



ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

Bayesian estimation of a multidimensional additive graded response model for correlated traits

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Bayesian estimation of a multidimensional additive graded response model for correlated traits / I. Martelli; M. Matteucci; S. Mignani. - In: COMMUNICATIONS IN STATISTICS. SIMULATION AND COMPUTATION. - ISSN 0361-0918. - STAMPA. - 45:5(2016), pp. 1636-1654. [10.1080/03610918.2014.932804]

This version is available at: <https://hdl.handle.net/11585/315115> since: 2016-05-26

Published:

DOI: <http://doi.org/10.1080/03610918.2014.932804>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

(Article begins on next page)

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

This is the final peer-reviewed accepted manuscript of:

Irene Martelli, Mariagiulia Matteucci & Stefania Mignani (2016) Bayesian Estimation of a Multidimensional Additive Graded Response Model for Correlated Traits, Communications in Statistics - Simulation and Computation, 45:5, 1636-1654, DOI: 10.1080/03610918.2014.932804

The final published version is available online at:
<https://doi.org/10.1080/03610918.2014.932804>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

BAYESIAN ESTIMATION OF A MULTIDIMENSIONAL ADDITIVE GRADED RESPONSE MODEL FOR CORRELATED TRAITS

Irene Martelli, Mariagiulia Matteucci, and Stefania Mignani

Department of Statistical Sciences

University of Bologna

Bologna, Italy

irene.martelli2@unibo.it, m.matteucci@unibo.it, stefania.mignani@unibo.it

Key Words: item response theory; additive graded response model; MCMC.

ABSTRACT

The aim of the paper is to propose a Bayesian estimation through Markov chain Monte Carlo of a multidimensional item response theory model for graded responses with an additive structure with correlated latent traits. A simulation study is conducted to evaluate the model parameter recovery under different conditions (sample size, test and subtest length, number of response categories, and correlation structure). The results show that the parameters are well reproduced when the sample size is sufficiently large ($n = 1000$), while the worst recovery is observed for small sample size ($n = 500$), and 4 response categories with a short number of test items.

1. INTRODUCTION

In the field of educational and psychological measurement, typically a test consisting of a set of items is submitted to a sample of individuals to infer the subjects' unobservable characteristics. To this aim, item response theory (IRT) (Hambleton and Swaminathan, 1985; van der Linden and Hambleton, 1997) represents the main methodological approach that allows to estimate both the item psychometric properties and the subjects' scores. In the past, one of the most common assumption was unidimensionality, i.e. the presence of a

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

single latent trait, commonly denoted ability in the IRT literature, underlying the response process. However, real data often suggest a multidimensional structure and, consequently, tests should be composed of different subtests in order to infer several latent traits.

For this reason, the attention has been recently devoted to models that include more than one latent trait, the so called multidimensional IRT (MIRT) models (see e.g., Reckase, 2009). These models perform better than separate unidimensional models in fitting the subtests because they are able to describe the data complexity, taking into account correlated traits and also a possible hierarchical structure of abilities.

Within the multidimensional context, different approaches are possible: explorative models where all latent traits are allowed to load on all item response variables or confirmatory models where the relations between the observed and the latent variables are specified in advance. By adopting a confirmatory approach, it is also possible to assume the concurrent presence of general and specific latent traits underlying the response process (Sheng and Wikle, 2008). Another distinction is among non compensatory and compensatory models, where in the compensatory model a lack in one trait can be compensated by another (Reckase, 2009).

Moreover, in many studies, data show a hierarchical structure with dimensions on different levels that vary in their degree of generality and abstraction. On the highest levels of these hierarchies, dimensions represent overall trait levels while, on lower levels of the hierarchy, dimensions represent more specific traits. Additive models and higher-order models are two alternative approaches for dealing with items that assess several related domains that are hypothesized to comprise a general structure. Additive models are potentially applicable when there is a general factor and there are multiple domain specific factors, each of which is hypothesized to account for the unique influence of the specific domain over the general factor (Chen et al., 2006). In addition, researchers may be interested both in the domain specific factors as well as in the general factor. Higher-order models are potentially applicable when the lower-order factors are substantially correlated with each other and there is a higher-order factor that is hypothesized to account for the relationship among

the lower-order factors. In the additive model, we can directly examine the strength of the relationship between the domain specific constructs and their associated items, as the relationship is reflected in the loadings, whereas these relationships cannot be directly tested in the higher-order model as the domain specific factors are represented by the disturbances of the first-order factors (Chen et al., 2006).

Finally, different models can be distinguished according to the nature of the observable variables. In educational testing, there are typically dichotomous items (correct/wrong) while in psychological and behavioral researches the item scores are often ordinal, expressing a judge or an agreement. Different models have been developed according to the number of estimated parameters (partial credit models, graded response models, etc...). In a multidimensional context, while models for binary data are frequently applied, models for ordinal data are still uncommon and unfortunately were developed only for uncorrelated traits.

For these reasons, in this work we propose an extension of the unidimensional graded response model (Samejima, 1969) for ordinal data to a multidimensional additive structure with correlated traits. A further innovative and important aspect of our proposal deals with the estimation procedure, in fact, we propose a Markov chain Monte Carlo (MCMC) procedure for parameter estimation which has been implemented by using the software OpenBUGS (Lunn et al., 2000).

For years, the standard methodology has been mainly involving marginal maximum likelihood (MML). Unfortunately, this estimation method may be computational heavy due to the approximation of integrals involved in the likelihood function, especially for increasingly complex models. Moreover, the success of MML estimation based on the EM algorithm strongly depends on the choice of starting values. On the other hand, MCMC in a fully Bayesian framework is powerful for complicated models where the probabilities or expectations are intractable by analytical methods or other numerical approximation. Moreover, it has the advantage of estimating item parameters and individual abilities jointly and it is proved to be more accurate and efficient in parameter estimation compared with the usual MML method for IRT models (Albert, 1992). Lastly, the method is suitable to model the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

dependencies among parameters and sources of uncertainty and, at the end of the analysis, the user has access to the entire posterior distribution of every parameter, not just to a point estimate and standard error.

In order to study the performance of the additive graded response model, we conducted a simulation study by varying different conditions (sample size, test and subtest lengths, number of response categories, and correlation structure).

The paper is structured as follows. In Section 2, we introduce MIRT models and especially the new additive graded response model with correlated traits. In Section 3, we describe the simulation study conditions and the results. A discussion in Section 4 concludes the paper.

