The Hellinger Distance within Posterior Predictive Assessment for Investigating Multidimensionality in IRT Models

The Hellinger Distance within Posterior Predictive Assessment for Investigating Multidimensionality in IRT Models

Abstract

Under the Bayesian approach, posterior predictive model checking (PPMC) has become a popular tool for fit assessment of item response theory (IRT) models. In this study, we propose the use of the Hellinger distance within PPMC to quantify the distance between the realized and the predictive distribution of the model-based covariance for item pairs. Specifically, the case of multidimensional data analyzed with a unidimensional approach is taken into account. The results of the simulation study show the effectiveness of the method in detecting model misfit and the sensitivity to the trait correlations. An application to real data on tourism perceptions shows the feasibility of the method in practice and especially the capability of detecting potential misfit attributed to specific items.

Keywords: posterior predictive model checking, Hellinger distance, MIRT models, goodness of fit, MCMC.

1. Introduction

In the field of item response theory (IRT), the Bayesian approach via Markov chain Monte Carlo (MCMC) was proved to be very flexible for both estimating complex models and investigating goodness of fit through posterior predictive model checking (PPMC). Model checking is a crucial issue in statistical analysis. Bayesian model checking deals with the step of "assessing the fit of the model to the data and to our substantive knowledge" (Gelman et al., 2014; Chapter 6). Specifically, checking the fit of a model includes checking the sampling distribution, the prior distribution, the hierarchical structure and all the issues that are related

to the specification of the model, such as the presence of mixture components or not. As all models are essentially wrong, the typical question in model checking is not if the selected model is true or not. Instead, the researcher should investigate to what extent the weak aspects of the model are able to affect the inferential results. Within Bayesian model checking, PPMC is devoted to assessing the discrepancies between a model and data and understanding the model limitations in real applications (Gelman et al., 1996). Moreover, PPMC is used for assessing the fit of a single model to the data, in the absence of explicit alternative models.

The PPMC method is based on the comparison between the observed and the replicated data of a given discrepancy measure *D*. The main advantages of PPMC are that it does not rely on distributional assumptions and it is relatively easy to implement, given that the entire posterior distribution of all parameters of interest is obtained through MCMC algorithms.

In IRT, PPMC has been used initially to investigate differential item functioning (Hoijtink, 2001), person fit (Glas and Meijer, 2003), fit of unidimensional models (Sinharay, 2005; Zhu and Stone, 2011) and item fit (Sinharay, 2006). Successively, there was an increasing interest in checking specifically for the behaviour of unidimensional models fitted to potential multidimensional data (see, among others, Sinharay, Johnson, and Stern, 2006; Levy, Mislevy, and Sinharay, 2009; Levy, 2011) and in assessing the test dimensionality (Levy and Svetina, 2011; Levy, Xu, Yel, and Svetina, 2015). In real data applications, item response data often show multidimensionality due to the existence of different latent variables (e.g., cognitive abilities, perceptions). In these cases, estimating a unidimensional model yields an imprecise measure of individual traits as compared to fitting multiple unidimensional models with correlated traits (Wang, Chen, and Cheng, 2004).

The investigation of IRT model dimensionality is based on checking the local independence assumption. For this purpose, measures of association among item pairs (e.g. the odds ratio, depending on data only) or measures of correlation among item pairs (e.g. the model-based

covariance, depending on both data and model parameters) were found to be effective. Given the chosen discrepancy measure, PPMC is implemented first with graphical analyses and then with the estimation of the posterior predictive *p*-values (PPP-values). When the plots are ambiguous, i.e. they do not suggest neither good fit or possible sources of misfit, or when the practitioner is interested in more than one check on the model, it is possible to draw useful information from the PPP-value. However, the PPP-value simply counts the number of times the replicated *D* is equal or higher than the realized *D* without addressing the magnitude of the difference between the two distributions. In fact, an important limitation of the PPP-value is that it can be equal to 0.5 (far for being extreme) also when the realized and the predictive distributions are very different (see Wu, Yuen, and Leung, 2014; Section 4.3). The PPP-value approach was also found to be conservative as it could fail to reject an inadequate model.

To overcome the limitations of graphical plots and PPP-values, Wu et al. (2014) proposed the use of relative entropy (RE) within PPMC. The RE was used to measure the difference between the predictive and the realized distribution of global fit measures. However, the RE suffers from two main drawbacks that may weaken its usefulness as a pairwise distance measure in applied settings: it is asymmetric and not upper bounded (Kang and Chang, 2016).

The main novelty of our proposal is the use of the Hellinger distance to quantify the magnitude of the difference between the predictive and the realized distribution of measures for item pairs. In fact, similarly to the RE, the Hellinger distance provides a quantitative measure of the degree of misfit by overcoming the approach based on PPMC. Unlike the RE, the Hellinger distance satisfies the metric properties, including symmetry, and it is bounded between 0 and 1. These properties make the Hellinger distance a suitable measure for improving the interpretation of results in applied settings. Moreover, the Hellinger distance can be used for model comparison purposes, considering both single discrepancy measures on item pairs and overall. The performance of the Hellinger distance is investigated for detecting the

misfit of an IRT unidimensional model with both simulated and real multidimensional response data. We compare our proposal to the classical PPMC based on graphical analysis and PPP-values and the approach based on the RE. The simulation results in different scenarios allow to explore the features (strong points and weaknesses) of the proposed solution. The Hellinger distance seems to be an effective tool in highlighting the presence of possible misfit and determining plausible thresholds for classifying the misfit levels. The most important feature of the proposal based on the Hellinger distance is to be able to report and quantify misfit when data are multidimensional but a unidimensional model is used. Moreover, unlike PPP-values and RE, the Hellinger distance is effectively employed for model comparison.

The paper is organized in the following way. Section 2 briefly reviews the IRT models used in the paper. In Section 3, PPMC is discussed both in general terms and in reference to the specific issue of fit for unidimensional models when data are multidimensional. Here, the proposal of the Hellinger distance within PPMC is introduced. Section 4 describes the simulation study while Section 5 discusses an application to real data to show the effectiveness of the method in use. Concluding remarks end the paper.

2. MIRT Models

A fundamental assumption of IRT models is the conditional or local independence defined as

$$P(Y = y|\theta) = \prod_{j=1}^{k} P\left(Y_j = y_j|\theta\right), \qquad (1)$$

where Y_j is the response variable vector for item j, with j=1,…,k items, and θ is the set of latent traits. Equation (1) holds conditionally to a single latent trait in unidimensional models, and to the specified dimensionality structure in multidimensional IRT (MIRT) models (see Zhang and Stout, 1999; Levy and Svetina, 2011). For this reason, the fit and the dimensionality assessments involve checking for the assumption of local independence. When data are multidimensional and a unidimensional IRT model is used, the assumption of local

independence is not met. A different situation occurs with locally dependent unidimensional models (Ip, 2010). Usually, MIRT models outperform separate unidimensional models for test consisting of multiple subtests, because they allow description of the data's complexity, direct estimation of the correlation among the traits, and taking into account potential hierarchy in the latent variables. As underlined by Gibbons, Immekus, and Bock (2007) and Ip (2010), if there is a predominant overall factor, the presence of multidimensionality has little effect on the unidimensional estimates. On the contrary, if there are strong specific factors beyond the general one, the unidimensional parameterization is seriously compromised, except when the traits are highly correlated. However, many IRT applications are only possible with unidimensional models, and when the amount of multidimensionality seems negligible, a unidimensional approach may be preferred. MIRT models involve increased complexity, and they may require a large sample size in order to estimate all parameters effectively.

