made to account for capitalization on chance, it is perhaps not surprising that occasional disorders like this may occur.

5.1.2 M and personality self-report

The IIP personality self-report consists of 127 items marked on a five-point, 0 to 4, Likert scale. Although the items are marked on a Likert scale, there is evidence in the wording of some items on self-reports like this that the Likert labelling of responses is not necessarily comonotone with a single latent variable measuring the level of personality disorder, as discussed in Section 3.2.

Because, in the data set we will consider, there are only $n = 97$ subjects, it makes little sense to try to come to grips with the monotonicity properties of all 127 items at once. Instead, in a preliminary scan through the data, groups of 10 items at a time were considered as separate personality subscales. This leads to consideration of 12 subscales of 10 items each and one subscale of 7 items. The subscales are relatively well related to one another; for example the minimum correlation among the total scores on all 13 subscales is 0.19, the maximum correlation is 0.81, the median is 0.49 and half the correlations are between 0.34 and 0.63; see Figure 4.

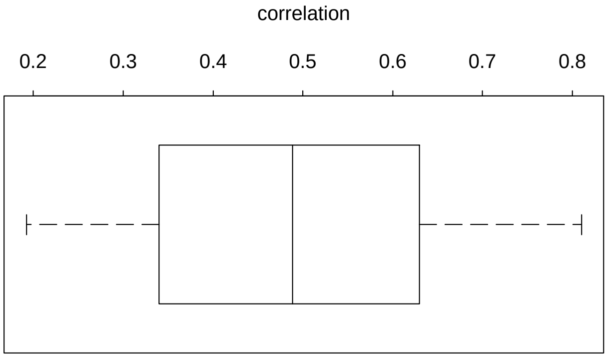

Figure 4: Summary of correlations among 13 subscale total scores from the IIP.

The order of the item categories in each subscale were investigated, again using Guttman scaling with the matrix of raw counts $\boldsymbol{Y}^T\boldsymbol{Y}$ and with the residual adjacency matrix $\boldsymbol{A}$. Most subscales showed some level of violation of M for particular items, when item categories were labelled according to the Likert scheme. We will concentrate on the subscales consisting of items 1 through 10 and items 91 through 100 as useful illustrations of the method.

Figure 5 and Figure 6 shows items 1 through 5 of the IIP with categories labelled according to the Likert scheme, in the leftmost column of each figure. Table 1 shows how the original categories were relabelled after sorting with Guttman scaling using raw counts ($\boldsymbol{Y}^T\boldsymbol{Y}$) and Guttman scaling using residual adjacencies ($\boldsymbol{A}$): For example, category 0 for item 1 was relabelled to be category 4 by both Guttman scaling methods. In Figure 5 and Figure 6, the center column shows the estimated ISRF's using the category orders for each item determined by Guttman scaling using raw counts, and the rightmost column shows the estimated

item:	item1:	item2:	item3:	item4:	item5:
Orig. order	-1 0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Guttman (YTY)	-1 4 2 1 0 3	4 3 2 1 0	4 2 1 0 3	3 1 0 2 4	4 3 2 0 1
Guttman (A)	-1 4 2 0 1 3	4 2 3 1 0	4 3 1 0 2	4 1 0 3 2	4 2 3 0 1
item:	item6:	item7:	item8:	item9:	item10:
Orig. order	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Guttman (YTY)	4 3 2 1 0	4 3 2 0 1	4 3 2 0 1	4 3 2 1 0	4 2 1 0 3
Guttman (A)	4 3 2 0 1	4 3 1 0 2	3 4 1 0 2	4 3 2 1 0	4 2 0 1 3

Table 1: Original item labels and labels after resorting by each of the Guttman scaling methods, items 1 to 10, IIP.

ISRF's using the category orders determined by Guttman scaling using residual adjacencies. In each graph, the nonparametric ISRF's are calculated using the theta estimate $g(S)$ where S is the rest score after dichotomizing all items at level 2 (this dichotomization tends to maximize total score variance for these ten items), and $g(S)$ is a monotone transformation for which $\theta = g(S)$ has a standard normal distribution.

[Note that in Figure 5, item 1 has an extra ISRF for category 0, and in Table 1 there is an extra category -1 for item 1. Category -1 indicates missing responses in this data; therefore in Figure 5 the places where the category 0 ISRF dips indicate θ values for which the corresponding subjects tended to omit responding to this item. In data such as this, not responding can also be informative, so missing response was treated as just another category to be scaled and ordered with the other categories for this item.]

It is interesting to note that for these ten items the category order chosen by Guttman scaling is essentially the reverse of the order indicated by the Likert category labels. Figures 5 and 6 show that with this ordering, more items exhibit good montonicity features than with the original Likert ordering.

Item 4 (Figure 5) is especially interesting, since the plot on the left shows that nonzero responses occur most for subjects with estimated θ in the 0 to 2 range, approximately; subjects with θ less than 0 or greater than 1 tend to respond 0 to this item—this can be seen from the humped shape of most of the ISRF's. Both methods of reordering tend to improve monotonicity for this item, though the residual adjacency method, which produces