2. MULTIDIMENSIONAL ITEM RESPONSE THEORY MODELS

Multidimensional item response theory (MIRT) models are used whenever two or more latent traits are encompassed in the observed response for a test item, reflecting a complex latent space. Depending on the specification of the relationship among the latent and the observed variables, MIRT models can be classified by using different criteria.

With reference to the available information at the model specification step, exploratory or confirmatory multidimensional models can be developed. By following an exploratory approach, no prior knowledge in terms of relationship between items and latent traits is included in the model. Conversely, when the number of latent abilities is specified in advance, the method is not merely explorative and we are in a confirmatory context. In line with the confirmatory approach, not only the number of latent variables is pre-specified but also their relationships with the items: the researcher uses prior knowledge to define which items load on which traits.

Another classification criterion of MIRT models depends on the way the elements of the θ -vector of latent abilities are combined to obtain the probability of response to the items. In compensatory models, a linear combination of latent ability values is used in the specification of the response probabilities. This approach implies that different combinations of elements in the ability vector can yield the same sum, raising a compensation effect: if a

θ -value is low, but another one is appropriately high, the sum can be the same. Differently, in noncompensatory models, distinct latent abilities used to solve an item are separated and each part is used as an unidimensional model. The global probability is then obtained as the product of the probabilities of each unidimensional part, nonlinearity raises in relation to the use of the product of such probabilities and the compensation property does not hold (Reckase, 2009).

By adopting a confirmatory approach, three most common multidimensional underlying latent structures can be distinguished. These structures are illustrated in Figure 1 for a bidimensional case, where circles represent latent traits, squares represent observed item responses and subtests are indicated with dashed lines.

[FIGURE 1]

When each latent ability affects only the response to items belonging to its specific subset, and the abilities are allowed to correlate, we are in the context of the so called multiunidimensional models (Sheng and Wikle, 2007), represented in Figure 1 (a). Letters (b) and (c) in Figure 1 show the latent structure assumed for MIRT hierarchical models, where a general ability is introduced in the framework: items in the same subtest measure a specific ability and, in turn, each specific ability is influenced by a general ability or viceversa, constituting an actual hierarchy. These kind of models have been widely applied within the economic and social fields, where typical examples are customer satisfaction and attitude studies. Different hierarchical models can be specified depending on the relation between specific and overall abilities: if each specific ability is a linear function of the overall ability we are in the case illustrated in (b), while if each specific ability linearly combines to form the overall ability we are in the case showed in (c) (Schmid and Leiman, 1957; Sheng and Wikle, 2008). Finally, the additive model is presented in Figure 1 (d). In this model, the response to a test item is directly affected by both the general and the specific latent traits, that are allowed to correlate, resulting in a compensatory, additive structure (Sheng and Wikle, 2009). An additive structure is quite common in psychological and educational contexts,

where the existence of a dominant trait and several specific traits can be assumed at the same time (e.g. general and specific abilities or cognitive factors). The additive model has an high capability of fitting the data well, thanks to its general structure. In the following, the additive models for binary and polytomous ordinal data are presented.

2.1 Additive Model for Binary Data

Given a sample of n subjects responding to a test consisting of p items, we assume that the test is divided into m subtests each containing p_v binary items (e.g. correct-incorrect response). The additive model for binary data describes the probability of a correct response for the individual i to item j belonging to the subtest v ($Y_{vij} = 1$), with $v = 1, \dots, m$, $i = 1, \dots, n$ and $j = 1, \dots, p$, as follows

$$P(Y_{vij} = 1 | \theta_{0i}, \theta_{vi}, \alpha_{0vj}, \alpha_{vj}, \delta_{vj}) = \Phi(\alpha_{0vj}\theta_{0i} + \alpha_{vj}\theta_{vi} - \delta_{vj}), \quad (1)$$

where θ_{0i} and θ_{vi} are the overall and the m specific abilities for individual i , respectively, α_{0vj} , α_{vj} , and δ_{vj} are the item parameters for item j , and Φ is the standard normal cumulative distribution function.

In model (1) it is assumed that the general ability directly affects the candidates responses, and that this effect is summed to the effect of specific factors in order to determine the probability of success to a given test item. For each item j of the subtest v , the additive model involves the estimation of a general and a specific discrimination parameter α_{0vj} and α_{vj} , respectively, and a threshold (or difficulty) parameter δ_{vj} . It is also possible to score subjects on both general and specific dimensions and the abilities are allowed to correlate.

MCMC estimation of the additive model for binary data with correlated abilities was proposed by Sheng and Wikle (2009) and implemented in the software MATLAB by Sheng (2010).

By adopting a fully Bayesian approach, the first step is the specification of the prior distributions. Given the vector of the abilities $\boldsymbol{\theta}_i = (\theta_{0i}, \theta_{1i}, \dots, \theta_{mi})'$ for the i -th individual, a multivariate normal prior distribution $\boldsymbol{\theta}_i \sim N_{m+1}(\mathbf{0}, \mathbf{P})$ is assumed, where $\mathbf{0}$ is a vector of length $m+1$ of zeros and $\mathbf{P}$ is the constrained covariance matrix with diagonal elements being

1 and off-diagonals elements being the ability correlations. The joint posterior distribution of interest becomes

$$P(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\Sigma} | \mathbf{Y}) \propto f(\mathbf{Y} | \mathbf{Z}) P(\mathbf{Z} | \boldsymbol{\theta}, \boldsymbol{\xi}) P(\boldsymbol{\xi}) P(\boldsymbol{\theta} | \mathbf{P}) P(\boldsymbol{\Sigma}), \quad (2)$$

where $\boldsymbol{\Sigma}$ is the unconstrained variance-covariance matrix of abilities.

The posterior distribution (2) has an intractable form so the resort to simulation through the Gibbs sampler is needed. The algorithm simply works by sampling iteratively until convergence from the tractable conditional distribution of each variable. Details on the parameter estimation procedure for the additive model for binary data can be found in Sheng and Wikle (2009) and Sheng (2010).

2.2 Additive Model for Ordinal Data

The additive model presented in the previous section is referred to dichotomous responses. Nevertheless items with multiple response options exist and their use is quite common in educational assessment and psychological field. The interest in adopting MIRT models for ordinal data raises to face the widespread use of Likert items (Likert, 1932), and in general other ordered scales, on questionnaires in sociological and psychological measurement.