Several MIRT models have been proposed to account for data structures involving multiple abilities in both an exploratory and a confirmatory setting (see, e.g., Adams, Wilson, and Wang, 1997; Reckase, 2009; Sheng and Wikle, 2007, 2008; Huo et al., 2015). In the present study, we consider a confirmatory approach with two-parameter normal ogive (2PNO) models for binary data, where Y_{ij} represents the 0-1 response variable for respondent i to item j, with i=1,…,n and j=1,…,k. If a simple structure for a test assessing a set of m specific domains is assumed with v=1,..,m, the 2PNO multi-unidimensional model (see, e.g., Sheng and Wikle, 2007), can be defined as follows

$$P\left(Y_{vij} = 1|\theta_{vi}, \alpha_{vj}, \delta_j\right) = \Phi\left(\alpha_{vj}\theta_{vi} - \delta_j\right), \qquad (2)$$

where α_{vj} denotes the v-specific item discrimination parameter for item j, δ_j is the difficulty parameter and θ_{vi} is the v-specific ability for respondent i. According to model (2), abilities may be correlated. This approach is more efficient at improving the measurement precision

than estimating different unidimensional models separately, especially when the tests are short (Wang et al., 2004).

In the case all items are further related to an overall trait θ_0, the 2PNO additive model (Sheng and Wikle, 2009) can be used

$$P\left(Y_{vij} = 1|\theta_{0i}, \theta_{vi}, \alpha_{0j}, \alpha_{vj}, \delta_j\right) = \Phi\left(\alpha_{0j}\theta_{0i} + \alpha_{vj}\theta_{vi} - \delta_j\right), \quad (3)$$

where the total number of latent abilities is m+1. With respect to the trait correlations, the most common approach in the literature is to assume orthogonal dimensions by fitting a bi-factor model (Gibbons & Hedeker, 1992) or to allow only the specific traits to correlate. On the other hand, in the approach proposed by Sheng and Wikle (2009), all correlations among the traits could be estimated. Many other multidimensional approaches are possible as well (see, e.g., de la Torre and Song, 2009; Sheng and Wikle, 2008; Reckase, 2009).

Bayesian estimation of these models resorts to MCMC algorithms as the joint posterior distribution of interest has an intractable form. In particular, the use of the Gibbs sampler (Geman and Geman, 1984) was proposed by Béguin and Glas (2001) for model (2), and Sheng and Wikle (2009) for model (3) with correlated traits by extending the original work of Albert (1992) for the unidimensional 2PNO model. See Sheng and Wikle (2007; 2008) and Sheng (2008a; 2008b; 2010) for further details on the Gibbs sampler implementation, and Fontanella, Fontanella, Valentini, and Trendafilov (2019) for recent developments on Bayesian MIRT models.

3. Posterior Predictive Assessment for Investigating Local Independence in IRT Models

3.1 Posterior Predictive Model Checking

PPMC techniques (Rubin, 1984; Gelman, Meng, and Stern, 1996) work by comparing observed data with replicated data generated or predicted by the model by using diagnostic measures that are sensitive to model misfit (Sinharay et al., 2006).

Given the data y, let $p(y|\omega)$ and $p(\omega)$ be the likelihood for a model depending on parameters ω and the prior distribution of ω, respectively. In the IRT context, ω includes the item parameters, the person parameters, and the trait correlations. The replicated data y^{rep} are drawn from the posterior predictive distribution (PPD) given by

$$p(y^{rep}|y) = \int_{\omega} p(y^{rep}|\omega)\, p(\omega|y) d\omega. \quad (4)$$

Once a suitable discrepancy measure or test quantity $D(\cdot)$ is chosen, the method works by comparing the posterior distribution of $D(y,\omega)$ to the posterior predictive distribution of $D(y^{rep},\omega)$, where substantial differences indicate poor model fit. Discrepancy measures based on data only (test statistics) can be used as well. The discrepancy measures should be able to capture relevant features of the data and differences among data and the model (Levy et al., 2009).

Due to difficulties in finding the PPD analytically, Rubin (1984) suggested drawing a number *R* of replications via MCMC and then drawing *R* replicated datasets to compare the predictive and realized discrepancies. As a first step in the application of PPMC to real data, a graphical analysis is conducted by plotting $D(y^{rep,r},\omega^{r})$ versus $D(y,\omega^{r})$, with $r=1,\ldots,R$, or, in the case of test statistic, comparing the histogram of $D(y^{rep,r})$ to the observed $D(y)$. If the model fits, then observed data should look similar to replicated data under the model.

When the plots are ambiguous or multiple checks are needed, it is possible to resort to the posterior predictive *p*-value (PPP-value or Bayesian *p*-value) defined as "the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity" (Gelman et al., 2014) and expressed as follows:

$$\text{PPP-value} = p(D(y^{rep},\omega) \geq D(y,\omega)|y) = \int_{D(y^{rep},\omega)\geq D(y,\omega)} p(y^{rep}|\omega)p(\omega|y)dy^{rep}d\omega. \quad (5)$$

PPP-values close to 0 or 1, depending on the chosen discrepancy, denote a failure of the model to describe the data correctly as the realized values fall far in the tails of the distribution of the

discrepancy measure based on PPD. On the other hand, values of approximately 0.5 indicate good fit. The PPP-values are estimated by the proportion of MCMC simulations for which the replicated discrepancy equals or exceeds the realized one. Despite they provide a quantitative measure of lack of fit, the PPP-values are not able to address the magnitude of the difference between the two distributions properly.

As underlined by Levy et al. (2009), PPMC has several advantages over traditional techniques. The method is easy to apply and flexible because the reference distribution is built empirically and it does not require regularity conditions or asymptotic results. Moreover, PPMC relies on Bayesian estimation, which is based on the full posterior distribution: compared with maximum likelihood techniques, which are based on a point estimate, the method is able to directly incorporate uncertainty into the estimation. However, it should be highlighted that PPMC is about assessing the discrepancies between a model and data, and not testing the correctness of a model (Gelman et al., 1996).

3.2 Discrepancy Measures Checking for Local Independence in IRT Models

In PPMC, the role and choice of the discrepancy measure or test quantity is fundamental. As underlined in Gelman et al. (2014), "test quantities play the role in Bayesian model checking that test statistics play in classical testing". Under the Bayesian umbrella, test quantities may depend on data only or, thanks to the posterior distribution, they may be generalized to depend on the model parameters too. The definition of a proper test quantity depends on the particular aspects of the data one wants to check and, for this reason, is strongly influenced by the goal of the research. While in the traditional hypothesis testing the choice of test statistics depends on power, e.g. the probability of rejecting the null hypothesis when in fact it is not true, model checking focuses on how the model fits in relation to the specific aspects that are relevant in practice. For this reason, the choice of the discrepancy measure should reflect the researcher's

inferential interests (Gelman et al., 1996). As discussed in Sinharay (2005) and Sinharay et al. (2006), although any function of data and model parameters could be used as discrepancy measure, not all measures are useful. Defining and defending the choice of a discrepancy measure is crucial in PPMC. As suggested by Gelman et al. (2014), ideally the discrepancy measure should not only reflect aspects of the model relevant for the research goal, but it should also measure features of the data that are not directly addressed by the model. This last feature is able to make the difference between a useful or an unsuitable discrepancy measure. Clearly, the chosen discrepancy measure should have adequate power. For a discussion on the role of PPMC test quantities in IRT, see Sinharay (2005; page 379).

Previous studies using PPMC for checking the assumption of local independence worked with diagnostic measures based on the association or on covariance/correlation among item pairs (Sinharay et al., 2006; Levy et al., 2009; Levy, 2011; Levy and Svetina, 2011; Levy et al., 2015). In fact, discrepancy measures investigating associations between item pairs can hopefully detect the misfit of models that do not take into account the correct dimensionality and dependency structure in the data.

The Mantel-Haenszel (MH) statistic (e.g., Agresti, 2002, p. 234) is based on data only and it is defined as the odds ratio conditionally to the rest score *s*, i.e., the raw test score obtained by excluding the two items. For each pair of items *j* and *j'*, with *j, j'*=1,…,*k*, the MH statistic is

$$\mathrm{MH}_{jj'} = \frac{\sum_s n_{11s} n_{00s}/n_s}{\sum_s n_{10s} n_{01s}/n_s}, \qquad (6)$$

where the generic term $n_{tt's}$ is the number of subjects with rest score *s* who score *t* on item *j* and *t'* on item *j'*, with *t*, *t'*=0,1, and n_s is the number of subjects with rest score *s*. The effectiveness of the MH statistic in investigating the local independence assumption relies on conditioning to the rest score, which is a proxy of the latent ability. If the local independence assumption holds, the MH statistic is close to one. If the local independence does not hold, the MH statistic is greater than one for within-cluster items and smaller than one for between-cluster items (see

Sinharay et al., 2006). Bad fit is reported with PPP-values close to zero (when MH > 1) or close to one (when MH < 1).