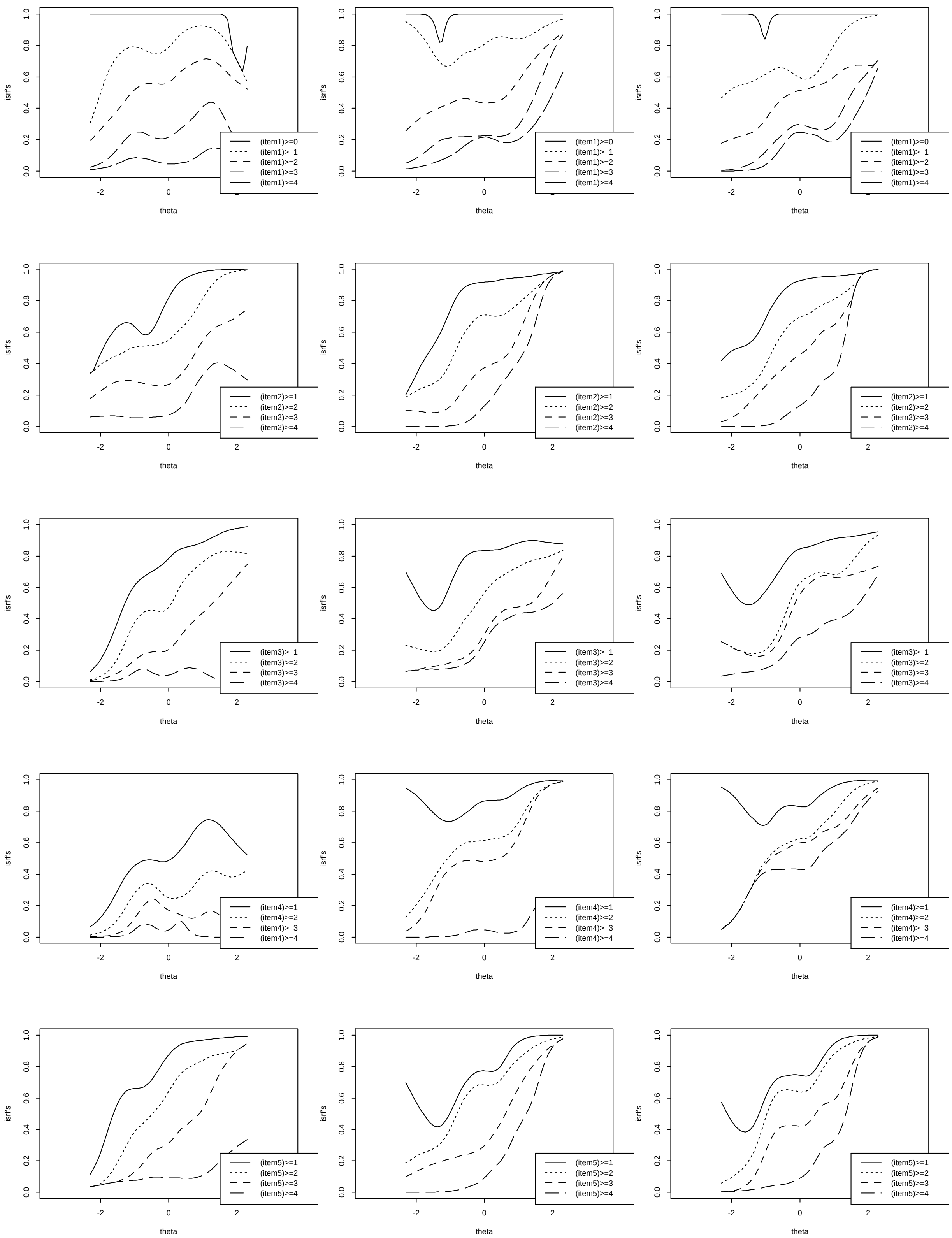

Figure 5: Items 1 through 5 of the IIP, considered as part of the subscale of items 1 through 10. *Left column:* Items before re-labelling. *Center column:* Items after relabelling using Guttman scaling with raw counts. *Right column:* Items after relabelling using Guttman scaling with residual adjacencies. Nonparametric-regression ISRF's were calculated using rest scores with items dichotomized at level 2 in all plots.

