Measurement and Research Department Reports 2002-2

Are Attitude Items Monotone Or Single-Peaked? An Analysis Using Baybesian Methods

Gunter Maris

Cito groep

Measurement and Research Department Reports 2002-2

ARE ATTITUDE ITEMS MONOTONE OR SINGLE-PEAKED? AN ANALYSIS USING BAYESIAN METHODS.

Gunter Maris

CITO NATIONAL INSTITUTE FOR EDUCATIONAL MEASUREMENT

ARNHEM

Eric Maris

DEPARTMENT OF MATHEMATICAL PSYCHOLOGY

NIJMEGEN INSTITUTE FOR COGNITION AND INFORMATION (NICI)

UNIVERSITY OF NIJMEGEN

Citogroep
Arnhem, September 16, 2002

Abstract

A methodology is presented for evaluating whether the item response function of an attitude statement is monotone or single-peaked. A new model for attitude measurement is introduced. This model has item response functions that are single-peaked. Depending on the item parameters, the peak of the item response function can be at plus or minus infinity which corresponds to a monotone increasing or decreasing item response function. That is, the new model contains a model with monotone item response functions as a special case. This implies that testing whether the item response function of a statement is monotone, boils down to evaluating the goodness-of-fit of a restricted model relative to a more general (unrestricted) model. The goodness-of-fit of the restricted model is evaluated with a posterior-predictive check. It is found that the power of the traditional posterior-predictive check is unsatisfactory. This problem is solved by computing the posterior-predictive p-value using the posterior distribution of the nuisance parameters under the unrestricted instead of the restricted model. The methodology is applied to a real data set.

Key words: MCMC methods, Gibbs sampling

This paper is about probabilistic models for attitude measurement. Attitudes are usually measured by the responses to a number of statements expressing various attitudes with respect to a particular topic. Examples of such topics are capital punishment, environmental issues, and health care. In this paper, the responses are assumed to be dichotomous, indicating agreement or disagreement with the statement. We use Y_{vi} to denote the response of the v-th subject on the i-th statement. The variable Y_{vi} is defined as follows:

$$Y_{vi} = \begin{cases} 1 & \text{if subject } v \text{ agrees with statement } i \\ 0 & \text{if subject } v \text{ does not agree with statement } i \end{cases}$$

In the following, statements will also be called *items*.

In probabilistic models for attitude measurement the probability that subject v agrees with statement i is governed by the unknown attitude of subject v, denoted by θ_v. Models differ in the precise nature of the relation between this probability and θ_v. Usually, this probability is assumed to be a parametric function of θ_v. This parametric function also depends on one or more item parameters, which are denoted by $\boldsymbol{\omega_i}$. The probability that person v agrees with item i is denoted by $P(Y_{vi} = 1|\theta_v, \boldsymbol{\omega}_i)$, and is called the item response function (IRF).

Most probabilistic models for attitude measurement are inspired by Coombs' (1964) deterministic parallelogram model. Some examples are the models developed by Andrich and Luo (1993), Hoijtink (1990), Verhelst and Verstralen (1993), and Roberts (1995). The model of Coombs assumes that the response (agree or disagree) depends on the distance between the attitude of the subject and the content of the item, denoted by δ_i ($\delta_i = \omega_{i1}$). Specifically, if the attitude of the subject is sufficiently close to the content of the item, the subject agrees with the item, otherwise he

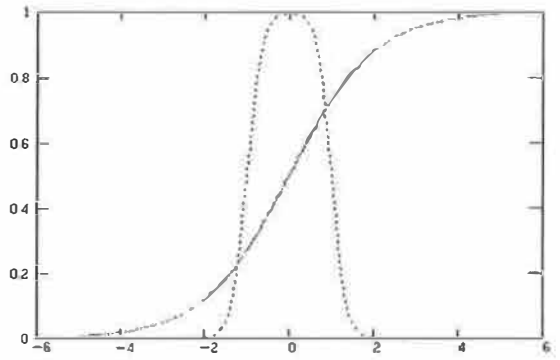

FIGURE 1.
Example of a monotone (solid line) and a single-peaked (dashed line) IRF.

disagrees. Formally, this can be written as follows:

$$Y_{vi} = \begin{cases} 1 & \text{if } |\theta_v - \delta_i| \leq \xi \\ 0 & \text{if } |\theta_v - \delta_i| > \xi \end{cases},$$

in which ξ is some positive number. That is, the response is a decreasing function of the distance between the subject's attitude and the item's content. It follows from the model that, if the subjects are ordered according to their attitude, and the items are ordered according to their content, then the 1-responses in the data matrix have the shape of a parallelogram. Probabilistic models for attitude measurement that are inspired by the parallelogram model have the property that the *probability* that a person agrees with an item is a decreasing function of the distance between the attitude of the subject and the content of the item. This implies that the IRF as a function of θ is first increasing (when θ is smaller than δ_i) and then decreasing (when θ is larger than δ_i). In other words, the IRF is a *single-peaked* function of θ. An example of such an IRF is given in Figure 1.

A second class of probabilistic models for attitude measurement are inspired by the deterministic scalogram model of Guttman (1944). Examples of such models are the Rasch (1980) model, which was proposed by Jansen (1983) as a model for the

measurement of attitudes, and the One Parameter Logistic Model (OPLM) (Verhelst & Glas, 1995) which was proposed by Klinkenberg (2001) for the same purpose. The basic assumption of the scalogram model is that the response of a subject depends on whether or not his attitude dominates the content of the item. Specifically, if a subject's attitude is larger than the item's content, the subject agrees with the item. Formally, this can be expressed as follows:

$$Y_{vi} = \begin{cases} 1 & \text{if } \theta_v \geq \delta_i \\ 0 & \text{if } \theta_v < \delta_i \end{cases}$$

That is, the response of a subject is an increasing function of his attitude. It follows from the model that, if the subjects are ordered according to their attitude, and the items are ordered according to their content, then the 1-responses in the data matrix have a triangular shape. Probabilistic models for attitude measurement that are inspired by the scalogram model have the property that the *probability* that a subject agrees with an item is an increasing function of the attitude of the subject. An example of such an IRF is given in Figure 1.

A slight extension of models with monotonically increasing IRFs involves that the IRFs are assumed to be monotone, but not necessarily increasing. This gives rise to a classification of attitude statements as belonging to one of two classes. For some statements, which are said to be *positively worded*, the IRF is increasing, and for other statements, which are said to be negatively worded, the IRF is decreasing. Recoding the responses to the negatively worded statements produces a data set that is in agreement with a model with only monotonically increasing IRFs.

This recoding is closely related to a scaling method known as the method of *summated ratings* (Likert, 1932). This method involves that the responses to a subset of the items are first recoded, and after this recoding, the total number of items a subject endorses (his sum score) is taken to reflect his attitude. It is known that,

if all IRFs are monotonically increasing, the expected value of the sum score is monotonically related to the attitude (e.g., Lord & Novick, 1968). Thus, we see that the method of summated ratings makes sense if (a) the IRFs of all items are monotone, and (b) it is known for each item whether its IRF is increasing or decreasing.