The model-based covariance (MBC; Reckase, 1997) depends on both data and model parameters as follows:

$$\mathrm{MBC}_{jj'} = \frac{\sum_{i=1}^{n}\left(Y_{ij}-E(Y_{ij})\right)\left(Y_{ij'}-E(Y_{ij'})\right)}{n}, \qquad (7)$$

where E(Y_{ij}) is the expected value of the response variable depending on the specific IRT model. The MBC is found to be effective as it measures the covariance among item pairs by explicitly conditioning on the latent variable (Levy et al., 2009). If the local independence assumption holds, the MBC is close to zero. If the local independence does not hold, the MBC is greater than zero for within-cluster items (PPP-values are close to zero) and smaller than zero for between-cluster items (PPP-values are close to one). Another useful measure which produces comparable results to the MBC in terms of graphical analysis and PPP-values is the Yen's Q_3 (Yen, 1993), defined as the correlation among the residuals of item pairs.

Lastly, Levy and Svetina (2011) proposed an overall measure, namely the generalized dimensionality discrepancy measure (GDDM)

$$\mathrm{GDDM} = \frac{1}{k(k-1)/2}\sum_{j>j'}\left|\frac{\sum_{i=1}^{n}\left(Y_{ij}-E(Y_{ij})\right)\left(Y_{ij'}-E(Y_{ij'})\right)}{n}\right|. \qquad (8)$$

The GDDM is a unidirectional measure of average conditional covariance defined as the mean of the absolute values of MBC over unique item pairs. When the GDDM is equal to zero, a "weak" local independence for all the item pairs is assumed (Levy and Svetina, 2011). If the assumption of local independence is violated, the GDDM is greater than zero and the PPP-value will be close to zero. Levy et al. (2015) proposed a standardized version of the GDDM (SGDDM) defined as a mean absolute conditional correlation to improve the interpretability of the results. The results of SGDDM in terms of graphical analysis and PPP-values are comparable to the ones obtained by the GDDM.

3.3 PPMC Based on the Hellinger Distance

Previous studies (see, e.g., Sinharay et al., 2006; Levy et al., 2009; Levy, 2011; Levy and Svetina, 2011; Levy et al., 2015) investigated the potentialities of PPMC in terms of graphical analysis and PPP-value estimates. While PPP-values only count the number of replications for which the predictive discrepancy exceeds the realized one, the researcher may be interested in measuring the size of the difference itself. To reach this goal, Wu et al. (2014) proposed the use of relative entropy (RE; Kullback and Leibler, 1951) with full and limited information statistics as discrepancy measures. In fact, the RE (also known as Kullback-Leibler divergence or information) quantifies the information a distribution *P* loses while approximating to another distribution *Q*. In the case of the distributions *P* and *Q* of a continuous random variable, the RE is defined as follows:

$$D_{KL}(P \parallel Q) = \int \ln \frac{p(y)}{q(y)} p(y) dy, \qquad (9)$$

where $p(y)$ and $q(y)$ are the probability density functions of y. The RE measures the relation between the two distributions asymmetrically, meaning that, depending on the order of the arguments, the values of the RE may differ. Moreover, the RE is always non-negative. The smaller the RE value, the smaller the information loss and the more the two distributions are similar.

In the PPMC setting, given a discrepancy measure $D(y,\omega)$ depending on both data and model parameters, the RE becomes

$$D_{KL}(P \parallel Q) = \int \ln \frac{p\left(D(y,\omega)\right)}{p\left(D(y^{rep},\omega)\right)} p\big(D(y,\omega)\big)\, dy d\omega. \qquad (10)$$

In this case, the RE displays the information the realized distribution loses while approximating the predictive distribution of a given discrepancy measure. The more similar the realized and predictive distributions are, the smaller the RE values is, and consequently, the

better the model fit is. Despite the RE is a very powerful information measure, it has several drawbacks that make its use difficult in applied settings. Firstly, the RE is not symmetric. Moreover, RE ≥ 0. As a consequence of its unboundedness, it is difficult to interpret the outcomes. What can be said about the size of the difference between the two distributions?

To evaluate the magnitude of the discrepancy between the realized and the predictive distributions, we propose the use of the Hellinger (H) distance which is defined in terms of the Hellinger integral (Hellinger, 1909) as follows:

$$\mathrm{H}(P,Q) = \sqrt{1 - \int \sqrt{p(y)q(y)}\, dy}, \qquad (11)$$

where the term $\int \sqrt{p(y)q(y)}\, dy$ is the Bhattacharyya coefficient (Bhattacharyya, 1943). The H distance is used to quantify the distance between two probability measures and it is a proper distance metric in the mathematical sense, by satisfying the properties of nonnegativity, symmetry, and triangle inequality. The H distance is also bounded between 0 and 1, where 0 means that the two distributions are indiscernible and 1 that they are maximally distant. This feature makes the interpretation of the H distance more convenient than the RE in real data applications.

In the context of PPMC, we propose to use (11) as follows:

$$\mathrm{H}(P,Q) = \sqrt{1 - \int \sqrt{p\big(D(y,\omega)\big)p(D(y^{rep},\omega))}dyd\omega}. \qquad (12)$$

Analogously to the RE, the H distance is used with discrepancy measures that depend on both data and model parameters, as the MBC or the GDDM. The direct calculation of (12) is computationally demanding and it is usually done via MCMC. Specifically, given the MCMC simulations, it is estimated by using the normal kernel density estimates to represent the probability density functions of the realized and the predictive discrepancy measures.

In order to check for local independence, we propose the use of the H distance with the MBC discrepancy measure (MBC-H) to take into account a fit measure for each item pair and with the GDDM (GDDM-H) to evaluate the overall fit based on item pairs. The H distance is then able to quantify the misfit as “the amount of the difference between the realized and the predictive distribution of the chosen discrepancy". Moreover, the H distance values can be compared directly, so this measure is suitable to be used for model comparison.

4. Simulation Study

A simulation study is conducted to examine the performance of the proposed MBC-H and GDDM-H at detecting the misfit of the unidimensional 2PNO model when data follow the multi-unidimensional model (2) or the additive model (3) with m=2. In order to compare the results with the existent solutions in the literature under PPMC, we estimated the PPP-values for the MH statistic, the MBC, and the GDDM. Additionally, the RE for MBC (MBC-RE) and for GDDM (GDDM-RE) were estimated.

The results of the simulations are presented for a test with k=10 items consisting of two subtests (k_1=k_2=5) and a sample size of n=1,000, by manipulating the correlations among the traits. We also manipulated the sample size by using n=2,000 and the test length by using k=20 with k_1=k_2=10. The results are included in Appendix A and Appendix B, respectively.

For each simulation condition, 100 replications are used. For each replication, the simulation procedure can be summarized as follows.

1. A n×k dataset with binary item responses is simulated according to the data-generating model and the simulation conditions.
2. The parameters of the data-analysis model are estimated via the Gibbs sampler, where all the MCMC draws defining the posterior distribution are recorded. Specifically, for each model parameter, a vector of the MCMC sampled values is created with length equal to the effective number R of iterations (total minus burn-in iterations).

3. The *R* iterations are thinned by a constant to be used within PPMC. Even if there is a debate in the literature on the usefulness of thinning the MCMC chains, this procedure is adopted in this work to reduce the computational time of PPMC.
4. The final synthesis of the output consists of the following elements: two $k \times k$ matrices with the PPP-values for the MH statistic and the MBC for each item pair, the PPP-value for GDDM, two $k \times k$ matrices containing the MBC-RE and the MBC-H for each item pair, the GDDM-RE and the GDDM-H.