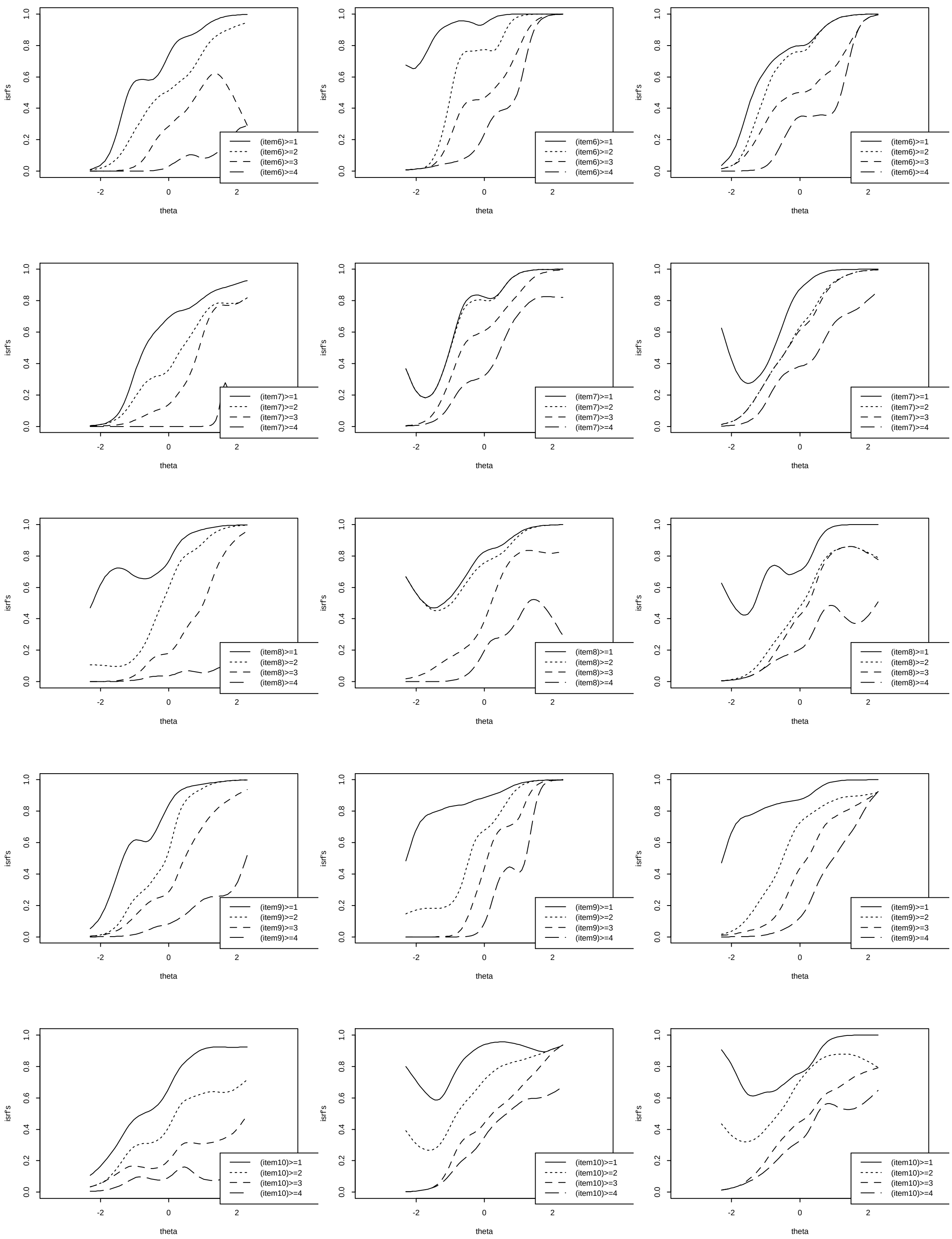

Figure 6: Items 6 through 10 of the IIP, considered as part of the subscale of items 1 through 10. *Left column:* Items before re-labelling. *Center column:* Items after relabelling using Guttman scaling with raw counts. *Right column:* Items after relabelling using Guttman scaling with residual adjacencies. Nonparametric-regression ISRF's were calculated using rest scores with items dichotomized at level 2 in all plots.

item:	item91:	item92:	item93:	item94:	item95:
Orig. order	0 1 2 3 4	0 1 2 3 4	-1 0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Guttman (YTY)	0 1 2 3 4	0 1 2 4 3	0 -1 2 3 1 4	0 2 1 3 4	0 2 1 3 4
Guttman (A)	4 3 2 0 1	4 3 1 0 2	3 4 1 0 2 -1	4 0 3 2 1	4 0 2 3 1
item:	**item96:**	**item97:**	**item98:**	**item99:**	**item100:**
Orig. order	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4	0 1 2 3 4
Guttman (YTY)	0 1 3 4 2	0 1 2 3 4	0 1 4 2 3	0 3 2 4 1	0 1 3 4 2
Guttman (A)	4 1 0 2 3	4 3 2 0 1	4 3 0 1 2	4 2 1 0 3	4 3 1 0 2

Table 2: Original item labels and labels after resorting by each of the Guttman scaling methods, items 91 to 100, IIP.

more nearly-reversed categories, seems to do a bit better.

As with the transitive reasoning items, we pay for the improvement of monotonicity in some items by a degradation of monotonicity of other items. However, for those items where monotonicity is degraded—items 3, 5, 7 and 10, the nonmonotonicity tends to happen near or below -2 on the estimated θ scale, where there are relatively few examinees.

Finally we compare H for each of the three orders of item categories. For the original Likert order, $H = 0.2783$. For Guttman scaling with $\boldsymbol{Y}^T\boldsymbol{Y}$, $H = 0.3453$. For Guttman scaling with $\boldsymbol{A}$, $H = 0.3240$. This is consistent with our earlier experience that $\boldsymbol{Y}^T\boldsymbol{Y}$-based scaling tends to produce a somewhat larger H value than $\boldsymbol{A}$-based scaling. The fact that the two H values are close together is consistent with the visual impression from Figures 5 and 6 that there is not much difference between the monotonicity properties of these two re-orderings.

Turning to the subscale consisting of items 91 through 100, Table 2 shows how the original categories were relabelled after sorting with Guttman scaling using raw counts ($\boldsymbol{Y}^T\boldsymbol{Y}$) and Guttman scaling using residual adjacencies ($\boldsymbol{A}$): The corresponding ISRF's are plotted in Figure 7 and Figure 8. Again, the leftmost column shows the ISRF's using the original Likert labelling, the center column shows the estimated ISRF's using the category orders for each item determined by Guttman scaling using raw counts, and the rightmost column shows the estimated ISRF's using the category orders determined by Guttman scaling using residual adjacencies. In each graph, the nonparametric ISRF's are calculated using the theta estimate $g(S)$ where S is the rest score after dichotomizing all items at level 1 (this dichotomization tends to maximize total score variance for these ten items), and $g(S)$ is a monotone transformation for which $\theta = g(S)$ has a standard normal distribution.

[Note that in Figure 7 and in Table 2, item 93 again has a missing-response category that is scaled and ordered with the other categories.]

In these figures it appears visually that Guttman scaling using residual adjacencies $\boldsymbol{A}$ may be doing somewhat better at identifying an ordering of the item categories that is consistent with monotone ISRF's than Guttman scaling with raw counts $\boldsymbol{Y}^T\boldsymbol{Y}$. For example, item 94 (Figure 7) shows extreme non-monotonicities in all ISRF's like that of item 4 (Figure 5) above. Guttman scaling with raw counts does not improve the situation; but Guttman scaling with residual adjacencies makes this items approximately monotone in the latent trait being estimated. The probability of responding 1 or more to item 95 increases monotonically with θ estimate, but the categories 2, 3 and 4 are used infrequently by subjects, except in the θ range from 0 to 2, approximately. Guttman scaling with raw counts produces new categories 1 and 2 that exhibit reasonably good monotonicity, but categories 3 and 4 are still relatively uninformative. Guttman scaling using residual adjacencies produces a more sensible result (see Table 2): for example, since category 4 is relatively little-used by those who score high on these items, it is re-labelled to be category 1 in by this second Guttman method. The relabelled ISRF's exhibit a fair amount of monotonicity that was not obvious before (item 95, rightmost column, Figure 7). Many similar examples can be found in Figure 8.

Once again, we pay for some better monotonicity in some items with a degradation of monotonicity of other items. Item 91 is especially bad for Guttman scaling with residual adjacencies in this respect; however in general there are fewer items with degraded monotonicity properties after re-ordering among items 91–100 than there were among items 1–10.

Finally, we note that H is rather insensitive to the re-orderings done for items 91–100. For the originally-labelled items $H = 0.2926$; for the items re-labelled by Guttman scaling with raw counts, $H = 0.3120$; and for the items re-labelled by Guttman scaling using residual adjacencies, $H = 0.3058$.