The main difference between the two approaches is whether the IRF is single-peaked or monotone. This difference is relevant as it gives rise to a different interpretation of the resulting scale. If the responses are governed by a single-peaked IRF, a subject can disagree with an item for one of two reasons, whereas if the responses are governed by a monotone IRF, a subject can only disagree for a single reason. In particular, with a single-peaked IRF, a subject can disagree because his attitude is too far to the left of the item's content (disagreeing from below), *as well as* because his attitude is too far to the right of the item's content (disagreeing from above). In contrast, with a monotone increasing IRF, a subject can only disagree because his attitude is too far to the left of the item's content. And with a monotone decreasing IRF, a subject can only disagree because his attitude is too far to the right of the item's content. In this paper a methodology by means of which one can study whether the IRFs of a set of items are monotone or single-peaked is presented. Since a monotone IRF can be regarded as a special case of a single-peaked IRF, with the peak at plus or minus infinity, this involves testing the null hypothesis that the peak is at, respectively, plus or minus infinity. In the first section, we introduce a new model with single-peaked IRFs that contains as a special case a model with monotone IRFs. In the second section, we show how the parameters can be estimated in a Bayesian framework. In the third section, we deal with the problem of testing the assumption that the IRFs are monotone. In the fourth section, the results of a simulation study are presented. In the fifth section an application is presented,

and in the last section some conclusions are drawn.

1. A Model Allowing for Monotone as well as Single-Peaked IRFs

1.1. The Model

To introduce the new model, we first consider the model of Coombs. In this model, a subject agrees with an item if his attitude is in a symmetric interval of width 2ξ about the item's location δ_i. We allow the width of this interval to be different for different items. In particular, we consider the model

$$Y_i = \mathcal{I}_{(\delta_i-\xi_i,\delta_i+\xi_i)}(\theta)$$

and reparameterize the model as follows: $\alpha_i = \delta_i - \xi_i$, and $\beta_i = \delta_i + \xi_i$. This reparameterization gives the following model:

$$Y_i = \mathcal{I}_{(\alpha_i,\beta_i)}(\theta) \tag{1}$$

As β_i goes to infinity, this model becomes identical to the Guttman model, and as α_i goes to minus infinity, this model becomes identical to a Guttman model with decreasing IRFs. We see that this deterministic model has both the scalogram model of Guttman and the parallelogram model of Coombs as special cases.

We now describe how this deterministic model can be turned into a probabilistic model. Assume that a person's attitude is subject to some random variation about his or her true attitude. In particular, let the *realized* attitude X be a random variable with expectation θ. Now, rather than evaluating whether θ is in a symmetric interval about the item's location, a person evaluates whether his realized attitude is in this interval. Since the realized attitude is not observable, it is integrated out of the model to obtain the probability that a person agrees with an item. If we assume that the realized attitude X has a logistic distribution with expectation θ and scale

parameter one, we obtain the following IRF:

$$P(Y_i = 1|\theta, \alpha_i, \beta_i) = \int \mathcal{I}_{(\alpha_i,\beta_i)}(x) \frac{\exp(x - \theta_v)}{[1 + \exp(x - \theta_v)]^2} dx$$
$$= \frac{\exp(\theta - \alpha_i)}{1 + \exp(\theta - \alpha_i)} - \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \quad , \tag{2}$$

where α_i is smaller than or equal to β_i. We also introduce an item precision parameter σ_i that influences the steepness of the IRF. In particular, by letting $1/\sigma_i$ be the scale parameter of the logistic distribution of X, we obtain the following IRF:

$$P(Y_i = 1|\theta, \alpha_i, \beta_i, \sigma_i) = \frac{\exp[\sigma_i(\theta - \alpha_i)]}{1 + \exp[\sigma(\theta - \alpha_i)]} - \frac{\exp[\sigma_i(\theta - \beta_i)]}{1 + \exp[\sigma(\theta - \beta_i)]} \quad . \tag{3}$$

As β_i goes to infinity, this IRF becomes identical to the IRF of the two-parameter logistic (2-PL) model (Birnbaum, 1968). Similarly, as α_i goes to minus infinity, this IRF becomes identical to the complement of the IRF of the 2-PL model. That is, if $\beta_i = \infty$ we obtain a monotone increasing IRF, whereas if $\alpha_i = -\infty$ we obtain a monotone decreasing IRF. As the item precision parameter σ_i goes to infinity, the IRF becomes identical to the deterministic model of Coombs or Guttman, depending on whether the boundary parameters α_i and β_i are finite or not.

In the context of the model introduced in this section, the null hypothesis that the IRF of item i is an increasing function of θ can be written as follows:

$$H_{01} : \alpha_i = -\infty$$

Similarly, the null hypothesis that the IRF of item i is a decreasing function of θ can be written as follows:

$$H_{02} : \beta_i = +\infty$$

1.2. Related Models

We now consider how model (3) is related to other models for attitude measurement. We first consider the relation between model (3) and the graded response

model (Samejima, 1969). Second, we consider its relation with the stochastic unfolding model derived from the partial credit model by Verhelst and Verstralen (1993). Third, we consider the relation with the model considered by Klinkenberg (2001). Fourth, we show that the model belongs to the class of latent response models (LRMs) (Maris, 1995).

It is readily seen that the model in (3) is equivalent to a graded response model for trichotomous responses, in which the two extreme responses are collapsed into a disagree response. That is, if U_{vi} satisfies the following graded response model for trichotomous responses:

$$P(U_{vi} = j|\theta_v, \delta_{i1}, \ldots, \delta_{i4}) = \frac{e^{\theta_v - \delta_{ij}}}{1 + e^{\theta_v - \delta_{ij}}} - \frac{e^{\theta_v - \delta_{i(j+1)}}}{1 + e^{\theta_v - \delta_{i(j+1)}}} \tag{4}$$

with $\delta_{i1} = -\infty$ and $\delta_{i4} = \infty$, then the following recoding of the responses:

$$Y_{vi} = \begin{cases} 1 & \text{if } U_{vi} = 1 \\ 0 & \text{if } U_{vi} = 0 \text{ or } U_{vi} = 2 \end{cases} \tag{5}$$

takes the graded response model into the model (3) with $\alpha_i = \delta_{i2}$ and $\beta_i = \delta_{i3}$.