For each replication in the case of k=10 and n=1,000, the total number of MCMC iterations is equal to 5,000, with 1,000 as burn-in iterations (R=4,000). The convergence of the Gibbs sampler was assessed by using the Gelman-Rubin statistic starting from a single chain divided into sub-chains (see Sheng, 2010). The effective samples are thinned by 4 so that 1,000 samples are used in PPMC.

For each simulation condition, the results are synthetized by computing the arithmetic mean of each measure over the 100 replications. We use the proportion of extreme PPP-values (below 0.05 or above 0.95) among the item pairs to summarize the results for the MH statistic and the MBC. We report some descriptive statistics for the MBC-RE and the MBC-H. All these results are further reported for the pairs where both items belong to subtest 1 (“within1”), both items belong to subtest 2 (“within2”) and items belonging to different subtests (“between”). MATLAB packages provided by Sheng (2008a; 2008b; 2010) are used for data generation and for MCMC estimation of models (2) and (3). The Authors wrote MATLAB specific programs for estimating the unidimensional model and performing PPMC.

The case of correspondence between the data-generating model and the analysis model for all models is discussed before introducing the case of multidimensional data analyzed with the unidimensional model.

Same Data-Generating and Data Analysis Model

Using the same model to simulate and analyze the data is fundamental to investigate the capability of the proposed approach to report good fit correctly.

For the unidimensional model, the item parameters are drawn from $\alpha \sim U(1,2)$ and $\delta \sim U(-2,2)$, while the trait scores are $\theta \sim N(0;1)$. Model parameters are estimated via the Gibbs sampler using standard normal priors for the latent trait and the difficulty parameters. A standard normal distribution truncated at zero to the left is used as prior distribution for the discrimination parameters to ensure positivity. With respect to the multi-unidimensional model (2), the cases of no correlation ($\rho_{12}=0$), weak correlation ($\rho_{12}=0.3$), moderate correlation ($\rho_{12}=0.6$) and strong correlation ($\rho_{12}=0.9$) between θ_1 and θ_2 are taken into account. Item parameters are drawn from $\alpha_1 \sim U(1,2)$, $\alpha_2 \sim U(1,2)$, and $\delta \sim U(-2,2)$, and latent scores from a multivariate normal $\theta=(\theta_1, \theta_2) \sim MN(0;\Sigma)$, where Σ is the correlation matrix. We specified conjugate standard normal priors for the item parameters, and a conjugate multivariate normal as a prior for the covariance matrix of the latent traits (see Sheng, 2008b). For the additive model (3), two specific latent traits θ_1 and θ_2, and an overall trait θ_0, related to all item responses are considered ($m+1=3$). Correlations among all traits are set equal $\rho_{01}=\rho_{02}=\rho_{12}$ and are again fixed equal to 0, 0.3, 0.6 and 0.9. The item parameters are drawn from $\alpha_0 \sim U(1,2)$, $\alpha_1 \sim U(1,2)$, $\alpha_2 \sim U(1,2)$ and $\delta \sim U(-2,2)$, while the trait scores from $\theta=(\theta_0, \theta_1, \theta_2) \sim MN(0;\Sigma)$. Conjugate standard normal priors are specified for the item parameters, and a conjugate prior is used for the covariance matrix of the latent traits (see Sheng, 2010).

For all conditions, no extreme PPP-values are reported for the MBC and MH statistic computed on the 45 item pairs indicating good fit. The PPP-value for the GDDM is equal to 0.512 for the unidimensional model and it is in the range [0.411 - 0.558] for different correlations in the additive model, indicating good fit. For the multi-unidimensional model, the PPP-value for GDDM is in the range [0.480 - 0.532] for ρ_{12} = 0.0, 0.3, 0.6 while it is equal to

0.277 when $\rho_{12} = 0.9$. Even if this last PPP-values is not extreme, the deviation from 0.5 may be due to the strong correlation between the traits which make the data nearly unidimensional.

Table 1 reports the results on the RE. Some descriptive statistics are used to synthetize the MBC-RE for the item pairs and the median value is reported for the 10 "within1", 10 "within2", and 25 "between" item pairs.

INSERT TABLE 1 HERE.

In case of good fit, the realized and the predictive distributions of the discrepancy measure should be overlapping and the RE will be close to zero. From Table 1, it is observed that the MBC-RE slightly increases as the correlation between the two traits increases. The highest increase is observed for the multi-unidimensional model from $\rho_{12} = 0.6$ to $\rho_{12} = 0.9$. There are no meaningful differences by item pair type. The same behavior is observed for the GDDM-RE, where for the multi-unidimensional model with $\rho_{12} = 0.9$ the value is equal to 6.437 denoting a relevant change in the order of magnitude of the RE. According to the RE, what can be said about the model fit? Are the RE values close to zero enough to indicate good fit? As $RE \geq 0$, the answer is not straightforward.

INSERT TABLE 2 HERE.

Table 2 shows the results for the H distance. The average value for the MBC-H is 0.45 in the unidimensional model and slightly lower in the multi-unidimensional and additive models (the range is 0.319-0.457). Again, the highest value is reported for the multi-unidimensional model with strongly correlated traits and no meaningful differences emerge from the results by item pair type. The same behavior is observed for the GDDM-H. The measures are computed for all the item pairs and then averaged. For this reason, it is very difficult to obtain values of the H distance very close to zero, even when the correct model is fitted to the data. The results could be further deepened by looking at the quartiles for the different simulation conditions. In fact, the third quartile (Q_3) is around 0.5 for the unidimensional model and 0.4 for the

multidimensional ones, excluding the cases of strong correlations. We expect that MBC-H values below Q_3 are associated to item pairs reporting good fit. When the MBC-H values are lower than the Q_1 values (around 0.4 for the unidimensional model and around 0.3 for the multidimensional ones) we expect that the item pairs show very good fit. As $0 \leq H \leq 1$, fixing a threshold at 0.5 may be helpful in evaluating model fit. The value of 0.5 was chosen following, as usual in complex situations, a rule of thumb obtained by a detailed analysis of the simulation results under different conditions. In particular, we considered the mean and the quartiles presented in Table 2 under the "null" conditions. On average, all MBC-H and GDDM-H values are smaller than 0.5, indicating good fit. Moreover, the Q3 are lower than 0.5 for all conditions, except for the unidimensional one (but the value is 0.524) so we believe that by considering the threshold of 0.5 we could expect not more than ¾ of the item pairs under the threshold when the model fits the data. We believe that this threshold is rather flexible to minimize the risk of assessing model fit incorrectly. Also, a value of the H lower than 0.5 means that the distance between the realized and the replicated distribution of the discrepancy measure is closer to the case of H=0 (the two distributions overlap) than to the case of H=1 (maximum distance).

The proportion of item pairs with MBC-H equal or higher than 0.5 (Prop ≥ 0.5) is equal or very close to zero for the conditions involving a multidimensional approach, meaning that the threshold of 0.5 can be effectively used in practice. However, when data are unidimensional, the approach based on the H distance shows a weakness by reporting about 30% of item pairs with MBC-H ≥ 0.5.

Additional simulations were conducted by increasing the sample size to n=2,000 (Appendix A) and the total test length to k=20 (Appendix B). The results confirm the main findings. In fact, when increasing the sample size, no extreme PPP-values are reported both for the MBC and the MH statistics. The results on the RE and the H distance (see Tables A1 and A2) are

very similar to the ones reported in Tables 1 and 2. Only for the multi-unidimensional model with ρ=0.9, an increase in the MBC-RE, the GDDM-RE, and the GDDM-H is observed, meaning that the latent structure may be conceived as unidimensional.

When increasing the test length, again no extreme PPP-values are reported. For the RE (Table B1) we can generally observe that the values are slightly higher than the ones of the baseline setting. The values of the MBC-H (Table B2) also show on average an increase with respect to the case of *k*=10. This may be due to the presence of a large number of item pairs (190) which make much more difficult to obtain overlapping realized and predictive distributions of the discrepancy measure.