The Guttman scaling methods have been fairly successful at improving the monotonicity properties of these personality self-report items. If we are happy with the re-ordered item categories, we would probably want to allow subjects to continue to mark the items using the present Likert labels, and then re-score the items afterwards using the re-labellings discovered here. Once again, however, the appropriateness of the new item category order (or integer weights) depends on subject-matter expertise. It is also worth noting that, for

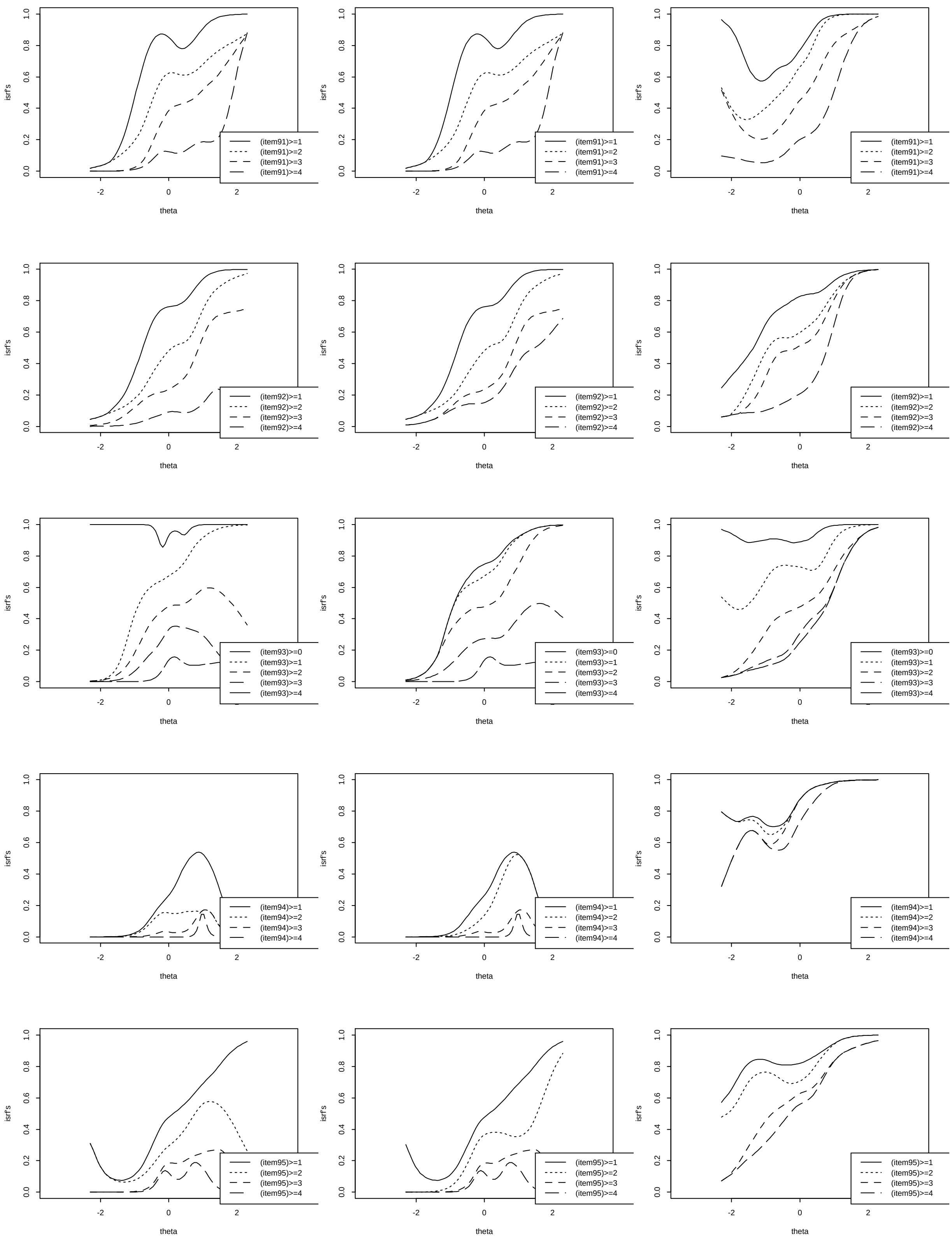

Figure 7: Items 91 through 95 of the IIP, considered as part of the subscale of items 91 through 100. *Left column:* Items before re-labelling. *Center column:* Items after relabelling using Guttman scaling with raw counts. *Right column:* Items after relabelling using Guttman scaling with residual adjacencies. Nonparametric-regression ISRF's were calculated using rest scores with items dichotomized at level 1 in all plots.