The recoding that takes the graded response model into the model in (3) is equivalent to the one used by Verhelst and Verstralen in their formulation of a stochastic unfolding model from the partial credit model. Specifically, these authors assume that the random variable U_{vi} is in accordance with the partial credit model of Masters (1982) for trichotomous responses. That is, U_{vi} satisfies the following partial credit model for trichotomous responses:

$$P(U_{vi} = j|\theta_v, \delta_{i0}, \delta_{i1}, \delta_{i2}) = \frac{\exp\left(j\theta_v - \sum_{g=0}^{j} \delta_{ig}\right)}{\sum_{h=0}^{2} \exp\left(h\theta_v - \sum_{g=0}^{h} \delta_{ig}\right)}$$

Klinkenberg (2001) considers the following monotone IRT model for attitude measurement:

$$P(Y_{vi} = 1|\theta_v, \delta_i; a_i) = \frac{\exp\left(a_i[\theta_v - \delta_i]\right)}{1 + \exp\left(a_i[\theta_v - \delta_i]\right)}$$

where a_i is a known constant. If the a_i's are known *positive* constants, this model is known as the OPLM. In Klinkenbergs version of the model, the a_i's are assumed to be known, but are not required to be positive. It is easily seen that a negative a_i corresponds to a decreasing IRF. To see how this model is related to our model it is useful to reparameterize Klinkenbergs model as follows:

$$\alpha_i = -\infty$$

$$\beta_i = \delta_i$$

$$\sigma_i = |a_i|$$

if $a_i < 0$, and

$$\alpha_i = \delta_i$$

$$\beta_i = \infty$$

$$\sigma_i = |a_i|$$

if $a_i > 0$. With this reparameterization, we see that this model differs from our model with monotone IRFs in that the discrimination parameters σ_i is assumed to be known beforehand. That is, this model involves a further restriction on the model in (3).

The model in (3) belongs to the class of LRMs. LRMs are characterized by the property that the observed responses are the result of a mapping that takes unobserved (latent) responses as its argument. One way to formulate the model in (3) as a LRM is to consider the random variable U_{vi} as the latent response. These latent responses are modeled with the graded response model in (4) and the mapping that takes the latent responses into the observed responses is given in (5). In this way, the model in (3) is formulated as a LRM with *discrete* latent responses. The model in (3) can also be formulated as a LRM with *continuous* latent responses. Specfically,

in the above, we considered the random variable X_{vi} as the latent response. These latent responses are modeled with the logistic distribution with expectation θ_v and scale parameter σ_i. The mapping that takes the latent responses into the observed responses is given in (1), with θ replaced by the latent response. The main reason for considering the latter formulation is that it enables us to use a Bayesian estimation procedure that was specifically developed for LRMs with continuous latent responses by Maris and Maris (2002).

2. Parameter Estimation

Before introducing the method of parameter estimation, we make two preliminary remarks. First, as is true for all models involving subject parameters, the number of parameters increases with the number of subjects. That is, the attitude parameters θ_v are *incidental* parameters (Neyman & Scott, 1948). It is known that joint estimation of the structural item parameters and the incidental subject parameters in general does not lead to consistent estimates of the structural item parameters. This problem is overcome by considering the subjects as a random sample from a population characterized by an attitude distribution G. That is, we integrate the attitude parameters out of the model. Rather than estimating each subject's attitude, only the parameters of the attitude *distribution* are estimated. The attitude distribution is assumed to be normal with expectation μ and variance ν^2.

Second, the parameters of the model in (3) are not identified. The type of non-identification is the same as for the 2-PL model. That is, adding a constant to the item parameters α_i and β_i, as well as to the attitude parameter θ_v does not affect the probability. Also, multiplying the item parameters α_i and β_i as well as the attitude parameter θ_v by a constant, and dividing the item precision parameter σ_i by the same constant does not affect the probability either. This location and scale

non-identification is removed by setting the expectation of the attitude distribution to zero and its variance to one. The model in (3) also exhibits a reflection non-identification. This is seen by replacing α_i by $-\beta_i$, β_i by $-\alpha_i$, and θ_v by $-\theta_v$. To see how the reflection non-identification can be removed, first observe that the IRF has its maximum value when θ_v equals $(\alpha_i+\beta_i)/2$. From this it follows that if we impose the restriction $(\alpha_1 + \beta_1) > 0$, the IRF of item 1 will take its maximum value at a positive value of θ_v. Subject to this constraint, the IRFs can no longer be reflected.

For estimating the item parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma})$, a Bayesian method is used. The key feature of the Bayesian framework is that the parameters are considered as random variables. This allows us to study the properties of the posterior by drawing a sample from it. Typically, one is interested in the posterior expectation and variance of the parameters. Within the Bayesian framework it is also possible to do hypothesis testing. In the next section, this will be described in some detail.

In general, drawing *independent* samples from a high-dimensional distribution is a complicated problem. However, in the last decade, several methods have been developed for drawing *dependent* samples from such a distribution, which turns out to be much easier for many distributions. Such methods are called *Markov chain Monte Carlo* (MCMC) methods (Gelman, Carlin, Stern, & Rubin, 1995; Tanner, 1996). These methods involve (a) setting up a Markov chain which in the limit generates a dependent identically distributed (*did*) sample from the posterior, and (b) the use of the Monte Carlo method for estimating properties of the posterior from properties of the *did* sample. The particular MCMC method we use here is the DA-T-Gibbs sampler (Maris & Maris, 2002).

The prior distribution of the item parameters β_i and α_i is taken to be a bivariate normal distribution with expectation zero, standard deviation 10, and correlation zero, truncated to the set $\alpha_i \leq \beta_i$. The prior distribution of the item precision pa-

rameter σ_i is taken to be the normal distribution with expectation 1 and standard deviation 10, truncated to the positive real numbers. As before, the attitude parameters θ_v are assumed to be a random sample from a normal distribution with expectation one and standard deviation 1.

With this prior distribution, the following posterior distribution is obtained:

$$f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}|\mathbf{y}) \propto \prod_v \int \prod_i P(Y_{vi} = y_{vi}|\theta_v, \alpha_i, \beta_i, \sigma_i)\phi(\theta_v|0,1)d\theta_v$$

$$\prod_i \phi(\alpha_i|0,10)\phi(\beta_i|0,10)\mathcal{I}_{(-\infty,\beta_i)}(\alpha_i)\phi(\sigma_i|1,10)\mathcal{I}_{(0,\infty)}(\sigma_i)$$

The DA-T-Gibbs sampler requires that the observed data are augmented with continuous latent data. In particular, the DA-T-Gibbs sampler requires that the observed data can be conceived as the result of a mapping that takes continuous latent data as its argument. The goal of this data augmentation is to set up a joint posterior distribution (of latent data and parameters) from which it is easy to draw a sample. For generating a sample from this joint posterior, a Gibbs sampler is used. The basic idea behind the Gibbs sampler is to partition the complete set of random variables into a number of disjoint subsets, and to draw each subset in turn, conditional on the current values of all the other subsets and the observed data. This distribution of a subset, conditional on all the other subsets and the observed data, is called a *full conditional distribution.* For some models, sampling from the full conditional distributions becomes very simple after a transformation of the latent data. This transformation of variables distinguishes the DA-T-Gibbs sampler from other MCMC-methods.