Multidimensional Data-Generating Models vs. Unidimensional Estimated Model

The focus of this simulation study is on multidimensional data-generating structures, where the data-analysis model is unidimensional. The simulation settings follow the ones used in the previous study. When the correlation among the traits is very strong, the model is close to a unidimensional approach because the two dimensions nearly overlap. Of course, the latent traits, even if strongly correlated, can be different in their empirical interpretation.

INSERT TABLE 3 HERE.

The results in Table 3 show that the MH statistic and the MBC perform differently for the case of uncorrelated traits with multi-unidimensional data. In fact, while the MH statistic correctly reports model misfit based on a rather high proportion of extreme PPP-values (0.8) the MBC fails by reporting only 22.2% of extreme PPP-values. The behavior of the MBC is unexpected and may be attributed to the peculiarity of the generating model, which resembles two separate unidimensional models. To investigate this condition more in detail, we report in Appendix C the generating item parameters for the uncorrelated multi-unidimensional data (Table C1) and the item parameter estimates (and their standard deviations) according to the

unidimensional model (Table C2). The item discrimination estimates for the unidimensional model are close to zero for the first subtest (items 1-5) and higher than one for the second subtest (items 6-10). This means that the latent trait is defined by the items of the second subtest only, which are related to the same ability. This is true for all the replications of this simulation condition as the datasets were generated starting from the same simulated item parameters. For this reason, the MBC reports misfit only for the item pairs where both items belong to subtest 1 ("within1"). Unlike the MH statistic, the MBC is a model-based indicator and it is affected by the item parameter estimates. We also estimated the item parameters via marginal maximum likelihood to understand if the results could be affected by the choice of the prior distributions for the item parameters in MCMC, but we obtained similar estimates.

To deepen the results, we have checked when the MBC starts working appropriately by considering more correlation conditions (ρ_{12}=0.1, 0.2, 0.4). As the correlation ρ_{12} increases, the MBC reports misfit. At ρ_{12}=0.2, the MBC reports about 40% of extreme PPP-values, reaching the 100% at ρ_{12}=0.4. Differently, the MH statistic works well by reporting all extreme PPP-values, already at condition ρ_{12}=0.2.

The cases of strong correlation is again peculiar. At ρ_{12}=0.9, the proportion of extreme PPP-values is close to zero for the MH statistic and equal to zero for the MBC.

With respect to the additive data, both the MH statistic and the MBC report strong evidence of misfit of the unidimensional model for uncorrelated traits and for $\rho_{01}=\rho_{02}=\rho_{12}$=0.3. The performances of the discrepancy measures are similar as the proportion of extreme PPP-values decreases at $\rho_{01}=\rho_{02}=\rho_{12}$=0.6 and then is equal to zero at $\rho_{01}=\rho_{02}=\rho_{12}$=0.9. No meaningful differences are found for the results by item pair type. In fact, we found extreme PPP-values both for the within item pairs, where positive conditional covariance is expected and PPP-values will be close to zero, and for between item pairs, where negative conditional covariance is expected and PPP-values will be close to one.

For all the simulation conditions, the GDDM reports poor fit as the estimated PPP-value is always zero, with the only exception of strongly correlated traits where it is still extreme and close to zero.

INSERT TABLE 4 HERE.

Table 4 shows that, for the conditions involving zero or weak trait correlations, the mean MBC-RE among item pairs goes to infinity. The median values are always larger than the MBC-RE values in Table 1, where the same data-generating model and analysis-model was used. Differences in the median values for item pair type are noticed, for example in the case of uncorrelated multi-unidimensional data, where the "within1" items are characterized by a high median MBC-RE. Analogously, the GDDM-RE goes to infinity with the only exception of the conditions of strongly correlated traits meaning that the discrepancy between the predictive and the realized distribution of GDDM is maximum. Overall, the REs are far from zero, denoting poor fit and the results on MBC-RE show a large variability among the item pairs and for the different simulation conditions, making cumbersome the interpretation of the results.

INSERT TABLE 5 HERE.

The results in Table 5 show poor fit of the unidimensional model when data are multidimensional. In fact, for most conditions, the mean and median values of the MBC-H are higher than 0.8, denoting that the distance between the predictive and the realized distribution of the MBC is rather high. Coherently with the results on the PPP-values in Table 3, exceptions are observed for the multi-unidimensional data with $\rho_{12}=0$ and the cases of strongly correlated traits, where data are conceived as unidimensional. Excluding the case of strong correlation, it can be observed that the magnitude of misfit increases as the correlation increases for multi-unidimensional data but not for additive data. In fact, the specification of the additive model includes a general, overall trait. For this reason, as the correlation increases, data simulated by

an additive model may be conceived as close to unidimensional and the magnitude of misfit reported is lower. The proportion of item pairs with MBC-H equal or higher than 0.5 is equal or very close to 1 for all the conditions, except for multi-unidimensional data with ρ_{12}=0, meaning that the threshold of 0.5 works well for the H distance in detecting the violation of local independence. By looking at the quartiles, the Q_1 and Q_3 are higher than 0.7 and 0.9, respectively, excluding the case of multi-unidimensional data with ρ_{12}=0 and the cases where ρ_{12}=0.9 which are peculiar. This means that MBC-H values higher than 0.7 show large misfit.

The GDDM-H is equal to one, reporting the maximum possible distance between the predictive and the realized GDDM (lower values are reported for strongly correlated traits). On average all MBC-H and GDDM-H values are higher than the threshold 0.5, indicating bad fit.

INSERT FIGURE 1 HERE.

We checked the distribution of the MBC-H for each item pair under the different simulation conditions to understand if the estimated value is stable among the 100 replications. Figure 1 reports the histograms of the MBC-H values for specific item pairs under several simulation conditions. The upper figures represent the case of the same data-generating and data-analysis model (from left to right: unidimensional, multi-unidimensional, and additive) while the lower figures represent the case of multidimensional data analyzed with a unidimensional model (from left to right: multi-unidimensional and additive data). Clearly, the MBC-H in the upper figures shows good fit but it is associated with an higher variability in comparison to the MBC-H in the lower figures, showing bad fit especially in the case of additive data.

An important advantage of the H distance with respect to the PPP-values and the RE, is that it can be used for model comparison purposes. In fact, we compared the estimated MBC-H for all the 45 item pairs under all the simulation conditions in which the data are multidimensional. For all the conditions, when the data-analysis model is unidimensional, the estimated MBC-H is higher than in the case a multidimensional approach is used, for all the item pairs meaning

that the multidimensional approach is always able to fit the data better than the unidimensional one. Unlike the H distance, it is not possible to use the PPP-values for making direct comparisons as they can only be interpreted as extreme (indicating misfit) or not. A graphical analysis under PPMC is also not feasible due to the presence of multiple checks, e.g., a discrepancy measure for each item pair. The RE values cannot be directly compared either due to the unboundness of the measure. The advantage of the H distance over the PPP-values in making model comparison is clear by considering a fit measure for each item pair. Overall, it is possible to consider both the proportion of extreme PPP-values and the mean of the H distances among the item pairs for choosing among competing models.

The results of the simulations conducted with n=2,000 and k=20 are consistent with the ones obtained in the baseline approach (n=1,000 and k=10). When the sample size is increased, the proportion of extreme PPP-values (Table A3) is equal or higher than in the baseline study for all multidimensional data. The only exception is the uncorrelated multi-unidimensional data, where no extreme PPP-values are observed for the MBC, a result which goes in the same direction of what emerged in the case of n=1,000. Looking at Tables A4 and A5, overall the RE and H distance mean values are equal or higher than the corresponding results of the main study, denoting an improvement in the capability of detecting misfit. However, very similar results are obtained in the boundary cases (multi-unidimensional data with ρ=0.0 and 0.9, and additive data with ρ=0.9). Increasing the test length to k=20 confirms the main findings. Again, the peculiar case of uncorrelated multi-unidimensional data is associated to no extreme PPP-values (see Table B3), the values of the RE and the H distance (Tables B4 and B5) are usually higher than in the baseline case, with few exceptions.