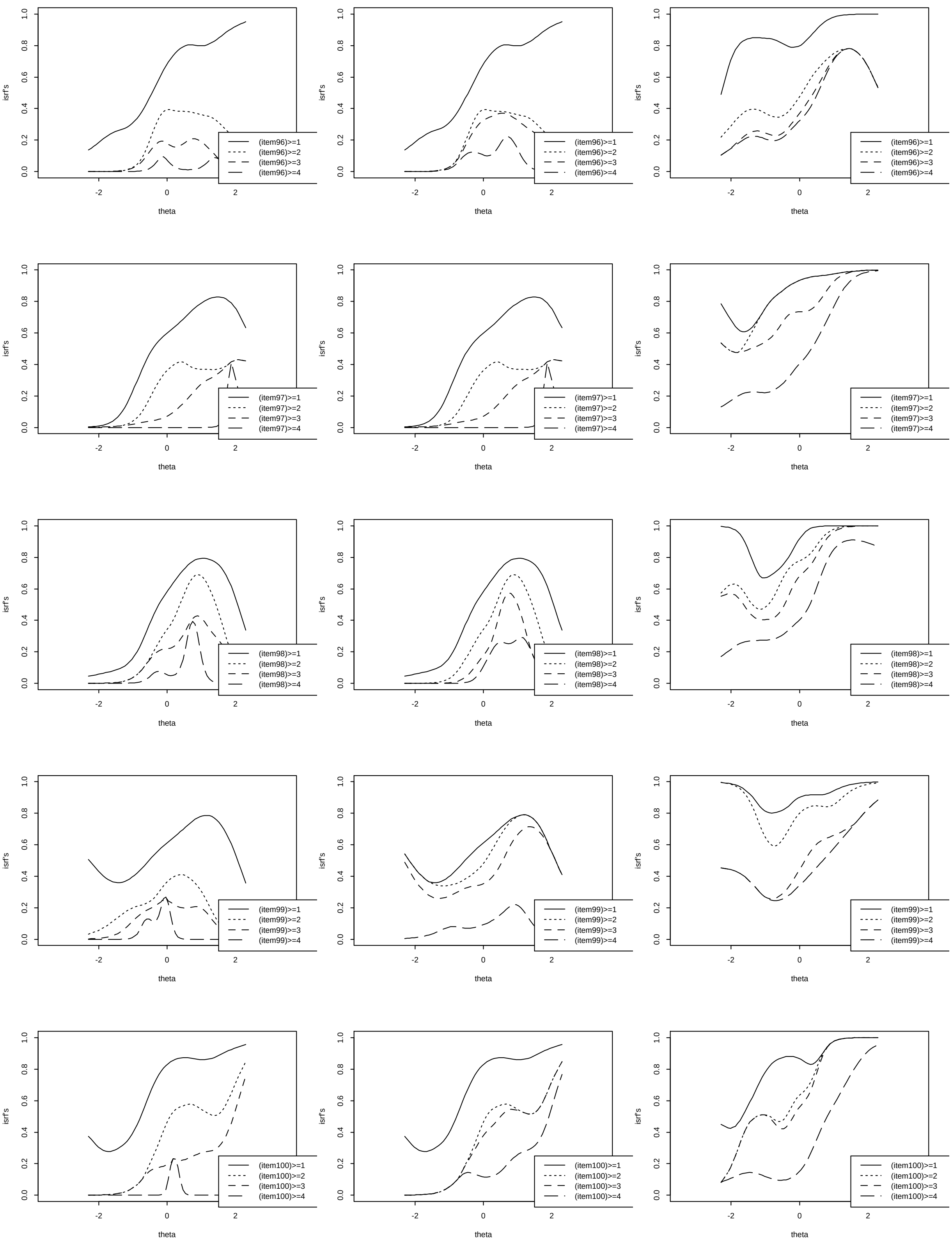

Figure 8: Items 96 through 100 of the IIP, considered as part of the subscale of items 91 through 100. *Left column:* Items before re-labelling. *Center column:* Items after relabelling using Guttman scaling with raw counts. *Right column:* Items after relabelling using Guttman scaling with residual adjacencies. Nonparametric-regression ISRF's were calculated using rest scores with items dichotomized at level 1 in all plots.

ease of interpretation, it may be a more useful exercise to re-write the items so that the Likert category labels do lead to monotone ISRF's.

5.1.3 M and working memory

The 27 items from the working memory load experiment were analyzed in the same way as the IIP items in the previous section. Recall that response category 1 indicates total failure at the cognitive task, response category 2 indicates success with a maximum number of reminders, and response category 5 indicates total success at the task with no reminders. Categories 3 and 4 indicate success with various reminders; they are not expected to be well-ordered with respect to one another. In this case the items were analyzed as a single scale, ignoring some details of the experiment that modify the conditions under which the items were administered. Table 3 gives the original category labels and the re-labelling accoring to the two Guttman methods. Though the pattern is not as clear in Table 3 as one might hope, one does see that categories 1 and 2 tend to remain "low" after re-ordering (by either method) and category 5 tends to remain "high". Moreover neither method of ordering categories tends to reverse the order of categories as much as seemed to be the case in the IIP data.

Neither of the Guttman scaling methods produces item category orders that were particularly compatible with monotone ISRF's. Indeed, H for the original data was only 0.1904; for Guttman scaling with raw counts $H = 0.2319$ and for Guttman scaling with residual adjacencies $H = 0.1937$. This tends to confirm the notion that monotonicity is a restrictive assumption—not all polytomous item response data sets can be relabelled so as to attain monotonicity conditions.

For the purposes of illustration, Figure 9 presents items 1 through 5 of the working memory item set, following exactly the pattern established for Figures 5 through In the figure, Guttman scaling using raw counts (center column) seems to fare poorly on these items. Guttman scaling using residual adjacencies (rightmost column) does slightly better on items 1 and 2 and about the same on items 3, 4 and 5 (ISRF's are calculated using rest scores after dichotomizing at category 5, which is the perfect response category before re-sorting).

item:	item1:	item2:	item3:	item4:	item5:
Orig. order	1 2 3 4 5	1 2 3 4 5	1 4 5	1 2 3 4 5	1 2 3 4 5
Guttman (YTY)	3 2 1 5 4	2 1 3 5 4	1 4 5	1 2 3 5 4	1 3 2 4 5
Guttman (A)	3 2 1 4 5	1 2 3 5 4	4 1 5	1 2 3 4 5	1 2 3 4 5
item:	item6:	item7:	item8:	item9:	item10:
Orig. order	4 5	3 4 5	1 2 4 5	1 2 3 4 5	1 2 3 4 5
Guttman (YTY)	4 5	3 4 5	1 4 5 2	1 2 3 5 4	1 2 4 3 5
Guttman (A)	4 5	4 3 5	1 4 2 5	1 3 4 2 5	1 2 4 3 5
item:	item11:	item12:	item13:	item14:	item15:
Orig. order	1 3 5	1 3 4 5	1 2 3 4 5	1 3 4 5	1 2 3 4 5
Guttman (YTY)	3 1 5	1 3 5 4	4 2 1 3 5	5 3 1 4	2 5 1 3 4
Guttman (A)	3 1 5	1 3 4 5	4 3 2 1 5	4 3 1 5	1 4 2 3 5
item:	item16:	item17:	item18:	item19:	item20:
Orig. order	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
Guttman (YTY)	2 1 3 5 4	1 5 2 3 4	1 4 2 3 5	1 2 5 4 3	1 4 2 3 5
Guttman (A)	2 1 3 4 5	1 4 2 3 5	1 3 2 4 5	1 4 5 2 3	1 4 3 2 5
item:	item21:	item22:	item23:	item24:	item25:
Orig. order	1 2 3 4 5	1 2 3 4 5	1 3 4 5	1 3 4 5	1 2 3 4 5
Guttman (YTY)	1 2 3 4 5	2 1 3 5 4	1 3 4 5	1 3 4 5	2 1 5 4 3
Guttman (A)	2 1 3 4 5	2 1 3 4 5	1 4 3 5	1 3 4 5	2 1 5 3 4
item:	item26:	item27:			
Orig. order	1 2 4 5	1 2 3 4 5			
Guttman (YTY)	1 2 4 5	1 5 3 2 4			
Guttman (A)	1 4 2 5	2 4 3 1 5			

Table 3: Original item labels and labels after resorting by each of the Guttman scaling methods, items 1 to 27, working memory items.