A multidimensional generalization of the IRT graded response model can be obtained from its unidimensional counterpart, first introduced by Samejima (1969). Let consider: (i) n individuals; (ii) a set of p ordinal items where the response Y_{ij} of the i -th subject to the j -th item can take values in the set $\{1, \dots, K_j\}$, where the lowest score on item j is 1 and the highest score is K_j . Each item thus has $K_j - 1$ thresholds $\kappa_{j1}, \dots, \kappa_{j, K_j-1}$ that have to satisfy the order constraint $\kappa_{j1} < \dots < \kappa_{j, K_j-1}$; and (iii) the existence of multiple, say m , latent abilities $\boldsymbol{\theta}_i = (\theta_{1i}, \dots, \theta_{mi})'$ underlying the responses to the items.

It is assumed that an individual can reach a specific category level of an ordinal test item only if he/she is also able to reach all the lower categories on the same item, i.e. the item necessitates an amount of steps and the accomplishment of a step requires the achievement of the previous one. This type of model is then appropriate for rating scales where a rating category includes all previous categories (Reckase, 2009).

We use a procedure adapted from the Samejima's approach. Our model is specified on the basis of the probability that an item response will fall in 'category k or lower': the probability π_{ijk} that the i -th subject will select the k -th category on item j is equal to the probability of answering below the upper boundary for the category (κ_k) minus the probability of answering below the category's lower boundary (κ_{k-1}). Generalizing, probabilities π_{ijk} are built on the basis of the cumulative probabilities $P_{ijk} = P(Y_{ij} \leq k|\theta_i)$, for $k = 2, \dots, K_j - 1$:

$$\pi_{ijk} = P_{ijk} - P_{i,j,k-1} = P(Y_{ij} \leq k|\theta_i) - P(Y_{ij} \leq k-1|\theta_i), \quad (3)$$

where $\pi_{ij1} = P_{ij1} = P(Y_{ij} \leq 1|\theta_i)$ and $\pi_{ijK_j} = 1 - P_{i,j,K_j-1} = 1 - P(Y_{ij} \leq K_j - 1|\theta_i)$, to guarantee that the probability of each category can be determined from (3).

The confirmatory model for graded responses we propose is characterized by the multi-dimensional additive latent structure, that leads to the existence of an added overall ability, denoted by θ_0 . Thus the vector of latent traits is represented by $\theta_i = (\theta_{0i}, \theta_{1i}, \dots, \theta_{mi})$. Assuming that the test is divided into m subtests each containing p_v ordinal items, the cumulative probability that the individual i will select the k -th category or lower on item j belonging to the v -th subtest ($v = 1, \dots, m$) can be expressed as:

$$\begin{aligned} P_{vij} &= P(Y_{vij} \leq k|\theta_{0i}, \theta_{vi}, \alpha_{0vj}, \alpha_{vj}, \kappa_{jk}) = \\ &= \Phi(\kappa_{jk} - \alpha_{0vj}\theta_{0i} - \alpha_{vj}\theta_{vi}) = \int_{-\infty}^{\kappa_{jk} - \alpha_{0vj}\theta_{0i} - \alpha_{vj}\theta_{vi}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad (4) \end{aligned}$$

where θ_{0i} represents the i -th overall ability and θ_{vi} represent the specific ability. For each item j of the subtest v : α_{0vj} reflects the item discrimination with reference to the overall ability, α_{vj} reflects the item discrimination with reference to the specific ability and κ_{jk} is an item parameter representing the threshold between categories k and $k+1$. The predictor becomes $\kappa_{jk} - \alpha_{0vj}\theta_{0i} - \alpha_{vj}\theta_{vi}$, and the probability π_{vij} that the i -th examinee will select the k -th category on item j in subtest v is obtained recursively from the cumulative probabilities:

$$\begin{aligned} \pi_{vij1} &= P_{vij1} \\ \pi_{vij} &= P_{vij} - P_{v,i,j,k-1} \quad \text{for } k = 2, \dots, K_j - 1 \\ \pi_{vijK_j} &= 1 - P_{v,i,j,K_j-1}. \end{aligned}$$

It has to be noticed that both general and specific abilities are involved in determining the response probability by following a compensatory approach.

The estimation of model (2) involves item parameters, person parameters and correlations between the latent traits. In a fully Bayesian approach, the application of the Gibbs sampler (Geman and Geman, 1984) can be successfully carried out through the well known open-source software OpenBUGS (<http://openbugs.net>). Briefly, this software permits a flexible and easy implementation of the algorithm (Lunn et al., 2000). The process requires the definition of four steps in the model building phase (Ntzoufras, 2011): identify the main variable of interest and the corresponding (observed) data, find a distribution that adequately describes the observed data and formulate the likelihood of the model, build a structure for the distribution parameters, and specify the prior distributions. The manifest variables are the responses given by a group of subjects to a test consisting of graded responses items, and a categorical (or generalized Bernoulli) distribution of parameters $\pi_{vij1}, \dots, \pi_{vijK_j}$ is assumed for responses, thus for $v = 1, \dots, m$, $j = 1, \dots, p$ and $i = 1, \dots, n$, it holds that:

$$P(Y_{vij} = k|\bullet) = \pi_{vij1}^{[k=1]} \cdot \pi_{vij2}^{[k=2]} \cdot \dots \cdot \pi_{vijK_j}^{[k=K_j]},$$

where the $\bullet$ represents all the parameters.

The model specification is developed according to expression (2), hence, the unknown parameters for those we need to specify a prior distribution are the person parameters $\theta_i = (\theta_{0i}, \theta_{1i}, \dots, \theta_{mi})$ and the item parameters α_{0vj} , α_{vj} and $\kappa_{j1}, \dots, \kappa_{j,K_j-1}$.

In the additive graded response model, we assume that the latent traits $\theta_1, \dots, \theta_n$ are independent among units and multivariate normally distributed. For $i = 1, \dots, n$, the prior distributions are defined as:

$$\theta_i \sim N_{m+1}(\mu, \Sigma), \quad (5)$$

where μ is the $(m+1)$ -dimensional ability mean vector and Σ is the $(m+1) \times (m+1)$ fixed correlation matrix with diagonal elements being 1 and off-diagonal elements being the ability correlations.

Moreover, considering that the parameters which reflect the power of an item to discriminate between examinees are evidently positive, truncated normal distributions are assumed

for item discrimination parameters α_{0vj} and α_{vj} , for $v = 1, \dots, m$ and $j = 1, \dots, p$:

$$\alpha_{0vj} \sim N(\mu_{\alpha_0}, \sigma_{\alpha_0}^2) I(\alpha_{0vj} > 0) \quad \alpha_{vj} \sim N(\mu_{\alpha}, \sigma_{\alpha}^2) I(\alpha_{vj} > 0), \quad (6)$$

where $I(\bullet)$ is the indicator function which, in this case, ensures positive discrimination parameters.