5. Empirical Application

The empirical data come from a survey conducted by the University of Bologna (Italy) to investigate the perceptions of the residents in tourist cities on the impact of the tourism industry (Bernini, Matteucci, and Mignani, 2015). The data consist of the responses given by 794 residents to 10 items of a questionnaire, where five items address the perceived benefits and five items address the costs. Items on benefits concern economic support, quality of life, public services, job opportunities, and cultural activities, while items on costs address general cost of life, crime rate, environmental damage, traffic, and pollution. Item responses are binary, where Y_{ij}=1 denotes a positive perception (high benefit or low cost) and Y_{ij}=0 indicates a negative perception (low benefit or high cost).

The unidimensional, the multi-unidimensional, and additive models are fitted via the Gibbs sampler with a total of 10,000 iterations (5,000 as burn-in iterations). Two specific factors are assumed: a first one related to the benefit items 1-5 and a second one related to the cost items 6-10.

The deviance information criterion (DIC) suggests that the additive model (DIC=6916.38) fits the data better than the unidimensional model (DIC=7742.08) and the multi-unidimensional model (DIC=7183.11). To investigate goodness of fit model by model, 1,000 MCMC posterior samples drawn for each model parameter (5,000 effective iterations thinned by 5) are used to implement PPMC with the techniques discussed in this paper. Table 6 reports the proportion of extreme PPP-values for the MH statistic and the MBC and the PPP-value estimate for the GDDM under the three models. The MATLAB function ppmcplt, provided by Sheng (2010), is used to show a graphical representation of the extreme PPP-values for the 45 item pairs to give an immediate picture of model fit.

INSERT TABLE 6 HERE.

INSERT FIGURE 2 HERE.

For the unidimensional model, Figure 2 shows that the number of extreme PPP-values is equal to 37 (82.2%) for the MH statistic and 35 (77.8%) for the MBC, suggesting that the local independence assumption does not hold for the unidimensional model. Clearly, we see that both the MH statistic and the MBC have PPP-values close to zero for the within-cluster items and close to one for between-cluster items, meaning that positive or negative conditional covariance is observed, respectively.

For both discrepancy measures, item 10 (pollution) is associated with all extreme PPP-values, meaning that this item contributes to the misfit for this model. Another critical item is 7 (crime rate), which involves all extreme values for MH and eight extreme values for MBC. The remaining items show extreme PPP-values for most pairs, with the only exception being item 5 (cultural activities).

For the multi-unidimensional model, a correlation of ρ_{12}=0.53(0.00)[1] is estimated between the two specific traits. Looking at the item pairs in Figure 2, not more than four extreme PPP-values are reported for each item and about the 30% of item pairs show bad fit.

Within the additive model, the estimated correlations among the overall perception and the specific perceptions on benefits and costs are equal to ρ_{01}=0.29(0.05) and ρ_{02}=0.21(0.04), respectively, while the correlation between the two specific traits is ρ_{12}=0.10(0.03). Only 15.6% and 8.9% of PPP-values are extreme according to the MH statistic and the MBC, respectively. Some extreme PPP-values are observed, especially for pairs involving items 9 and 10 (traffic and pollution).

The PPP-value for the GDDM is always extreme (0.000) meaning that, according to this generalized discrepancy measure, all the three models show poor fit. It seems that this measure is very sensitive to the presence of only few item pairs showing positive or negative conditional covariance.

[1] Monte Carlo standard error in brackets.

Table 7 provides the results for the RE and the H distance.

INSERT TABLE 7 HERE.

With respect to the RE, the results show a great variability, also for item pair type, and it is not possible to make a direct comparison among the three models. Given that the RE is unbounded, it is difficult to interpret the results. The GDDM-RE goes to infinity for the unidimensional model, but it is also far from zero for the other two models.

Unlike the RE, the results on the H distance are more easily interpretable. The MBC-H values show a large variability on the item pairs. However, we can say that, on average, the distance between the predictive and the realized MBC is 0.83 for the unidimensional model, 0.54 for the multi-unidimensional model and 0.44 for the additive model. According to the threshold of 0.5, only the additive model shows good fit. The proportion of item pairs with MBC-H $\geq$ 0.5 is 0.911 for the unidimensional model, 0.511 for the multi-unidimensional model, and 0.378 for the additive model. We also compared the values of the MBC-H for all the item pairs on the three models. Overall, 40 out of 45 item pairs have MBC-H lower with the additive model than with the unidimensional one. Moreover, 30 item pairs out of 45 present lower MBC-H for the additive model with respect to the multi-unidimensional model. The GDDM-H show the maximum possible distance for the unidimensional and the multi-unidimensional model, and it is also very high (0.98) for the additive model. Figure 3 highlights the MBC-H values higher than 0.5 or 0.8 for the item pairs.

INSERT FIGURE 3 HERE.

There is a considerable improvement from the unidimensional to the additive model. This means again that it is likely that the response data follow an additive structure.

By using discrepancy measures for item pairs, it is possible to deepen the investigation of model fit by considering each item pair separately. For example, in Figure 4 we compare the realized (MBC obs) and the predictive distribution (MBC rep) of the MBC in the additive

model for the items 7 and 10 (on the left) and the items 1 and 5 (on the right). The PPP-value for the $MBC_{7,10}$ is 0.9650 and MBC-$H_{7,10}$=0.7125: the item pair shows poor fit. On the other hand, items 1 and 5 show good fit with a PPP-value for $MBC_{1,5}$ equal to 0.5180 and MBC-$H_{1,5}$=0.1168.

INSERT FIGURE 4 HERE.

The investigation of fit for each item pair makes it clear the advantage of using the H distance with respect to the approach based on PPP-values. Let us take the item pair (5, 6) as an example, where item 5 belongs to the first subtest while item 6 to the second one. For the unidimensional model, the PPP-value for the $MBC_{5,6}$ is equal to 0.96 denoting misfit. Moreover, the PPP-values for the $MBC_{5,6}$ are equal to 0.66 and 0.50 for the multi-unidimensional and the additive model, respectively. Both PPP-values are not extreme, but it is not possible to choose which model fits the item pair data best. On the other hand, the MBC-$H_{5,6}$ is equal to 0.72 for the unidimensional model, 0.40 for the multi-unidimensional model, and 0.11 for the additive model. By using the Hellinger distance on the item pair, not only we can say that the unidimensional model shows misfit while the two multidimensional models show good fit by considering the threshold of 0.5, but also we can compare the H distance values directly. Consequently, between the two "fitting" multidimensional models, the additive one is the best as it is associated with the lowest H distance value for the item pair.

6. Concluding Remarks

The use of the Hellinger distance within posterior predictive assessment is proposed to investigate the assumption of local independence for 2PNO models by focusing on multidimensional data analyzed with the unidimensional model. The results of the simulation studies show that the PPP-values and the H distance are coherent in judging model fit, while the outcomes based on the RE are much more difficult to interpret. However, the H distance

outperforms the existing methods in providing a quantitative, easily interpretable measure of the amount of the difference between the predictive and the realized distribution of the discrepancy measure (MBC). The H distance is found to be effective in detecting misfit of the unidimensional approach for multidimensional data and to investigate if the misfit is due to specific items that prevent from fulfilling the local independence assumption. In the empirical study, the item response data are explicitly multidimensional and the results based on the H distance suggest that the additive model fits the data best in comparison to the unidimensional and the multi-unidimensional models.