2.1. Data Augmentation

The IRF in Equation 3 can be written as follows:

$$P(Y_i = 1|\theta, \alpha_i, \beta_i, \sigma_i) = \int \mathcal{I}_{(\alpha_i,\beta_i)}(x_{vi})\sigma_i^{-1} \frac{\exp\left[\sigma_i(x_{vi} - \theta_v)\right]}{\{1 + \exp\left[\sigma_i(x_{vi} - \theta_v)\right]\}^2} dx_{vi}$$

That is, the latent data are logistically distributed with location parameter θ_v and scale parameter σ_i^{-1}.

The joint posterior distribution can be written as follows:

$$\begin{aligned} f(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\theta},\mathbf{x}|\mathbf{y}) \propto & \prod_v \prod_i \mathcal{I}_{(\alpha_i,\beta_i)}(x_{vi})^{y_{vi}}(1-\mathcal{I}_{(\alpha_i,\beta_i)}(x_{vi})^{1-y_{vi}} \\ & \frac{\exp\left[\sigma_i(x_{vi}-\theta_v)\right]}{\left\{1+\exp\left[\sigma_i(x_{vi}-\theta_v)\right]\right\}^2} \\ & \prod_v \phi(\theta_v|0,1)\prod_i \phi(\alpha_i|0,10)\phi(\beta_i,0,10)\mathcal{I}_{(-\infty,\beta_i)}(\alpha_i) \\ & \phi(\sigma_i|1,10)\mathcal{I}_{(0,\infty)}(\sigma_i) \end{aligned} \tag{6}$$

2.2. Transformation

The key feature of the DA-T-Gibbs sampler is that, after some transformation of the latent data, the distribution of the transformed latent data does not depend on the parameters. It can be shown that, after this transformation, the full conditional distributions are all truncated distributions (Maris & Maris, 2002). Sampling from a truncated distribution is usually simple.

For the model in (3), the following transformation removes the parameters from the distribution of the latent data:

$$z_{vi} = \sigma_i(x_{vi}-\theta_v)$$

The joint posterior can now be written as follows:

$$\begin{aligned} f(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\theta},\mathbf{z}|\mathbf{y}) \propto & \prod_v \prod_i \mathcal{I}_{(\alpha_i,\beta_i)}(z_{vi}\sigma_i^{-1}+\theta_v)^{y_{vi}}(1-\mathcal{I}_{(\alpha_i,\beta_i)}(z_{vi}\sigma_i^{-1}+\theta_v))^{1-y_{vi}} \\ & \frac{\exp(z_{vi})}{(1+\exp(z_{vi}))^2} \\ & \prod_v \phi(\theta_v|0,1)\prod_i \phi(\alpha_i|0,10)\phi(\beta_i,0,10)\mathcal{I}_{(-\infty,\beta_i)}(\alpha_i) \\ & \phi(\sigma_i|1,10)\mathcal{I}_{(0,\infty)}(\sigma_i) \end{aligned} \tag{7}$$

Using a Gibbs sampler to sample from the joint posterior in (7) is much easier than for the joint posterior in (6). The reason for this is that the parameters and the latent data only co-occur in the range restrictions.

2.3. Gibbs Sampling

We now consider the Gibbs sampler used for sampling from the joint posterior distribution in (7). The t-th iteration of this Gibbs sampler involves the following steps:

1. *Imputation step:* For $v = 1 \cdots N$ and $i = 1 \cdots M$, sample $z_{vi}^{(t)}$ conditionally on $\left(\mathbf{y}, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\sigma}^{(t-1)}\right)$
2. *Posterior step:*
    1. For $v = 1 \cdots N$ sample $\theta_v^{(t)}$ conditionally on $\left(\mathbf{y}, \mathbf{z}^{(t)}, \boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\sigma}^{(t-1)}\right)$
    2. For $i = 1 \cdots M$ sample $\alpha_i^{(t)}, \beta_i^{(t)}$ conditionally on $\left(\mathbf{y}, \mathbf{z}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\sigma}^{(t-1)}\right)$
    3. For $i = 1 \cdots M$ sample $\sigma_i^{(t)}$ conditionally on $\left(\mathbf{y}, \mathbf{z}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}\right)$

We use $u^{(t)}$ to denote the realization of the random variable U in the t-th iteration. In general, the conditioning is on (a) the parameter values of the previous steps of the current cycle, and (b) the parameter values of the later steps of the previous cycle. For the Gibbs sampler described above, the full conditional of a subject (item) parameter does *not* depend on the other subject (item) parameters. This is reflected in our notation by suppressing those random variables on which the full conditional does not depend.

All full conditionals are truncated distributions. We only consider the full conditional of θ_v. The full conditional distribution of θ_v is a truncated normal distribution:

$$\phi_{\Xi}(\theta_v|0,1) \propto \left(\prod_i \mathcal{I}_{(\alpha_i,\beta_i)}(z_{vi}\sigma_i^{-1} + \theta_v)^{y_{vi}} \left[1 - \mathcal{I}_{(\alpha_i,\beta_i)}(z_{vi}\sigma_i^{-1} + \theta_v)\right]^{1-y_{vi}}\right) \phi(\theta_v|0,1) \tag{8}$$

This is a normal distribution truncated to the set Ξ:

$$\left\{\theta_v : \prod_i \mathcal{I}_{(\alpha_i,\beta_i)}(z_{vi}\sigma_i^{-1} + \theta_v)^{y_{vi}} \left[1 - \mathcal{I}_{(\alpha_i,\beta_i)}(z_{vi}\sigma_i^{-1} + \theta_v)\right]^{1-y_{vi}} = 1\right\} \quad ,$$

which depends on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\sigma}$, and $z_{v1}, \ldots, z_{vn}$. It can be shown that Ξ restricts the range of θ_v to the union of a number of disjoint intervals. This is easily seen by considering the contribution of a single item to this range restriction. Specifically, if y_{vi} equals one, we obtain that θ_v is in a closed interval, whereas if y_{vi} equals zero, it is restricted to the complement of this closed interval. The intersection of a number of closed intervals and a number of complements of closed intervals is the union of a number of disjoint intervals.We do not show the algebra behind this simplification because it would lead us too far.

We have shown how a *did* sample from the joint posterior distribution $f(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$ can be obtained. From this sample, one can obtain a sample from the posterior distribution of any subset of the random variables by dropping the other random variables from the chain.

2.4. An Item-Level Statistical Test for Monotonicity of the IRF

For the construction of an item-level statistical test for monotonicity of the IRF, we stay within the Bayesian framework. In particular, as a reference distribution for our test we make use of the *posterior-predictive distribution.* This type of model evaluation is called a *posterior-predictive check* (PPC) (Rubin, 1984).