The prior distributions for the threshold parameters must take into account the order constraint $\kappa_{j1} < \dots < \kappa_{j,K_j-1}$. Curtis (2010) and Plummer (2010) suggest to introduce priors on unconstrained auxiliary variables $\kappa_{j1}^*, \dots, \kappa_{j,K_j-1}^*$, such that:

$$\kappa_{jk}^* \sim N(\mu_{\kappa}, \sigma_{\kappa}^2), \quad (7)$$

for $j = 1, \dots, p$ and $k = 1, \dots, K_j - 1$. Then, thresholds for the j -th item can be obtained considering the order statistics for the auxiliary variables, that is:

$$\begin{aligned} \kappa_{j1} &= \kappa_{j,[1]}^* \\ \kappa_{j2} &= \kappa_{j,[2]}^* \\ &\vdots \\ \kappa_{j,K_j-1} &= \kappa_{j,[K_j-1]}^*, \end{aligned}$$

where with $\kappa_{j,[s]}^*$ is denoted the s -th order statistic of $\kappa_{j1}^*, \dots, \kappa_{j,K_j-1}^*$.

Particular attention should be paid to the restrictions that have to be imposed on hyperparameters in order to ensure the model identification. In general, Bayesian item response models can be identified by imposing restrictions on the hyperparameters or via a scale transformation in the estimation procedure (Fox, 2010).

According to the first approach, for identification purposes we set $\boldsymbol{\mu} = \mathbf{0}$, $\mu_{\alpha_0} = \mu_{\alpha} = \mu_{\kappa} = 0$, and $\sigma_{\alpha_0}^2 = \sigma_{\alpha}^2 = \sigma_{\kappa}^2 = 1$. Moreover, a multivariate normal prior distribution with a fixed correlation structure is assumed for abilities: $\boldsymbol{\theta}_i \sim N_m(\mathbf{0}, \boldsymbol{\Sigma})$, for $i = 1, \dots, n$, where $\boldsymbol{\Sigma}$ is the fixed correlation matrix defined before. Considering these identification constraints, expressions (5), (6) and (7) are replaced by (8), (9) and (10):

$$\boldsymbol{\theta}_i \sim N_{m+1}(\mathbf{0}, \boldsymbol{\Sigma}) \quad (8)$$

$$\alpha_{0vj} \sim N(0, 1) I(\alpha_{0vj} > 0) \quad \alpha_{vj} \sim N(0, 1) I(\alpha_{vj} > 0) \quad (9)$$

$$\kappa_{jk}^* \sim N(0, 1) . \quad (10)$$

Even if this choice can be viewed as very restrictive, it reflects the common beliefs and usual assumptions we find in literature. In fact, a point of strength of the Bayesian approach is the possibility to formulate particular prior distributions depending on the information available a priori.

Once the the model is fully specified, we can perform the Bayesian estimation of the parameters of interest through an easy implementation in OpenBUGS, which automatically runs the Gibbs sampler algorithm. In particular, the main advantage is due to the fact that we do not have to deal directly with the joint posterior distribution, which presents an untractable form: $P(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\kappa}, \boldsymbol{\Sigma} | \mathbf{Y}) \propto f(\mathbf{Y} | \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\kappa}) P(\boldsymbol{\theta} | \boldsymbol{\Sigma}) P(\boldsymbol{\alpha}) P(\boldsymbol{\kappa})$, where f is the likelihood function and $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are assumed to be independent.

3. SIMULATION STUDY

Several works on MIRT models focus on the accuracy of parameter estimation under different simulation conditions (Sheng, 2008; Sheng, 2010; Edwards, 2010; Matteucci, 2014). In this work, a simulation study has been designed with reference to the additive graded response model.

3.1 Simulation study design

In the present study, the assessment of parameter recovery focuses on the case of two specific latent traits, i.e. $m = 2$ and $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$, by varying the number of response categories, the sample size, the test and subtest lengths and the ability correlation structure.

The scheme used for each simulation is the following:

- simulate the vectors of ‘real’ parameters, taking into account the conditions we are testing. We perform this step using an R procedure;
- perform $Q = 10$ replications of the computation procedure. In each replication we

sample the data matrix using the parameters obtained at the previous step, and we run OpenBUGS through the R package BRugs (Thomas et al., 2006), which basically permits to recall OpenBUGS automatically from R;

- proceed to the evaluation of parameter recovery and the computation of the reproduced correlations between the latent traits by using the Q estimates gained at the previous step.

In order to evaluate the recovery of the generated item parameters, we compute the bias and the root mean square error (RMSE) for each estimated parameter, taking into account the Q replications for each simulation. If we denote with $\hat{\omega}$ a generic parameter estimate, i.e. the mean of the posterior distribution gained in each replication, and with ω^* the real generated value, biases and RMSE are computed as follow:

$$\text{Bias}(\omega) = \frac{1}{Q} \sum_{q=1}^Q (\hat{\omega}_q - \omega^*)$$

$$\text{RMSE}(\omega) = \frac{1}{Q} \sqrt{\sum_{q=1}^Q (\hat{\omega}_q - \omega^*)^2}.$$

Lower levels of bias and RMSE indicate better precision in parameter recovery.

Moreover, considering that in our additive model the latent traits are allowed to correlate, leading to different variance/covariance matrices Σ , we are interested not only in item parameters recovery, but also in the way the models are able to reproduce such ability correlations. For each simulation, we thus report also the estimated ability Pearson correlations $\hat{r}_{01}$, $\hat{r}_{02}$ and $\hat{r}_{12}$, computed from the values $\hat{\theta}_1, \dots, \hat{\theta}_n$.

3.2 Simulation conditions

In every replication performed, we considered a chain length of 30,000 iterations, with a burn-in phase of 15,000 iterations. Moreover, two chains were generated for each unknown model parameter to check the chains convergence through the R diagnostic statistics of

Gelman and Rubin (Gelman and Rubin, 1992; Brooks and Gelman, 1998). The chosen chain length and burn-in period widely ensure the convergence of all the generated chains.

We consider n individuals and a test consisting of p ordinal items, divided into 2 subtests, each one containing p_1 and p_2 items. According to the bidimensional additive latent trait structure, we assume the existence of an overall ability, which influences the responses to all the test items p , and 2 specific abilities, that affect, respectively, the responses to the p_1 items (Subtest 1), and the responses to the p_2 items (Subtest 2).