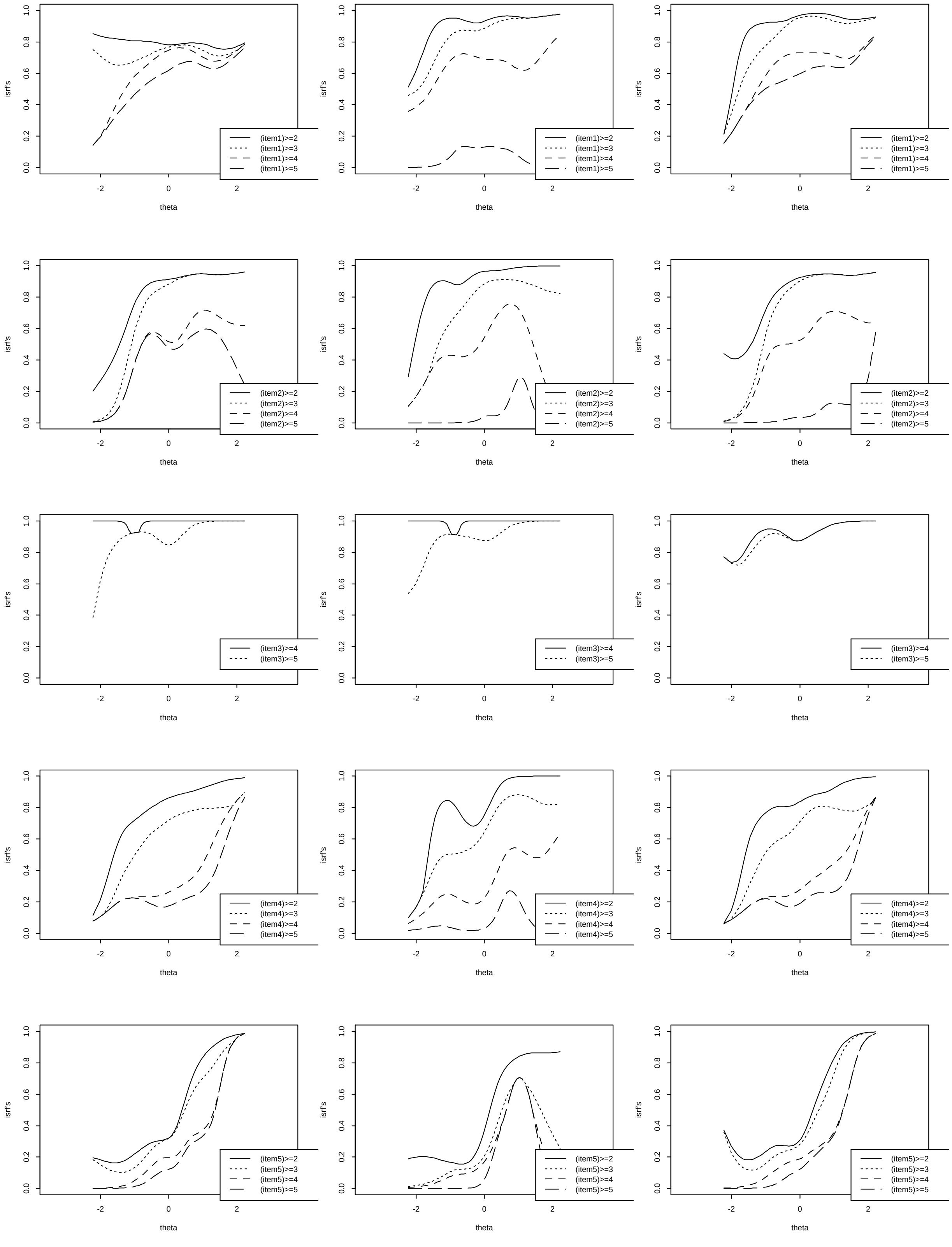

Figure 9: Items 1 through 5 from the working memory experiment. *Left column:* Items before re-labelling. *Center column:* Items after relabelling using Guttman scaling with raw counts. *Right column:* Items after relabelling using Guttman scaling with residual adjacencies. Nonparametric-regression ISRF's were calculated using rest scores with items dichotomized at level 5 in all plots.

5.2 Condition MILR and the np-PCM

Pairs of three-category items can be sorted to try to achieve MILR, or rather the observable condition (3) of Section 4.1, that all 2×2 subtables of each cross-classification of pairs of items have positive association. The sorting algorithm described in Section 4.6 was implemented in S-PLUS; recall that the algorithm repeatedly sweeps through the matrix of associations, and swaps rows or columns original two-way table if either of the following conditions holds:

Fully reversed labels If any row or column of the matrix of 2×2 associations has all-negative associations, swap the corresponding pair of rows or columns of the two-way table;

Look-ahead If any swap of adjacent rows or columns of the two-way table leads to fully-reversed labels on the next sweep, perform that swap.

making the first swap to satify the conditions that it finds, on each sweep.