A PPC evaluates whether the observed data are similar to replicated data that are generated under the model. To understand the rationale behind PPC's, it is useful to consider this method as a two-step procedure. In the first step, the parameters are assumed to be known, and one computes the probability of observing a test statistic that is more extreme than the one that is actually observed. This probability is called the predictive p-value. This predictive p-value can be computed

analytically or approximated via simulation. The latter option is almost always more convenient. In the second step, the predictive p-value (which is computed conditionally on some parameter value) is averaged over the posterior distribution of the parameters, resulting in a *posterior-predictive p-value.*

In its standard form, as originally conceived by Rubin (1984), the posterior distribution of the parameters under the null hypothesis is used. This is not always a good choice, as will be illustrated by a simple example. This example also points the way to the solution of the problem. Consider the problem of testing whether a random variable X is normally distributed with expectation zero and unknown variance σ^2. The sample mean $\overline{x}_n$ will be used as a test statistic. Bayarri and Berger (2000) show that, with the non-informative prior for σ^2, the posterior-predictive distribution of the sample mean is a scaled $\mathcal{T}$-distribution with n degrees of freedom and scale parameter

$$\sqrt{\frac{1}{n}\left(\frac{1}{n}\sum_i x_i^2\right)} \quad .$$

It is readily seen that the posterior predictive p-value of the observed sample mean is the following:

$$P\left(|T_n| > \frac{\sqrt{n}\overline{x}_n}{\sqrt{n^{-1}\sum_i x_i^2}}\right) \quad , \tag{9}$$

where T_n denotes a $\mathcal{T}$ distributed random variable with n degrees of freedom. If the null hypothesis does not hold, then $\sqrt{n^{-1}\sum_i x_i^2}$ can be large, resulting in a large posterior predictive p-value.

This is not a finite sample problem. To see this, start from the fact that, for large n, the posterior-predictive distribution of the sample mean is the normal distribution with expectation zero and variance $n^{-1}(\mu^2+\sigma^2)$, where μ denotes the unknown expectation of X. Consequently, for large n, the comparison is essentially between observed data that are normally distributed with expectation μ and variance σ^2,

and replicated data that are normally distributed with expectation zero and variance $(\mu^2 + \sigma^2)$. This comparison seems pointless unless μ equals zero (i.e., the null hypothesis holds).

The source of the problem is the posterior distribution of the nuisance parameter σ^2, which is an inverse gamma distribution with shape parameter equal to $n/2$ and scale parameter equal to $(\sum_i x_i^2)/2$. This scale parameter is too large unless μ equals zero. This also points to the solution of the problem: Use the posterior distribution of the nuisance parameter σ^2 under the *unrestricted* model (i.e., both μ and σ^2 unknown). This posterior is the inverse gamma distribution with shape parameter $(n-1)/2$ and scale parameter $(n-1)/2s_n^2$, in which s_n^2 is the sample variance on $(n-1)$ degrees of freedom. The posterior predictive distribution of the sample mean now becomes the scaled $\mathcal{T}$ distribution with $(n-1)$ degrees of freedom and scale parameter $\sqrt{n^{-1}s_n^2}$. The resulting posterior predictive p-value of the observed sample mean,

$$P\left(|T_{n-1}| > \frac{\sqrt{n}\bar{x}_n}{\sqrt{s_n^2}}\right) \quad , \tag{10}$$

is equal to the frequentist p-value. Consequently, for large n, the comparison is essentially between observed data that are normally distributed with expectation μ and variance σ^2, and replicated data that are normally distributed with expectation zero and variance σ^2. Thus, to avoid the problem that the replicated data might not be comparable to the observed data if the null posterior is used, we propose to use the posterior distribution of the parameters under the unrestricted model in (3).

Contrary to the frequentist framework, in the Bayesian framework, there is no problem with using a test statistic that depends on the nuisance parameters as well as on the data (Meng, 1994). Such a test statistic is called a *discrepancy measure*. The posterior predictive distribution of such a discrepancy measure is obtained in exactly the same way as the posterior-predictive distribution of a test statistic that

depends on the data only.

We now consider the discrepancy measure that is used for testing the two null hypotheses in which we are interested: The IRF is monotonically increasing or decreasing. This discrepancy measure is the covariance between the observed responses on item i and the latent trait Θ. In the Appendix, we show that, for given β_i and σ_i, this covariance is minimized if α_i equals minus infinity. Similarly, this covariance is maximized if β_i equals plus infinity.

To evaluate whether the observed value of the covariance between the responses and the attitude is too large (small) under the null hypothesis that the IRF is decreasing (increasing), we compare it with its posterior predictive distribution. The posterior predictive distribution of this discrepancy measure is computed with the nuisance parameters obtained from their posterior distribution under the *unrestricted* model in (3).

3. Simulation Study

We now consider the results of a small simulation study conducted to evaluate the operating characteristics of the item-level statistical test for monotonicity introduced in the previous section. The simulation study is set up as follows. A set of 24 items is used, of which 6 items have an increasing IRF, 6 a decreasing one, and 12 are single-peaked. The items differ in their location on the attitude continuum and their discriminatory power. The 12 items with a single-peaked IRF also differ with respect to the degree of single-peakedness. By this, we mean the extent to which the maximum of the IRF falls in one of the tails of the population distribution of the attitudes (here, the standard normal). It is easily seen that, if an IRF's mode becomes more extreme, it becomes more difficult to distinguish it from a monotone IRF. Details concerning the item parameters used can be found in Table 1. With

the item parameters in this table, and a random sample of size 245 from a standard normal distribution[1], 100 data sets were generated according to the model in (3).

3.1. Convergence of the Markov Chain

We first consider a preliminary technical issue involved in the evaluation of a posterior-predictive p-value. The Monte Carlo estimate of this p-value is valid only if the parameters used for generating the replicated data are a draw from the posterior distribution. The MCMC method presented in the above achieves this goal only in the limit. Therefore, the initial iterations, the so-called burn-in cycles, are usually discarded and the rest of the Markov chain is considered to be a *did* sample from the posterior. To evaluate whether a sufficient number of iterations has been discarded in order for the Markov chain to have converged, we use the $\sqrt{\widehat{R}}$ statistic of Gelman and Rubin (1992). This convergence diagnostic is computed using multiple independent replications of the Markov chain. Specifically, the Markov chain is considered to be converged if the value of $\sqrt{\widehat{R}}$ is below 1.2 for each parameter of the model.

For each of the 100 data sets, a Markov chain was run for 60000 iterations. Starting values were drawn from the prior distribution. Of the 60000 iterations, the first 20000 were discarded to ensure convergence. To evaluate whether a burn-in of 20000 was sufficient to reach convergence, for 20 of the 100 data sets we ran 10 replications of the Markov chain using independent draws from the prior distribution as starting values. For these 20 data sets, we computed the convergence diagnostic proposed by Gelman and Rubin (1992). The largest of these convergence diagnostics was 1.05, indicating that a burn-in of 20000 is sufficient for the Markov chain to have converged.

[1]The numbers of 245 subjects and 24 items are identical to the numbers of subjects and items in the application on which is reported in the next section.

TABLE 1.