We start from a first block of simulations referred to the case where a test length of $p = 15$ is divided into a first subtest made of $p_1 = 5$ items and a second subtest made of $p_2 = 10$ items. A further distinction has been made about the number of item categories, varying in the first block from $K = 3$ to $K = 4$. Furthermore, each case was analyzed by using two different correlation matrices among the abilities: Σ_A and Σ_B . Σ_A is a 3×3 identity matrix, where all the correlations among the abilities are set to zero ($r_{01} = r_{02} = r_{12} = 0$). In this case, the additive model with orthogonal traits has the same latent structure of the well known bi-factor model and the three latent traits (the general and specific abilities) are separate and well distinguished from each other. The second correlation matrix Σ_B introduce moderate correlations between all the latent abilities ($r_{01} = 0.4$, $r_{02} = 0.3$, $r_{12} = 0.2$). The choice to consider not particularly high levels of correlation has been driven by the consideration that high correlations among the latent abilities may lead to the existence of a dominant latent trait, redirecting to a unidimensional model. Moreover, in addition to the two correlation matrices taken into account and described in this simulation study, other correlation matrices for abilities have been considered in order to perform a sensitivity analysis, leading to similar results.

In order to investigate further conditions, we designed a second block of simulations, where we increase both the length of the test and the number of item categories. We consider a case characterized by a test length of $p = 50$ (divided into $p_1 = 20$ and $p_2 = 30$ items for subtest 1 and 2, respectively) and $K = 4$ categories for each test item; and a last case where the test length is $p = 30$ ($p_1 = 10$ and $p_2 = 20$) and items have $K = 5$ categories.

Again, with respect to the correlation matrix, the two cases of Σ_A and Σ_B are distinguished as above.

A final varying condition is the sample size n : for each case considered we perform two simulations, setting the sample size at $n = 500$ and $n = 1000$. Considering the simulation conditions, we refer to 16 different scenarios to investigate the parameter recovery for the proposed model. The simulation conditions are reported in Table 1.

[TABLE 1]

As the number of parameters is quite large, for each scenario, we summarize the results on the item parameters within each subtest by using median absolute bias and median RMSE.

3.3 Results

The average computation time for each replication is about 16 hours and the convergence of each Markov chain for every parameter is reached. The diagnosis of convergence has been performed through the calculation of the Gelman and Rubin statistic, available in OpenBUGS.

Tables 2 and 3 show the item parameter recovery for the first block of simulations where $p = 15$ ($p_1 = 5$ and $p_2 = 10$), respectively for subset 1 and subtest 2. It emerges that all parameters are quite well recovered when the number of categories for each item is $K = 3$ and a sample size of $n = 500$ is enough to get accurate estimates. Results are slightly better for the Σ_A correlation matrix, rather than Σ_B .

On the other hand, when the number of item categories is $K = 4$ we obtain less accurate estimates, for both Σ_A and Σ_B ability correlation structures. Estimates get better after increasing the sample size, but median RMSEs and biases remain rather high, especially for α_0 and α_v discrimination parameters. Considering that this result is more evident for the first subtest where $p_1 = 5$, rather than the second one where $p_2 = 10$, this may be due to the small number of items compared to the increased number of categories.

[TABLE 2]

[TABLE 3]

Results about the second block of simulations are reported in Tables 4 and 5. Focusing on the case where $p = 50$ ($p_1 = 20$ and $p_2 = 30$) and $K = 4$, we observe that in both subtests the item parameters are not well recovered, particularly the discrimination parameters. Nevertheless, these shortcomings are overtaken by increasing the sample size. In fact, when $n = 1000$ all the parameters are recovered very precisely. Different correlation structures seem not to affect parameter recovery, with an exception of the discrimination parameters for the second subtest, where we register higher median RMSEs in association to the more complex correlation structure.

Analogously, the cases where $p = 30$ ($p_1 = 10$ and $p_2 = 20$) and $K = 5$ benefit from the enlarged sample size. For $n = 1000$, item parameters are recovered with extreme care, with slightly better accuracy with respect to Σ_A correlation matrix.

[TABLE 4]

[TABLE 5]

Table 6 illustrates the estimated ability correlations for each scenario. Their correspondent true values are also reported and we can observe how the correlations are reproduced. In particular, the results are coherent with the ones observed in relation to the item parameters: the best performance is associated to the cases of the highest sample size, a reasonable number of items (totally 50) and a number of categories equal to 4, even in case of slightly high correlations.

[TABLE 6]

5. ILLUSTRATIVE EXAMPLE

As illustrative example with real data, we implemented our proposed model on data collected with the aim to investigate the residents' perception and attitudes towards the tourism industry in the Romagna area and in the State of San Marino (Italy). Data analyzed

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

are a subset of results of a research conducted by the University of Bologna with the aim to study the subjective well-being (Bernini et al., 2013). A total of 794 questionnaires were collected. In particular, we focus on the perception of benefits and costs associated to the tourism industry. The perceived benefits of tourism were assessed by five items: the support in local economic development [B1], quality of life [B2], public services improvement [B3], employment prospects [B4], and opportunities for cultural activities [B5]. Respondents were asked to indicate whether those items would improve for their community as a result of increasing tourism activity on a 7-point anchor scale, from ‘strongly disagree’ to ‘strongly agree’. On the other hand, the perceived costs of tourism were assessed by other five items: the cost of living [C1], crime [C2], environment damage [C3], traffic congestion [C4], and pollution [C5]. In this case residents were asked to express if those aspects would worsen for their community as a result of increasing tourism activity on the 7-point scale mentioned above. Scales of the items with respect to costs were inverted in order to eliminate reverse scoring and make the low and high scores be associated with high and low perceptions of costs, respectively (see Table 7 for response frequencies for items on benefits and costs).

[TABLE 7]

In this context, latent traits can be defined as ‘perceptions’. The investigation involves two distinct aspects of the phenomenon. Therefore, it is possible to identify two specific perceptions (θ_1 for benefits and θ_2 for costs) and the overall attitude towards tourism (θ_0) of respondents as latent variables. Within this framework, discrimination parameters represent the capability of the items to differentiate between respondents with different levels of agreement, whereas the threshold parameters can be interpreted as ‘criticity levels’ of the corresponding item. For a given item, high values for the criticity parameters correspond to lower probabilities to observe responses in positive categories.

4.1 Main results

We estimated the additive graded response model by using the proposed OpenBUGS estimation procedure (Table 8).

[TABLE 8]

Results show that items B3, B4 and B5 are characterized by high specific discrimination parameters, hence they represent the most informative items on the perception of the tourism advantages. On the other hand, items C2, C4 and C5 are the most informative items on the perception of the tourism costs. Concerning the general latent trait, items B3, B5, C4 and C5 principally influence the general residents' attitude towards tourism (high general discrimination parameters).