The main strengths of the H distance, compared to traditional approaches based on PPP-values and RE, rely on the possibility a) to directly quantify the amount of misfit; b) to be used for model comparison purposes, c) to make more informative analyses on item pairs. First of all, the estimate of the MBC-H for each item pair is a measure of the degree of observed misfit itself. The H distance is a proper distance metric and it is bounded in the range 0-1, where H=0 means that the predictive and the realized distributions of the discrepancy measure overlap, while H=1 means that they are maximally distant. As the value for the H distance increases, the amount of observed misfit increases as well. Therefore, the Hellinger distance is a quantitative measure of misfit. Unlike the H distance, the PPP-value is used to report misfit when extreme values are observed and cannot be used to quantify the amount of misfit. Moreover, the PPP-value was found to be conservative in the sense of failing to reject an inadequate model (see Wu et al., 2014). Besides, the RE is not symmetric and does not have an upper bound, so it is not easily interpretable and comparable. Secondly, unlike the PPP-value and the RE, the estimates of the H distance are directly comparable over different conditions. It is possible to use the results on MBC-H for model comparison as one can compare the MBC-H estimates for the item pairs for different models. This feature may be very useful when comparing several models, but also the same multidimensional model with a

different number of latent traits. Lastly, more informative analyses on item pairs can be done to discover unusual patterns, e.g. high MBC-H values for pairs involving the same item or items.

The advantage of the Hellinger distance metric emerges especially in practical applications where we are interested not only in detecting misfit (all models are wrong for real data) but especially in quantifying the amount of misfit in order to make model comparison and choose the model which fits the data better than competing models. In practical applications, the H-MBC can be used to: a) leave out the models that show serious misfit by using the threshold of 0.5; b) compare the amount of misfit of different competing models and choose the model which fits the data best; c) identify, also through graphical plots, critical items that may involve misfit which are associated to high MBC-H in several pairs. Different degrees of misfit can be established by looking at the estimates of the MBC-H. This feature makes the H distance within PPMC a useful tool not only to assess global model fit but also to understand whether potential misfit can be ascribed to specific items. In this case, the researcher can evaluate the exclusion of one or more items from the analysis. The simulation study was needed to understand which values of the H distance are expected under different conditions.

The present study provides additional evidence on PPMC and the performance of the H distance but several limitations should be underlined. First of all, the approach based on the H distance cannot be used with discrepancy measures the depend on data only, such as the MH statistic, which were proved to be effective in PPMC. Secondly, using measures based on item pairs, it is very difficult to deepen the results on single items when the number of item pairs is large. In this case, aggregate results are considered. A further limitation of the study is due to the simulation setting. The subtest length is fixed to be the same, while, in practice, the number of items for each test subscale may differ. Analogously, the trait correlations are all set equal. Lastly, the simulation study showed that, when unidimensional data are analyzed with the

correct model (see Table 2), the approach based on the Hellinger distance shows a weakness by reporting about 30% of the item pairs with MBC-H equal or higher than the threshold 0.5.

Future research could enhance the models and the simulation settings provided in the present study. Furthermore, the strength of the relationship among the observed and the latent variables could be manipulated through the discrimination parameters. The method should be further validated with highly dimensional data (e.g., m=3) and for models requiring ordinal polytomous instead of binary responses. Moreover, the case of over-fitting where unidimensional data are analyzed through different multidimensional models should be considered as well. Finally, with the view to facilitate and improve the application of posterior predictive assessment, an interesting development could be to find a single discrepancy measure that is able to account for different aspects of fit, such as item and person fit, as well as overall model fit. When considering a single fit measure instead of a measure for each item pair, the advantage of using the H distance for model comparison will be even more relevant with respect to the approach based on PPP-value. In fact, under different competing models, the H distance values are directly comparable while the PPP-values are not. Future research on this topic could contribute to interesting developments.

References

Agresti, A. (2002). *Categorical data analysis*. New York: Wiley. DOI: 10.1002/0471249688

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269. DOI: 10.3102/10769986017003251

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society,* 35, 99-109.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 471-488. DOI: 10.1007/BF02296195

Bernini, C., Matteucci, M., & Mignani, S. (2015). Investigating heterogeneity in residents' attitudes toward tourism with an IRT multidimensional approach. *Quality & Quantity, 49,* 805-826. DOI: 10.1285/i20705948v11n2p427

de la Torre, J., Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33,* 620-639. DOI: 10.1177/0146621608326423

Fontanella, L., Fontanella, S., Valentini, P., & Trendafilov, N. (2019). Simple structure detection through Bayesian exploratory multidimensional IRT models. *Multivariate Behavioral Research,* 54(1), 100-112. DOI: 10.1080/00273171.2018.1496317

Gelman, A., Meng, X. L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733-807.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (Third Edition).* Boca Raton, FL: CRC Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741. DOI: 10.1109/TPAMI.1984.4767596

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423-436. DOI: 10.1007/BF02295430

Gibbons, R. D., Immekus, J. C., & Bock, R. D. (2007). The added value of multidimensional IRT models. *Multidimensional and hierarchical modeling monograph 1*. Chicago: Center for Health Statistics, University of Illinois.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217–233. DOI: 10.1177/0146621603027003003

Hellinger, E. (1909). Neue Begrundung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die Reine und Angewandte Mathematik*, *36*, 210–271.

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays in item response theory*. New York: Springer-Verlag. DOI: 10.1007/978-1-4613-0169-1_6

Huo, Y., de la Torre, J., Mun, E.-Y., Kim, S.-Y., Ray, A.E., Jiao, Y., & White, H.R. (2015). A hierarchical multi-unidimensional IRT approach for analyzing sparse, multi-group data for integrative data analysis. *Psychometrika*, 80(3), 834-855. DOI: 10.1007/s11336-014-9420-2

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology*, *63*, 395-416. DOI: 10.1348/000711009X466835

Kang, H.-A., & Chang, H.-H. (2016). Parameter drift detection in multidimensional computerized adaptive testing based on informational distance/divergence measures. *Applied Psychological Measurement*, *40*(7), 534-550. DOI: 10.1177/0146621616663676

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79-86. DOI: 10.1214/aoms/1177729694

Levy, R. (2011). Posterior predictive model checking for conjunctive multidimensionality in item response theory. *Journal of Educational and Behavioral Statistics*, *36*(5), 672-694. DOI: 10.3102/1076998611410213

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 64*, 208-232. DOI: 10.1348/000711010X500483

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*(7), 519-537. DOI: 10.1177/0146621608329504

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A standardized generalized dimensionality discrepancy measure and a standardized model-based covariance for dimensionality assessment for multidimensional models. *Journal of Educational Measurement, 52*, 144-158. DOI: 10.1111/jedm.12070

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Reckase, M. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag. DOI: 10.1007/978-1-4757-2691-6

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer-Verlag. DOI: 10.1007/978-0-387-89976-3

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

Sheng, Y. (2008a). Markov chain Monte Carlo estimation of normal ogive IRT models in MATLAB. *Journal of Statistical Software, 25*(8), 1-15. DOI: 10.18637/jss.v025.i08

Sheng, Y. (2008b). A MATLAB package for Markov chain Monte Carlo with a multi-unidimensional IRT model. *Journal of Statistical Software, 28*(10), 1-20. DOI: 10.18637/jss.v028.i10

Sheng, Y. (2010). Bayesian estimation of MIRT models with general and specific latent traits in MATLAB. *Journal of Statistical Software*, *34*(10), 1-27. DOI: 10.18637/jss.v034.i03

Sheng, Y., & Wikle, C. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67*(6), 899-919. DOI: 10.1177/0013164406296977

Sheng, Y., & Wikle, C. (2008). Bayesian multidimensional IRT models with an hierarchical structure. *Educational and Psychological Measurement, 68*(3), 413-430. DOI: 10.1177/0013164407308512

Sheng. Y., & Wikle. C. (2009). Bayesian IRT models incorporating general and specific abilities. *Behaviormetrika*, *36*(1), 27-48. DOI: 10.2333/bhmk.36.27

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375-394. DOI: 10.1111/j.1745-3984.2005.00021.x

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429-449. DOI: 10.1348/000711005X66888

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321. DOI: 10.1177/0146621605285517

van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag. DOI: 10.1007/978-1-4757-2691-6

Wang, W-C., Chen, P-H., & Cheng Y-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116-136. DOI: 10.1037/1082-989X.9.1.116