We begin with an illustration of how the algorithm works. The first two transitive reasoning items have two-way table

```
          X2
       0   1  2
   0   3  11  0
X1 1  21  83  1
   2  57 238 11
```

For the purposes of illustration, we now consider the same two items, but with the labels permuted, as shown in the next two-way table:

```
            X2
        1   2  0
   0   11   0  3
X1 2  238  11 57
   1   83   1 21
```

When this algorithm is applied to our example case of a two-way table for X_1 and X_2 with permuted labels we get the following output:

```
Two-way table:
    1  2  0
0  11  0  3
2 238 11 57
1  83  1 21

2x2 associations:
     1,2  1,0  2,0
0,2  121  -87  -33
0,1   11  -18   -3
2,1 -675  267  174

Swapping labels 1 and 2 of X1...

Two-way table:
    1  2  0
0  11  0  3
1  83  1 21
2 238 11 57

2x2 associations:
    1,2  1,0  2,0
0,1  11  -18   -3
0,2 121  -87  -33
1,2 675 -267 -174

Swapping labels 0 and 1 of X2...

Two-way table:
   0  2   1
0  3  0  11
1 21  1  83
2 57 11 238

2x2 associations:
    0,2 0,1  2,1
0,2   3  18  -11
0,1  33  87 -121
2,1 174 267 -675

Swapping labels 0 and 2 of X2...
```

```
Two-way table:
   0   1  2
0  3  11  0
1 21  83  1
2 57 238 11

2x2 associations:
    0,1 0,2 1,2
0,1  18   3  11
0,2  87  33 121
1,2 267 174 675

All sorted!
```

This algorithm will halt if a sorting of the labels of X_1 and X_2 is possible, but it has no global stopping procedure if no sorting can be found. Currently the algorithm halts if no more swaps can be found, or if the number of swaps made so far exceeds $2 \times \left[\binom{r}{2} + \binom{c}{2}\right]$. In either case, the algorithm backtracks to the configuration with the greatest correlation between the items.

In the illustration above, determinants were used to measure 2×2 associations. We have also used the algorithm with the decision function δ_c, based on the signed chi-squared statistic, as described in Section 4.6 above.

5.2.1 MILR and transitive reasoning

The algorithm described above was applied to the the nine transitive reasoning items considered earlier. As with the first Guttman analysis in Section 5.1.1, we expect to see no change in the label order. Since the items are considered in pairs only, there is the distinct possibility that the best category order for item 1, say, when paired with item 2, will be different than the best order for item 1, when paired with item 3. This can be accounted for, to a limited extent, by incorporating an indifference region around 0 association, with the decision function δ_c described in Section 4.6 above.

In the interest of brevity we summarize the results here:

- Using determinants to measure 2×2 association, there are 22 violations of (3) in the $\binom{9}{2} = 36$ two-way tables. Nevertheless, all the item category orders are preserved, except for item 2 in the "item 2 × item 5" table and in the "item 2 × item 7" table. In both cases, categories 0 and 1 are swapped for item 2.

- Using δ_c with $c = 1.00$, there are only 8 violations. In the "item 2 $\times$ item 8" table, categories 0 and 1 of item 2 are swapped, and in the "item 4 by item 9" table, categories 0 and 1 of item 9 are swapped. [The items who categories are re-sorted here need not be a subset of the items chosen by determinants alone, since 0's are essentially ignored when the algorithm looks for an all-negative row or column].
- Using δ_c with $c = 1.00$, there are only 2 violations. Only in the "item 4 by item 9" table, categories 0 and 1 of item 9 are swapped.

These results are generally consistent with the results of Guttman scaling in Section 5.1.1, and moreover they are supportive of the idea that a (nonparametric) partial credit model might be applied to the transitive reasoning data.

5.2.2 MILR and simulated PCM data

To further illustrate the effectiveness of the sorting algorithm in achieving (3), we applied the algorithm twice to the simulation data described in Section 3.4. The results are summarized in Table 4. In the table, "*Yes*" means the original category orders were recovered from data with the category labels permuted; "*Reversed*" means the original category orders were recovered, but both are in reverse order; and "*No*" means that the original category orders were not recovered from the simulated data. In the "remarks" column, the notation "*Table insoluble*" means that the matrix of 2×2 determinants is such that there does not exist a sorting of the categories of the two items that will satisfy condition (3). Finally "*(No permutation)*" means that the random permutation of item categories left the categories of this item unchanged.

Since the data for Table 4 was simulated from a model that exactly satisfies MILR and condition (3), it follows that the existence of insoluble tables and/or inability to recover the original orders of categories is due to chance alone. In view of the rather large number of item pairs in which the original order (or a reversal of it) was not discovered in Table 4, the confirmation of the original order of categories for the transitive reasoning items in Section 5.2.1 is all the more convincing.

A more complete understanding of the category sorting procedure could be obtained with further simulations. In particular the tuning/indifference parameter c for the measure of association δ_c could be set, using simulation data, to balance type I and type II error in selecting item category orders.

Item pair	Recovered using Determinants?	Recovered using δ_C, $c=1$?	Remarks
1,2	Reversed	Yes	
1,3	Yes	Yes	
1,4	Yes	Yes	
1,5	Yes	Yes	
1,6	Yes	Yes	
1,7	Yes	Yes	
1,8	Yes	Yes	Table insoluble
1,9	No	Yes	
1,10	Yes	Yes	
2,3	No	Reversed	Table insoluble
2,4	Reversed	Reversed	
2,5	No	No	Table insoluble
2,6	Reversed	Reversed	
2,7	Reversed	Reversed	
2,8	Reversed	Reversed	
2,9	Reversed	Reversed	
2,10	Reversed	Reversed	
3,4	No	Yes	Table insoluble
3,5	Yes	No	
3,6	Yes	Yes	
3,7	—	—	(No permutation)
3,8	—	—	(No permutation)
3,9	No	No	Table insoluble
3,10	Yes	Yes	
4,5	Reversed	Reversed	
4,6	Reversed	Reversed	
4,7	No	No	Table insoluble
4,8	Yes	Yes	
4,9	No	Yes	Table insoluble
4,10	Reversed	Reversed	
5,6	No	No	Table insoluble
5,7	No	Yes	Table insoluble
5,8	No	No	Table insoluble
5,9	Reversed	Reversed	Table insoluble
5,10	Reversed	Reversed	
6,7	Yes	Yes	Table insoluble
6,8	No	Yes	
6,9	Yes	No	
6,10	Yes	Yes	
7,8	—	—	(No permutation)
7,9	Yes	Yes	
7,10	No	Yes	Table insoluble
8,9	No	Yes	Table insoluble
8,10	Yes	Yes	
9,10	No	Reversed	Table insoluble

Table 4: Results of applying two variations of the MLR sorting algorithm to the transitive reasoning data. *Yes* means the original category orders were recovered from data with the category labels permuted. *Reversed* means the original category orders were recovered, but both are in reverse order. *No* means that the original category orders were not recovered from the simulated data.