With reference to the items' criticism levels, within the group of items on benefits, items B1 and B4 are associated with higher probabilities of responses in higher categories, because the corresponding estimates for the criticism levels are generally lower than for the remaining items. This means that residents who have an average attitude toward tourism ($\hat{\theta}_0 = 0$) and an average specific perception of advantages ($\hat{\theta}_1 = 0$) consider the economic development and the job opportunities as the main advantages of tourism. Among the items on costs, item C1 is characterized by generally lower thresholds' parameters in comparison to the estimated criticism levels of other items, especially with reference to higher categories. Hence, for a typical respondent characterized now by $\hat{\theta}_0 = 0$ and $\hat{\theta}_2 = 0$, the cost of life can be regarded as a marginal negative aspect of tourism in comparison with the other issues.

Finally, the estimated correlations between the latent traits are $\hat{r}_{01} = 0.03$, $\hat{r}_{02} = 0.18$ and $\hat{r}_{12} = -0.62$. It has to be noticed that the estimated correlation between the two specific latent traits is negative and relatively high, indicating that the perception of a high economic advantage of tourism is associated with a strongly negative environmental impact.

5. CONCLUSIONS

In this paper, we proposed a MIRT model for ordinal data where both a general and specific latent traits are assumed to explain the item responses, and correlation among the traits is allowed. In particular, we proposed a MCMC procedure to estimate the item parameters, the ability scores, and the trait correlations jointly. Despite the computational heaviness of the estimation procedure, MCMC is very effective in dealing with complex data

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

in comparison to traditional maximum likelihood estimation techniques. The parameter recovery for the model was investigated through a simulation study where different conditions are tested. Within an approach with two specific and one general factors, we considered different sample sizes ($n = 500$ and $n = 1000$), test lengths ($p = 15$, $p = 30$, $p = 50$) with different combinations for the subtest lengths, number of response categories ($K = 3$, $K = 4$, $K = 5$), and correlation matrices (identity matrix or low-moderate correlations among the traits). Overall 16 distinct scenarios were taken into account. The main results showed that the algorithm is particularly sensitive to the sample size due to the model complexity and the high number of parameters to be estimated. In fact, when the sample size is sufficiently large ($n = 1000$), all the parameters are well reproduced. The results are also affected by the trade-off between the test length and the number of categories: the worst results are associated to an high number of categories and a low test length. Analogous evidences apply for the correlation estimates.

Because the MIRT additive graded response model is very complex, it is very sensible to several issues conditioning the estimation, such as sample size, test and subtest length, number of response categories, and correlations among the traits. From an applicative point of view, the additive model is very attractive to analyse data coming from the behavioral and social sciences, where the manifest variables are often in a ordinal scale. Usually, models for binary data are used because the estimation procedure is simple. However, the dichotomization of the variables causes a considerable loss of information and the resort to graded response models should be preferred. Moreover, within social phenomena, the presence of an hierarchy among the constructs is very common and the analysis can be considerably improved by applying an additive graded model.

In order to verify the performance of the additive graded model in an empirical setting, we reported an illustrative example in which we investigate the residents' perception and attitudes towards the tourism industry.

With this article we would like to encourage the diffuse use of MIRT models among social researchers and practitioners. In particular, the MCMC estimation method in a fully

Bayesian approach offers many advantages, which include, from a practical point of view, the relative ease of implementation and the availability of free software.

Concerning the estimation procedure, a crucial issue is represented by the use of a fixed correlation matrix as a prior distribution for abilities. Considering that the additive graded model is highly complex, notwithstanding this choice may be very restrictive, it reasonably works. The introduction of a prior distribution for the variance/covariance matrix characterizing the prior distribution on abilities, e.g. an inverse Wishart distribution, represents an appropriate future development.

Future developments could include the extension of the simulation study to test the sensitivity to different prior distributions and to a larger number of specific dimensions. Finally, further application fields could be examined, for example in educational and psychological assessments.

FUNDING

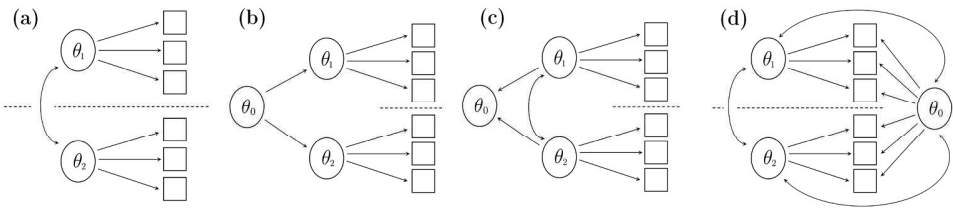
This research has been partially funded by the Italian Ministry of Education with the FIRB (“Futuro in ricerca”) 2012 project on “Mixture and latent variable models for causal-inference and analysis of socio-economic data”.

References

- Albert, J. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17(3):251–269.
- Bernini, C., Guizzardi, A., and Angelini, G. (2013). DEA-like model and common weights approach for the construction of a subjective community well-being indicator. *Social indicators research*, 114(2):405–424.
- Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

- Chen, F. F., West, S. G., and Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41(2):189–225.
- Curtis, S. (2010). BUGS code for item response theory. *Journal of Statistical Software*, 36(1):1–34.
- Edwards, M. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3):474–497.
- Fox, J.-P. (2010). *Bayesian item response modeling*. Springer.
- Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.
- Hambleton, R. and Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Kluwer Nijhoff Publishing, Boston.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*, 22(140).
- Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.
- Matteucci, M. (2014). An investigation of parameter recovery in MCMC estimation for the Additive IRT model. *Communications in Statistics - Theory and Methods*, 43(4):751–770.
- Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, volume 698. Wiley. com.
- Plummer, M. (2010). JAGS Version 2.2.0 user manual.

- Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer, London.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17*.
- Schmid, J. and Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1):53–61.
- Sheng, Y. (2008). A MATLAB package for Markov chain Monte Carlo with a multi-unidimensional IRT model. *Journal of Statistical Software*, 28(10):1–20.
- Sheng, Y. (2010). Bayesian estimation of MIRT models with general and specific latent traits in MATLAB. *Journal of Statistical Software*, 34(3):1–26.
- Sheng, Y. and Wikle, C. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 67(6):899–919.
- Sheng, Y. and Wikle, C. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, 68(3):413–430.
- Sheng, Y. and Wikle, C. (2009). Bayesian IRT models incorporating general and specific abilities. *Behaviormetrika*, 36(1):27–48.
- Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006). Making BUGS Open. *R News*, 6(1):12–17.
- van der Linden, W. and Hambleton, R. (1997). *Handbook of Modern Item Response Theory*. Springer-Verlag, New York.