Item parameters used to generate the data in the simulation study

item number	α_i	β_i	σ_i	
1	$-\infty$	-1.5	1	decreasing
2	$-\infty$	-1	2	decreasing
3	$-\infty$	-0.5	3	decreasing
4	$-\infty$	0	4	decreasing
5	$-\infty$	0.5	5	decreasing
6	$-\infty$	1	6	decreasing
7	-2.5	0	4	single-peaked
8	-2	-0.5	2	single-peaked
9	-1.5	1	5	single-peaked
10	-1	2	6	single-peaked
11	-0.5	1	1	single-peaked
12	0	0.5	8	single-peaked
13	0.5	1	5	single-peaked
14	1	3	6	single-peaked
15	1.5	2	3	single-peaked
16	0	1	6	single-peaked
17	0.5	1.5	9	single-peaked
18	1	2.5	9	single-peaked
19	-1.5	∞	1	increasing
20	-1	∞	2	increasing
21	-0.5	∞	3	increasing
22	0	∞	4	increasing
23	0.5	∞	5	increasing
24	1	∞	6	increasing

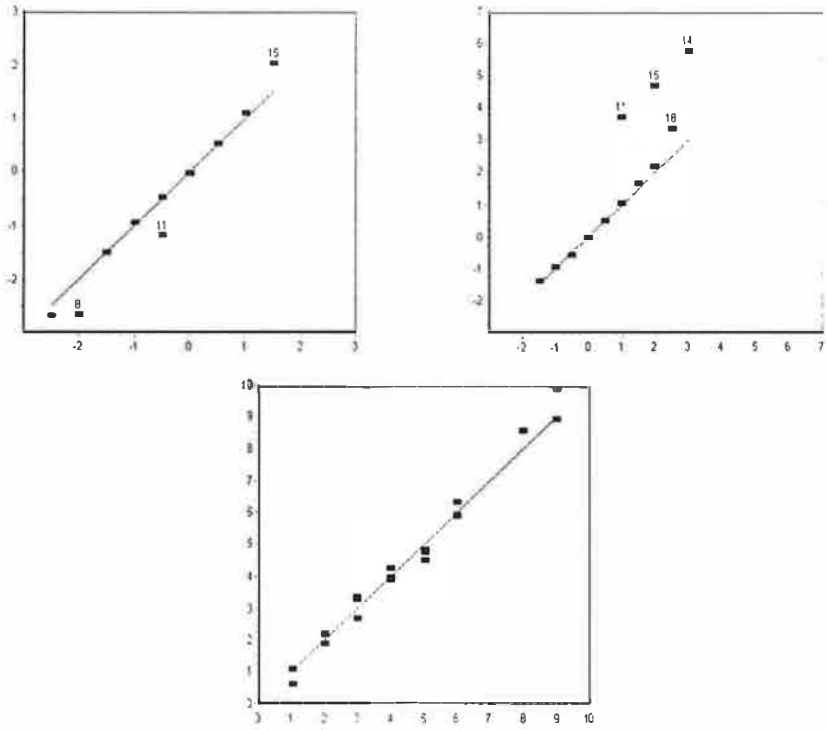

FIGURE 2.

True versus average EAP estimates of parameter: α (upper left), β (upper right), and σ (lower middle)

3.2. Parameter Recovery

Before we discuss the results of the item-level test, we first check the recovery of the parameters. In Figure 2, the true parameters are plotted against their average EAP estimate, with the average taken over replications of the data. Items with a true location parameter equal to plus or minus infinity are not included in these figures. For the six items with a decreasing IRF, the average EAP estimates of the α parameter are between -7.02 and -7.26. These estimates are sufficiently far in the tail of the standard normal population distribution for the IRF evaluated at the estimated parameters to be indistinguishable from a decreasing IRF. For the six items with an increasing IRF, the average EAP estimates of the β parameters are between 7.13 and 7.86. These estimates are sufficiently far in the tail of the population distribution for the IRF evaluated at the estimated parameters to be indistinguishable from an increasing IRF.

It is seen in Figure 2 that, except for the location parameters of items 8, 11, 14, 15 and 18, the recovery is excellent. The IRF evaluated at the true parameter values for items 8, 11, 14, 15 and 18 is given in Figure 3. For item 8, the poor recovery is due to the fact that the true α parameter is in the left tail of the population distribution. As a consequence, it is difficult to distinguish its IRF from a decreasing IRF. For items 14, 15 and 18 the poor recovery results from the fact that the true β parameter is in the right tail of the population distribution. As a consequence, it is difficult to distinguish its IRF from an increasing IRF. For item 11, both the true α and β parameter are *not* in the tail of the population distribution. The poor recovery of both the α and β parameter is due to the fact that the value of the IRF changes little over the range from -2 to +2. This implies that the responses to such an item provide little information about the item location parameters.

The results of the simulation study indicate that the recovery of the parameters

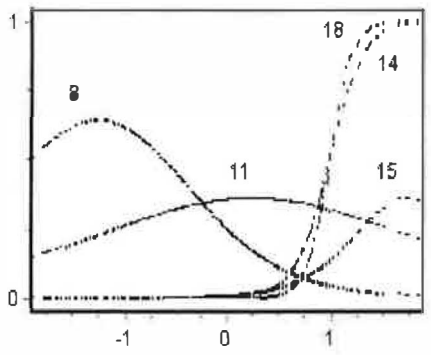

FIGURE 3.

True IRF evaluated at true parameter values for items 8, 11, 14, 15, and 18

is good, unless (a) one of the item location parameters is in the tail of the population distribution, or (b) the value of the IRF changes little over the attitude continuum. If one of the item location parameters is in the tail of the population distribution, the estimated parameter tends to be too extreme (positive or negative).

3.3. Item-level Tests

We now turn to the results of the item-level tests. In Table 2, a summary of the results is given. This table contains the mean and the standard deviation of the posterior-predictive p-values, and the proportion p-values smaller than or equal to 0.05, taken over replications of the data. We first examine whether the item-level tests succeed in detecting the 12 items with a single-peaked IRF. We reject a hypothesis if the posterior-predictive p-value is smaller than or equal to 0.05. It is seen in Table 2 that, except for items 8, 11, 14, 15 and 18, both p-values are on average smaller than 0.05. The performance of the item-level tests is worst for items 8, 11, 14, 15 and 18. This is not surprising because (a) item 8 is *almost* a decreasing item (its α parameter is located in the extreme left tail of the attitude distribution), (b) items 14, 15 and 18 are *almost* increasing items (their β parameters are all in the extreme right tail of the attitude distribution), and (c) item 11 has a low

discriminatory power.

Next, we consider the p-values for the items with a monotone IRF. It is seen in Table 2 that for the incorrect null hypothesis, the average p-value is 0 for all items, and for the correct hypothesis, the p-values are tightly concentrated about 0.5. The latter indicates that the probability that the null hypothesis is incorrectly rejected is much smaller than 0.05. This in contrast with a frequentist p-value which is uniformly distributed under the null hypothesis, and hence leads to 5 percent incorrect rejections.

We conclude that the item-level tests (a) correctly reject the null hypothesis, unless an item with a single-peaked IRF is too close to a monotone IRF, and (b) almost never incorrectly reject the null hypothesis.