5.2.3 MILR for more than three categories

There is a need for sorting higher order tables efficiently; brute force exhaustive search is not likely to succeed. Even for a 5×5 table, the total number of orderings to consider is $5!^2 = 14400$; at 0.5 seconds of computation to examine each table, that's 2 hours of computation per table. For a scale like the IIP, there are $\binom{127}{2} = 8001$ tables, leading to over 16,000 hours of computation in the worst case to find an optimal category ordering for all pairs of items.

The sorting algorithm described in Section 4.6 does not work well, as written, for more than two categories (this is hardly surprising, given its simplicity). Two approaches to generalizing to higher-order tables will be considered in the future:

- Directly generalizing the algorithm to more than three categories;
- Adapting principal components analysis to the problem of maximizing all 2×2 subtable associations in a cross-classification of pairs of items.

If the former approach above is successful, it will have to be combined with some sort of "chaining" procedure to move from one two-way table to the next, to obtain maximal pools of items within the original item set that can be ordered co-monotonically. To reduce the likely size of the region of the search space that must be explored, it will probably be a good idea to "seed" the algorithm with the output of one of the Guttman-scaling-based methods described above.

If the latter approach above is successful it can probably be adapted to a Guttman-like procedure, maximizing over all item pairs at once, as with Guttman analysis using the residual adjacency matrix $\boldsymbol{A}$.

6 Discussion

In this paper we have tried to describe and illustrate exploratory data analysis (EDA) techniques for examining the order of item responses so that the data is compatible with a monotone unidimensional latent variable model for polytomous data. Two such models were considered: the np-GRM, which in addition to local independence and unidimensionality of the latent trait makes the monotonicity assumption that $P[X_j > r|\theta]$ is nondecreasing in θ;

and the np-PCM, which makes the stronger monotonicity assumption that $P_{j(s+1)}(\theta)/P_{js}(\theta)$ is nondecreasing in θ.

Two conditions on the distribution of the observable item response data were derived, following the theoretical development of Junker (1993): the monotonicity assumption of the np-GRM corresponds to the observable, manifest condition that $P[X_j \geq r|S = s]$ should be nondecreasing in s if S is a "rest score" calculated after dichotomizing all the items. And the np-PCM model implies that all 2×2 subtables the table cross-classifying any two items X_i and X_j should be positively associated.

Some variations on a classical scaling technique, Guttman scaling—motivated by optimal scoring in archaelogical seriation and ordination problems—were used to obtain numerical category scores for each item; and these category scores were used to re-sort the item categories. Generally the re-sorted items had better monotonicity properties than the original items. A combinatorial algorithm to attain positive 2×2 subtable associations was also described and illustrated for items with three response categories.

The results of the paper seem to be promising for the development of a general EDA methodology for exploring monotonicity and the order of response categories for polytomous items. With the additional freedom of multiple categories in polytomous IRT, it is important to have a fully-developed EDA methodology, to confirm existing category orders as well as to search for new category orders. Whether the new category orders should be adopted remains, as always, a subject-matter question that is best answered in conjunction with the item writers and the users of the scale, test or item set being considered.

References

Abrahamowicz, M. and Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika, 57,* 5–27.

Basilevsky, A. (1994). *Statistical factor analysis and related methods: theory and applications.* New York: Wiley.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Ellis, J. L. (1994). Foundations of latent variable models. Doctoral Dissertation, Nijmegen Institute for Cognition and Information, Catholic University of Nijmegen, the Netherlands.

Ellis, J. L., and Wollenberg, A. L. van den (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika, 58,* 417–429.

Hemker, B. T., Sijtsma K., Molenaar, I. W. and Junker, B. W. (1995a). Polytomous IRT models and monotone likelihood ratio of the total score. Accepted, *Psychometrika.*

Hemker, B. T., Sijtsma K., Molenaar, I. W. and Junker, B. W. (1995b). Polytomous IRT models and stochastic ordering. Accepted, *Psychometrika.*

Holland, P. W., and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait models. *Annals of Statistics, 14,* 1523–1543.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56,* 255–278.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21,* 1359–1378.

Junker, B. W. (1996). Some remarks on Scheiblechner's treatment of ISOP models. Technical Report* #643, Department of Statistics, Carnegie Mellon University, Pittsburgh PA 15213. Submitted for publication.

Junker, B. W., and Ellis, J. L. (1995). A characterization of monotone unidimensional latent variable models. Technical Report* #614, Department of Statistics, Carnegie Mellon University, Pittsburgh PA 15213. Submitted for publication.

* Technical Reports from the Department of Statistics at Carnegie Mellon University are available on the World Wide Web at `http://www.stat.cmu.edu/cmu-stats/`.

Li, H.-H. and Stout, W. F. (1995). A version of DIMTEST to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data. Paper presented at the Annual NCME meeting, San Francisco CA, April 20, 1995.

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton.

Mokken, R. J., and Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 37,* 97–117.

Patz, R. J., Junker, B. W., Lerch, F. J. and Huguenard, B. R. (1996). Analyzing small psychological experiments with item response models. Technical Report* #644, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* Psychometrika Monograph Supplement to Volume 24, Monograph Number 17.

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika, 60,* 281–304.

Statistical Sciences, Inc. (1992). *S-PLUS, Version 3.1 Release 1 for HP Series 700, HP-UX 8.x.* Seattle, WA: author.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Sijtsma, K. and Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. Technical Report* #597, Department of Statistics, Carnegie Mellon University, Pittsburgh PA 15213. To appear, *British Journal of Mathematical and Statistical Psychology.*

Sijtsma, K., and Verweij, A. C. (1992). Mokken scale analysis: theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education, 5,* 355–378.

* Technical Reports from the Department of Statistics at Carnegie Mellon University are available on the World Wide Web at `http://www.stat.cmu.edu/cmu-stats/`.