Multidimensional underlying latent structures.
190x42mm (300 x 300 DPI)

Additive model: simulation conditions						
Simulation	p	p_1	p_2	K_j	n	Σ
# 1	15	5	10	3	500	Σ_A
# 2	15	5	10	3	500	Σ_B
# 3	15	5	10	4	500	Σ_A
# 4	15	5	10	4	500	Σ_B
# 5	15	5	10	3	1000	Σ_A
# 6	15	5	10	3	1000	Σ_B
# 7	15	5	10	4	1000	Σ_A
# 8	15	5	10	4	1000	Σ_B
# 9	50	20	30	4	500	Σ_A
# 10	50	20	30	4	500	Σ_B
# 11	30	10	20	5	500	Σ_A
# 12	30	10	20	5	500	Σ_B
# 13	50	20	30	4	1000	Σ_A
# 14	50	20	30	4	1000	Σ_B
# 15	30	10	20	5	1000	Σ_A
# 16	30	10	20	5	1000	Σ_B

Table 1: Varying conditions for the simulation study.

Simulations block 1 - Subtest 1														
			α_0		α_1		κ_1		κ_2		κ_3		κ_4	
n	(p_1, p_2)	K	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
500	(5,10)	3 Σ_A	0.08	0.05	0.08	0.03	0.08	0.01	0.07	0.04				
		Σ_B	0.09	0.02	0.15	0.13	0.09	0.03	0.09	0.02				
	(5,10)	4 Σ_A	0.13	0.07	0.12	0.09	0.15	0.12	0.15	0.10	0.13	0.04		
		Σ_B	0.17	0.05	0.23	0.10	0.16	0.16	0.09	0.02	0.15	0.03		
1000	(5,10)	3 Σ_A	0.07	0.03	0.09	0.03	0.08	0.06	0.07	0.03				
		Σ_B	0.09	0.03	0.14	0.07	0.08	0.03	0.06	0.03				
	(5,10)	4 Σ_A	0.09	0.02	0.16	0.12	0.15	0.05	0.06	0.03	0.16	0.10		
		Σ_B	0.14	0.08	0.16	0.12	0.08	0.05	0.08	0.04	0.15	0.10		

Table 2: Additive model: block 1 simulation results for subtest 1 (median RMSEs and median absolute biases).

Simulations block 1 - Subtest 2

n	(p_1, p_2)	K	α_0		α_2		κ_1		κ_2		κ_3		κ_4	
			RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
500	(5,10)	3 Σ_A	0.09	0.05	0.10	0.02	0.08	0.02	0.08	0.03				
		Σ_B	0.11	0.04	0.00	0.05	0.09	0.04	0.10	0.00				
	(5,10)	4 Σ_A	0.12	0.02	0.14	0.05	0.10	0.03	0.10	0.04	0.13	0.06		
		Σ_B	0.14	0.04	0.14	0.04	0.10	0.10	0.09	0.02	0.11	0.03		
1000	(5,10)	3 Σ_A	0.09	0.04	0.09	0.05	0.06	0.02	0.06	0.02				
		Σ_B	0.15	0.05	0.18	0.03	0.05	0.02	0.05	0.01				
	(5,10)	4 Σ_A	0.16	0.12	0.16	0.09	0.07	0.03	0.07	0.03	0.08	0.03		
		Σ_B	0.23	0.13	0.19	0.11	0.07	0.04	0.06	0.03	0.09	0.04		

Table 3: Additive model: block 1 simulation results for subtest 2 (median RMSEs and median absolute biases).

Simulations block 2 - Subtest 1														
			α_0		α_1		κ_1		κ_2		κ_3		κ_4	
n	(p_1, p_2)	K	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
500	(20,30)	4 Σ_A	0.14	0.07	0.14	0.08	0.10	0.06	0.10	0.05	0.10	0.04		
		Σ_B	0.15	0.07	0.18	0.10	0.09	0.03	0.08	0.03	0.09	0.03		
	(10,20)	5 Σ_A	0.20	0.05	0.21	0.06	0.09	0.01	0.09	0.03	0.08	0.04	0.09	0.04
		Σ_B	0.17	0.07	0.22	0.07	0.10	0.02	0.09	0.02	0.08	0.02	0.08	0.02
1000	(20,30)	4 Σ_A	0.07	0.05	0.08	0.04	0.06	0.01	0.06	0.01	0.05	0.01		
		Σ_B	0.06	0.02	0.08	0.04	0.06	0.04	0.06	0.04	0.06	0.03		
	(10,20)	5 Σ_A	0.07	0.04	0.07	0.04	0.08	0.02	0.06	0.02	0.05	0.02	0.05	0.01
		Σ_B	0.19	0.04	0.27	0.05	0.07	0.03	0.06	0.03	0.05	0.03	0.07	0.02

Table 4: Additive model: block 2 simulation results for subtest 1 (median RMSEs and median absolute biases).

Simulations block 2 - Subtest 2															
n	(p_1, p_2)	K	α_0		α_2		κ_1		κ_2		κ_3		κ_4		
			RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	
500	(20,30)	4	Σ_A	0.16	0.07	0.20	0.08	0.10	0.03	0.09	0.03	0.10	0.03		
		Σ_B	0.07	0.02	0.10	0.05	0.10	0.06	0.10	0.06	0.10	0.05			
	(10,20)	5	Σ_A	0.23	0.08	0.18	0.05	0.09	0.03	0.07	0.03	0.07	0.02	0.09	0.03
		Σ_B	0.20	0.09	0.19	0.08	0.08	0.03	0.07	0.03	0.08	0.02	0.08	0.02	
1000	(20,30)	4	Σ_A	0.06	0.05	0.06	0.02	0.06	0.02	0.06	0.02	0.06	0.01		
		Σ_B	0.14	0.06	0.17	0.03	0.07	0.02	0.05	0.02	0.07	0.02			
	(10,20)	5	Σ_A	0.07	0.05	0.07	0.03	0.05	0.01	0.05	0.01	0.05	0.01	0.06	0.02
		Σ_B	0.06	0.01	0.08	0.03	0.06	0.02	0.05	0.02	0.05	0.01	0.06	0.02	

Table 5: Additive model: block 2 simulation results for subtest 2 (median RMSEs and median absolute biases).