4. Application

We now apply our item-level tests to data collected by Roberts (1995)[2] with a questionnaire on capital punishment. The questionnaire consists of the 24 items in Table 3, originally published by Thurstone (1932) and later republished by Shaw and Wright (1967). Roberts (1995) asked 245 subjects to express their agreement with the items on a six point rating scale. The scale ranged from *strongly disagree* to *strongly agree*. The responses were dichotomized by recoding the first three response categories as zero (disagree) and the last three categories as one (agree).

We consider the results of the item-oriented tests introduced in the above. The p-values are computed as before. Ten replications of a Markov chain of length 60000 were generated to determine whether a burn-in of 20000 is sufficient for the Markov chain to have converged. The largest value of the convergence diagnostic was 1.18, indicating that the Markov chains have converged. Using the remaining 40000 draws

[2]These data were published on the internet at http://www.musc.edu/cdap/Roberts

TABLE 2.

Mean, standard deviation (sd), and proportion of p-values smaller than 0.05 (pr(reject)) of the distribution of the p-values for the item-level tests of monotonicity.

Item	Decreasing			Increasing		
	mean	sd	pr(reject)	mean	sd	pr(reject)
1	0.417	0.064	0.00	0.000	0.000	1.00
2	0.461	0.059	0.00	0.000	0.000	1.00
3	0.482	0.044	0.00	0.000	0.000	1.00
4	0.487	0.022	0.00	0.000	0.000	1.00
5	0.495	0.020	0.00	0.000	0.000	1.00
6	0.507	0.014	0.00	0.000	0.000	1.00
7	0.042	0.066	0.77	0.000	0.000	1.00
8	0.054	0.083	0.66	0.000	0.000	1.00
9	0.000	0.000	1.00	0.000	0.000	1.00
10	0.000	0.000	1.00	0.028	0.039	0.82
11	0.023	0.043	0.84	0.143	0.125	0.31
12	0.000	0.000	1.00	0.000	0.000	1.00
13	0.000	0.000	1.00	0.000	0.000	1.00
14	0.000	0.000	1.00	0.423	0.123	0.00
15	0.000	0.000	1.00	0.278	0.138	0.05
16	0.000	0.000	1.00	0.000	0.000	1.00
17	0.000	0.000	1.00	0.000	0.000	1.00
18	0.000	0.000	1.00	0.235	0.124	0.05
19	0.000	0.000	1.00	0.471	0.044	0.00
20	0.000	0.000	1.00	0.495	0.030	0.00
21	0.000	0.000	1.00	0.496	0.018	0.00
22	0.000	0.000	1.00	0.488	0.021	0.00
23	0.000	0.000	1.00	0.496	0.020	0.00
24	0.000	0.000	1.00	0.505	0.017	0.00

from the posterior, a set of replicated data was generated for every 50-th draw, resulting in 800 replicated data sets. For each of these replicated data sets the discrepancy measures are computed and compared with the discrepancy measures computed with the observed data. The resulting p-values are given in Table 3.

In Table 3, the items are grouped according to their p-values. The first group of items are all items for which the hypothesis of an increasing IRF is *not* rejected, whereas the hypothesis of a decreasing IRF *is* rejected. The second group of items are all items for which the hypothesis of a decreasing IRF is *not* rejected, whereas the hypothesis of an increasing IRF *is* rejected. The third group consists of items for which both tests are significant. These items have a single-peaked IRF. The fourth group consists of items for which neither of the tests is significant. These items have a flat IRF. We see that, of the 24 statements, 18 have a monotone IRF, 4 have a single-peaked IRF, and 2 items have a flat IRF.

We now examine whether these results are consistent with the content of the statements. The items in the first group in Table 3 are all favorable towards capital punishment, and the items in the second group are all opposed to capital punishment. Of the items in the third group, the first three (6, 9, and 18) express an ambivalent position. That is, to agree with these items, a subject has to agree both with a positively worded statement (e.g., Capital punishment is necessary in our imperfect civilization) and with a negatively worded statement (e.g., Capital punishment is wrong). For such items, it is not too surprising that their IRF is single-peaked. It should be observed that also items 1 and 3 express such an ambivalent position, but their IRF is nevertheless monotone. The fourth item in the third group (item 22) expresses an indifferent position towards capital punishment. Again, it is not surprising that the IRF of this statement is single-peaked. The items in the fourth group have in common that (a) very few subjects agree with them, and (b) they

express a very extreme position in favor of capital punishment.

TABLE 3.

Posterior-predictive p-values for the two item-level statistical tests based on the covariance between the item responses and the attitudes. The items are grouped according to their posterior-predictive p-values.

1	Capital punishment may be wrong but it is the best preventative to crime.	0	0.451
3	I think capital punishment is necessary but I wish it were not.	0	0.481
4	Any person, man or woman, young or old, who commits murder, should pay with his own life.	0	0.471
10	We must have capital punishment for some crimes.	0	0.501
17	Capital punishment is just and necessary.	0	0.45
20	Capital punishment gives the criminal what he deserves.	0	0.474
23	Capital punishment is justified only for premeditated murder.	0	0.335
24	Capital punishment should be used more often than it is.	0	0.475
2	Capital punishment is absolutely never justified.	0.515	0
5	Capital punishment cannot be regarded as a sane method of dealing with crime.	0.463	0
8	Capital punishment has never been effective in preventing crime.	0.51	0
12	I do not believe in capital punishment under any circumstances.	0.321	0
13	Capital punishment is not necessary in modern civilization.	0.5	0
14	We can't call ourselves civilized as long as we have capital punishment.	0.484	0

TABLE 3.
Continued

15	Life imprisonment is more effective than capital punishment.	0.44	0
16	Execution of criminals is a disgrace to civilized society.	0.504	0
19	Capital punishment is the most hideous practice of our time.	0.441	0
21	The state cannot teach the sacredness of human life by destroying it.	0.544	0
6	Capital punishment is wrong but is necessary in our imperfect civilization.	0	0
9	I don't believe in capital punishment but I'm not sure it isn't necessary.	0.006	0
18	I do not believe in capital punishment but it is not practically advisable to abolish it	0	0
22	It doesn't make any difference to me whether we have capital punishment or not.	0	0
7	Every criminal should be executed.	0.09	0.487
11	I think the return of the whipping post would be more effective than capital punishment.	0.369	0.204

5. Conclusion

In this paper, a methodology for evaluating whether the relation between the subject's attitude and the probability that a subject agrees with a statement is monotone or single-peaked was presented. These two possible relations are not mutually exclusive. Specifically, a monotone IRF is a special case of a single-peaked IRF (one with its maximum at plus or minus infinity). Consequently, it is possible to reject the hypothesis that the relation is monotone in favor of the more general hypothesis that it is single-peaked. With the item-level tests introduced in the above it is *not* possible to reject the hypothesis that the IRF is single-peaked.

Our method for evaluating whether the IRF of a single item is monotone or single-peaked differs from existing Bayesian procedures for evaluating model fit. The difference is that the posterior-predictive p-value is computed using the marginal posterior distribution of the nuisance parameters under the *unrestricted* model rather than the posterior distribution under the null hypothesis. In a small simulation study, we found that this new method gave satisfactory results. More research is needed however to establish the statistical properties of this alternative approach.

In our application, concerned with capital punishment, we found that most of the statements are in agreement with a model with monotone IRFs. We found that the content of these statements was in agreement with the results of the item-level tests. Moreover, the statements for which the IRF is not monotone reflect either an ambivalent position, or indifference towards capital punishment. Based on a single application one can, of course, not conclude that IRFs of attitude statements are typically monotone. However, we hope to have shown that it is important and worthwhile to evaluate whether attitude statements conform to the assumption of monotone IRFs.

Appendix

In this appendix we show that, under the model in (3), the covariance between the responses to an item and the latent trait θ is minimized at $\alpha_i = -\infty$. Showing that this covariance is *maximized* at $\beta_i = \infty$ proceeds along the same lines.

Our starting point is a reformulation of the expression for the covariance that turns out to be convenient for what is to be shown. First, assume, without loss of generality, that, in the population, Θ has zero expectation. The covariance can be expressed in the following way:

$$\begin{aligned} \mathrm{COV}(Y, \Theta|\alpha, \beta, \sigma) &= E(Y\Theta|\alpha, \beta, \sigma) \\ &= \int \theta P(Y = 1|\theta, \alpha, \beta, \sigma) f(\theta) d\theta \end{aligned}$$

Since $P(Y = 1|\theta, \alpha, \beta, \sigma)$ equals $P(\alpha \leq X \leq \beta|\theta, \sigma)$ the covariance can be rewritten as follows:

$$\begin{aligned} \mathrm{COV}(Y, \Theta|\alpha, \beta, \sigma) &= \int \theta \left(\int_{\alpha}^{\beta} f(x|\theta, \sigma) dx \right) f(\theta) d\theta \\ &= \int_{\alpha}^{\beta} \left(\int \theta f(\theta|x, \sigma) d\theta \right) f(x|\sigma) dx \\ &= \int_{\alpha}^{\beta} E(\Theta|x, \sigma) f(x|\sigma) dx \end{aligned} \tag{11}$$

To show that the covariance is smallest if $\alpha = -\infty$, we show that the covariance is a *single-peaked* function of α. This implies that it is smallest at the boundary points $\alpha = -\infty$ and $\alpha = \beta$. At $\alpha = \beta$, the covariance equals zero, whereas at $\alpha = -\infty$ it is smaller than zero (Ross, 1996, Proposition 7.2.1). That is, the covariance is smallest if $\alpha = -\infty$. That the covariance is a single-peaked function of α means that its first derivative with respect to α has one change of sign, and the change is from positive to negative. Differentiating the covariance with respect to α gives the

following result:

$$\frac{\partial}{\partial \alpha}\mathrm{COV}(Y,\Theta|\alpha,\beta,\sigma) = -E(\Theta|X=\alpha,\sigma)f(X=\alpha|\sigma) \tag{12}$$

Observe that $f(X=\alpha|\sigma)$ is always positive. If the distribution of the latent random variable X is a log-concave location family, as is the case for the logistic distribution in our model, then it follows from Lemma 1 that $E(\Theta|X=\alpha)$ has one change of sign, and the change is from negative to positive. As a consequence, the first derivative of the covariance with respect to α has one change of sign also, and the change is from positive to negative, as was to be shown.

Theorem 1. If the distribution of X is a log-concave location family with location parameter θ, then $E(\Theta|x)$ has one change of sign, from negative to positive.

Proof. It is known that a log-concave location family has *monotone likelihood ratio* (Lehmann, 1986, Example 1, p.509):

$$\frac{f(x_1-\theta_2)}{f(x_2-\theta_2)} \leq \frac{f(x_1-\theta_1)}{f(x_2-\theta_1)} \quad \text{for all } \theta_2<\theta_1 \text{ and } x_2<x_1$$

Multiplying the left and right hand sides with $f(x_2)/f(x_1)$ we see that also the conditional distribution of Θ (conditional on X) has monotone likelihood ratio:

$$\frac{\frac{f(x_1-\theta_2)g(\theta_2)}{f(x_1)}}{\frac{f(x_2-\theta_2)g(\theta_2)}{f(x_2)}} = \frac{g(\theta_2|x_1)}{g(\theta_2|x_2)} \leq \frac{g(\theta_1|x_1)}{g(\theta_1|x_2)} = \frac{\frac{f(x_1-\theta_1)g(\theta_1)}{f(x_1)}}{\frac{f(x_2-\theta_1)g(\theta_1)}{f(x_2)}} \quad \text{for all } \theta_2<\theta_1 \text{ and } x_2<x_1$$

It is also known that if the distribution of Θ conditionally on X has monotone likelihood ratio, then $E(\Theta|x)$ has a single change of sign, and the change is from negative to positive (Lehmann, 1986, lemma 2, p.85). This completes the proof.

References

Andrich, A., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied psychological measurement*,

17(3), 253-276.

Bayarri, M., & Berger, J. (2000). P values for composite null models. *Journal of the American Statistical Association*, *95*(452), 1127-1142.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading: Addison-Wesley.

Coombs, C. H. (1964). *A theory of data.* New York: John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139-150.

Hoijtink, H. (1990). *PARELLA, measurement of latent traits by proximity items.* Leiden: DSWO-press.

Jansen, P. G. W. (1983). *Rasch analysis of attitudinal data.* Unpublished doctoral dissertation, Katholieke Universiteit Nijmegen.

Klinkenberg, E. L. (2001). A logistic IRT model for decreasing and increasing item characteristic curves. In A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (p. 173-192). New York: Springer.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (Second ed.). New York: John Wiley & Sons.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 5-53.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523-547.

Maris, G., & Maris, E. (2002). A MCMC-method for models with continuous latent responses. *Psychometrika*, *67*, 335-350.

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

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142-1160.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1-32.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press.

Roberts, J. (1995). *Item response theory approaches to attitude measurement.* Unpublished doctoral dissertation, University of South Carolina, Columbia.

Ross, S. M. (1996). *Stochastic processes* (Second ed.). New York: John Wiley & sons.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika-Monograph-Supplement*, *34*.

Shaw, M., & Wright, J. (1967). *Scales for the measurement of attitudes.* New York: McGraw-Hill.

Tanner, M. A. (1996). *Tools for statistical inference* (Third ed.). New york: Springer-Verlag.

Thurstone, L. (1932). *Motion pictures and the attitudes of children.* Chicago: University of Chicago Press.

Verhelst, N., & Glas, C. A. W. (1995). The one parameter logistic model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (p. 215-238). New York: Springer Verlag.

Verhelst, N., & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitatieve Methoden*, *43*, 73-92.