Real and estimated ability correlations								
			r_{01}	$\hat{r}_{01}$	r_{02}	$\hat{r}_{02}$	r_{12}	$\hat{r}_{12}$
500	(5,10) 3	Σ_A	0.00	0.07	0.00	0.16	0.00	-0.07
		Σ_B	0.40	0.62	0.30	0.49	0.20	0.24
	(5,10) 4	Σ_A	0.00	0.09	0.00	0.29	0.00	-0.07
		Σ_B	0.40	0.60	0.30	0.56	0.20	0.27
1000	(5,10) 3	Σ_A	0.00	0.11	0.00	0.16	0.00	-0.05
		Σ_B	0.40	0.60	0.30	0.54	0.20	0.27
	(5,10) 4	Σ_A	0.00	0.13	0.00	0.36	0.00	-0.05
		Σ_B	0.40	0.58	0.30	0.65	0.20	0.30
500	(20,30) 4	Σ_A	0.00	0.00	0.00	0.29	0.00	-0.05
		Σ_B	0.40	0.50	0.30	0.36	0.20	0.21
	(10,20) 5	Σ_A	0.00	0.02	0.00	0.15	0.00	-0.03
		Σ_B	0.40	0.51	0.30	0.48	0.20	0.24
1000	(20,30) 4	Σ_A	0.00	0.03	0.00	0.07	0.00	-0.02
		Σ_B	0.40	0.45	0.30	0.37	0.20	0.20
	(10,20) 5	Σ_A	0.00	0.06	0.00	0.11	0.00	-0.05
		Σ_B	0.40	0.52	0.30	0.41	0.20	0.24

Table 6: Real (r) and estimated ($\hat{r}$) ability correlations.

		Responses						
		Low benefits ←			→ High benefits			
Item	Item description	1	2	3	4	5	6	7
B1	Econ. support	12	51	58	157	149	235	132
B2	Quality of life	24	49	78	184	227	155	77
B3	Public services	16	45	97	186	190	171	89
B4	Job opportunities	16	36	69	157	187	198	131
B5	Cultural act.	30	54	76	186	188	157	103

		Responses						
		High costs ←			→ Low costs			
Item	Item description	1	2	3	4	5	6	7
C1	Cost of life	64	151	182	139	119	100	39
C2	Crime rate	145	169	157	155	69	71	28
C3	Env. damage	117	151	166	187	96	59	18
C4	Traffic	193	152	158	130	89	45	27
C5	Pollution	158	173	164	136	63	81	19

Table 7: Response frequencies for items about tourism benefits (B1-B5) and items about tourism costs (C1-C5).

Item description		$\hat{\alpha}_v$	$SD(\hat{\alpha}_v)$	$MCSE(\hat{\alpha}_v)$	$\hat{\kappa}_1$	$SD(\hat{\kappa}_1)$	$MCSE(\hat{\kappa}_1)$	$\hat{\kappa}_2$	$SD(\hat{\kappa}_2)$	$MCSE(\hat{\kappa}_2)$	$\hat{\kappa}_3$	$SD(\hat{\kappa}_3)$	$MCSE(\hat{\kappa}_3)$
B1	Econ. support	1.047	0.074	0.001	- 3.049	0.169	0.002	- 2.014	0.105	0.001	- 1.469	0.087	0.001
B2	Quality of life	0.946	0.063	0.001	- 2.539	0.125	0.001	- 1.789	0.090	0.001	- 1.176	0.073	0.001
B3	Public services	1.247	0.082	0.001	- 3.278	0.188	0.002	- 2.200	0.119	0.001	- 1.306	0.088	0.001
B4	Job opportunity	1.290	0.082	0.001	- 3.342	0.187	0.002	- 2.390	0.125	0.001	- 1.595	0.094	0.001
B5	Cultural act.	1.194	0.077	0.001	- 2.713	0.141	0.001	- 1.901	0.103	0.001	- 1.256	0.084	0.001
C1	Cost of life	0.284	0.042	0.000	- 1.491	0.069	0.000	- 0.644	0.049	0.000	- 0.002	0.046	0.000
C2	Crime rate	1.534	0.109	0.002	- 1.824	0.123	0.002	- 0.534	0.085	0.002	0.461	0.083	0.001
C3	Env. damage	1.343	0.090	0.001	- 1.901	0.114	0.002	- 0.745	0.080	0.001	0.205	0.073	0.001
C4	Traffic	1.487	0.126	0.004	- 1.509	0.134	0.005	- 0.397	0.092	0.003	0.700	0.098	0.002
C5	Pollution	1.425	0.103	0.002	- 1.646	0.114	0.003	- 0.452	0.083	0.002	0.548	0.085	0.002
Item description		$\hat{\alpha}_0$	$SD(\hat{\alpha}_0)$	$MCSE(\hat{\alpha}_0)$	$\hat{\kappa}_4$	$SD(\hat{\kappa}_4)$	$MCSE(\hat{\kappa}_4)$	$\hat{\kappa}_5$	$SD(\hat{\kappa}_5)$	$MCSE(\hat{\kappa}_5)$	$\hat{\kappa}_6$	$SD(\hat{\kappa}_6)$	$MCSE(\hat{\kappa}_6)$
B1	Econ. support	0.013	0.012	0.000	- 0.507	0.066	0.001	0.204	0.064	0.001	1.458	0.086	0.001
B2	Quality of life	0.250	0.073	0.001	- 0.250	0.061	0.001	0.802	0.067	0.001	1.871	0.094	0.001
B3	Public services	0.446	0.095	0.002	- 0.264	0.071	0.001	0.760	0.078	0.001	2.066	0.116	0.002
B4	Job opportunity	0.144	0.083	0.002	- 0.560	0.073	0.001	0.405	0.071	0.001	1.644	0.097	0.002
B5	Cultural act.	0.343	0.094	0.002	- 0.228	0.069	0.001	0.732	0.074	0.001	1.856	0.103	0.001
C1	Cost of life	0.017	0.016	0.000	0.470	0.048	0.000	0.964	0.054	0.000	1.676	0.075	0.000
C2	Crime rate	0.074	0.060	0.001	1.470	0.106	0.002	2.068	0.128	0.002	3.083	0.179	0.003
C3	Env. damage	0.051	0.049	0.001	1.299	0.091	0.001	2.093	0.116	0.002	3.089	0.167	0.002
C4	Traffic	0.906	0.144	0.005	1.660	0.137	0.004	2.566	0.186	0.006	3.381	0.240	0.007
C5	Pollution	0.668	0.113	0.003	1.520	0.110	0.002	2.102	0.133	0.003	3.413	0.200	0.004

NOTE: $v = 1$ for the items on benefits and $v = 2$ for the items on costs, SD = standard deviation, MCSE = Monte Carlo standard error.

Table 8: Item parameter estimates for the additive graded response model.