Wu, H., Yuen, K.V., & Leung, S.O. (2014). A novel relative entropy-posterior predictive model checking approach with limited information statistics for latent trait models in sparse 2^k contingency tables. *Computational Statistics and Data Analysis, 79,* 261-276. DOI: 10.1016/j.csda.2014.06.004

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213. DOI: 10.1111/j.1745-3984.1993.tb00423.x

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129-152. DOI: doi.org/10.1007/BF02294532

Zhu, X., & Stone, C.A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement, 48*, 81-97. DOI: 10.1111/j.1745-3984.2011.00132.x

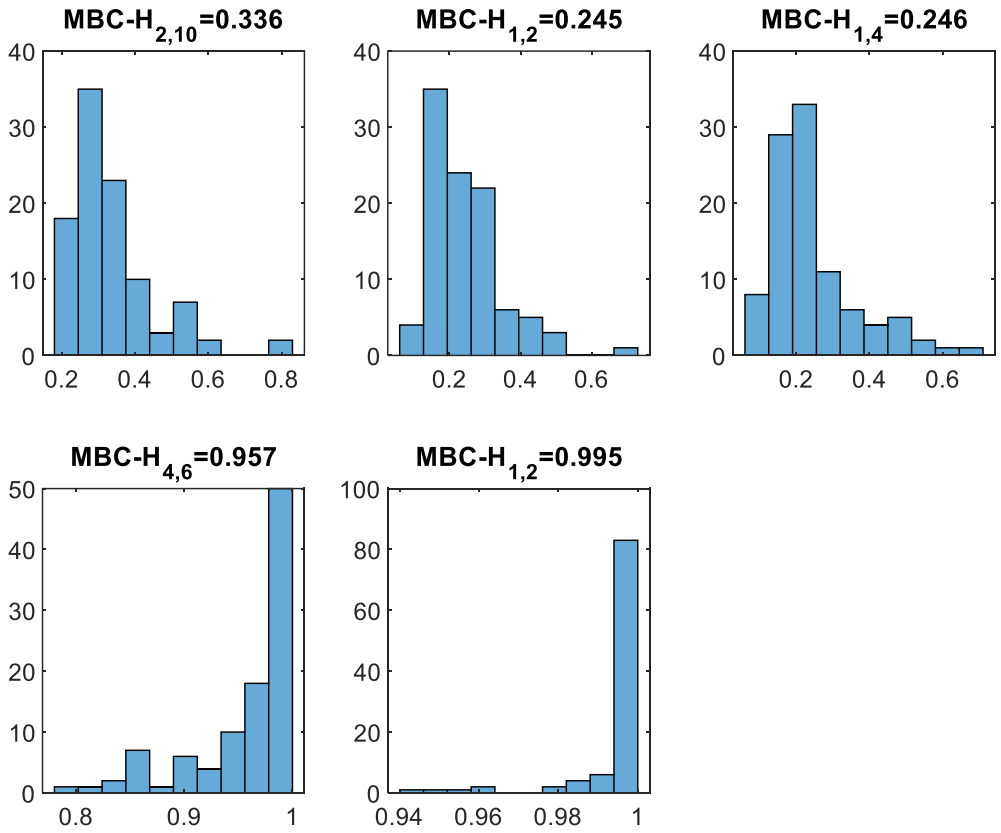

Figure 1.

Histograms of the MBC-H values on the 100 replications under different simulation conditions. Upper figures represent the cases in which the data-generating model is also the data-analysis model. In particular, from left to right, unidimensional data (item pair 2,10), multi-unidimensional data with ρ=0.3 (item pair 1,2), additive data with ρ=0.6 (item pair 1,4). Lower figures represent the cases in which multidimensional data are analyzed with a unidimensional model. In particular, from left to right, multi-unidimensional data with ρ=0.6 (item pair 4,6) and additive data with ρ=0.3 (item pair 1,2).

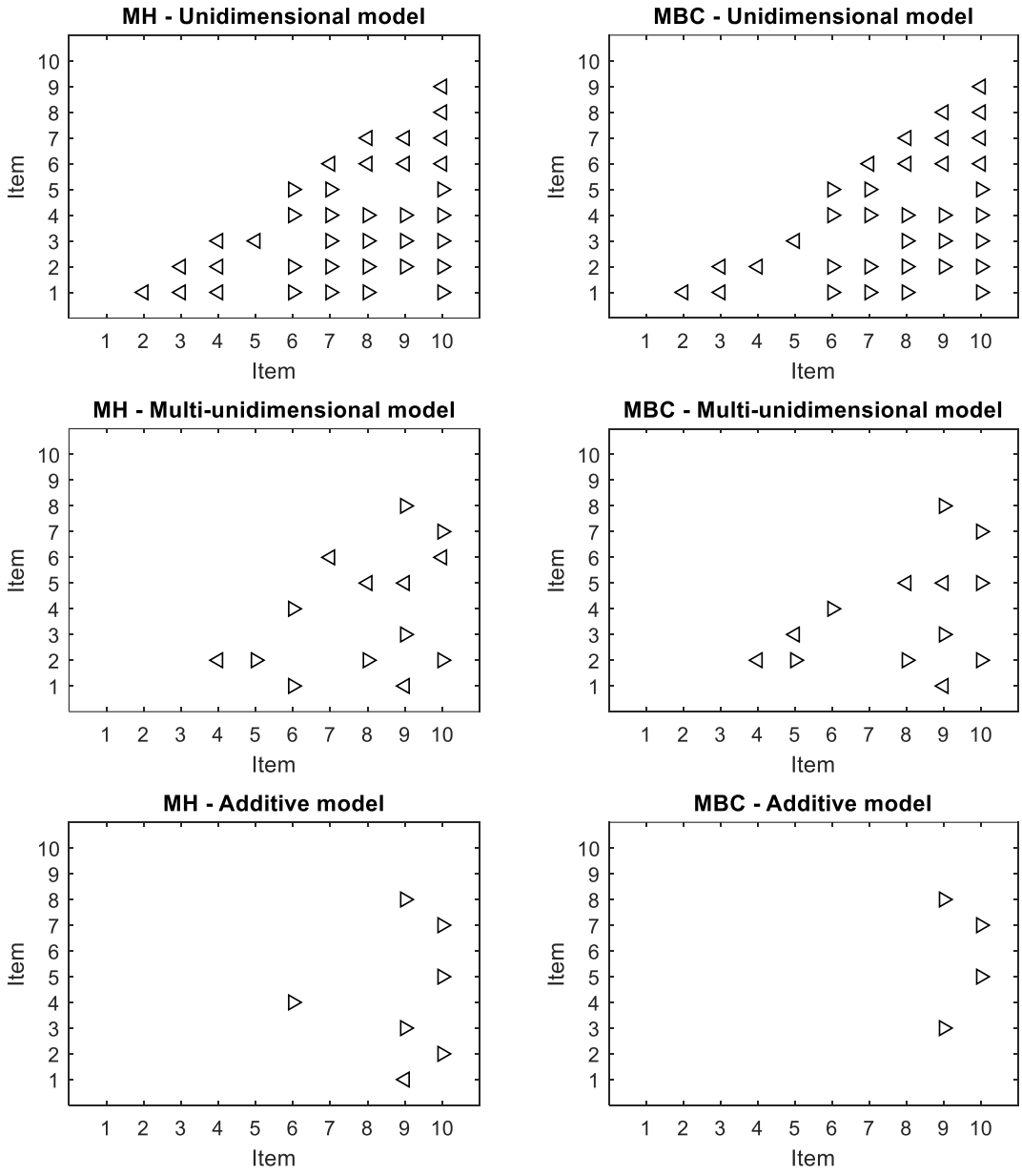

Figure 2.

Graphical representation of extreme PPP-values for MH (on the left) and MBC (on the right) for item pairs for different models (data on tourism perceptions). Right triangles indicate PPP-values greater than 0.95; left triangles indicate PPP-values lower than 0.05.

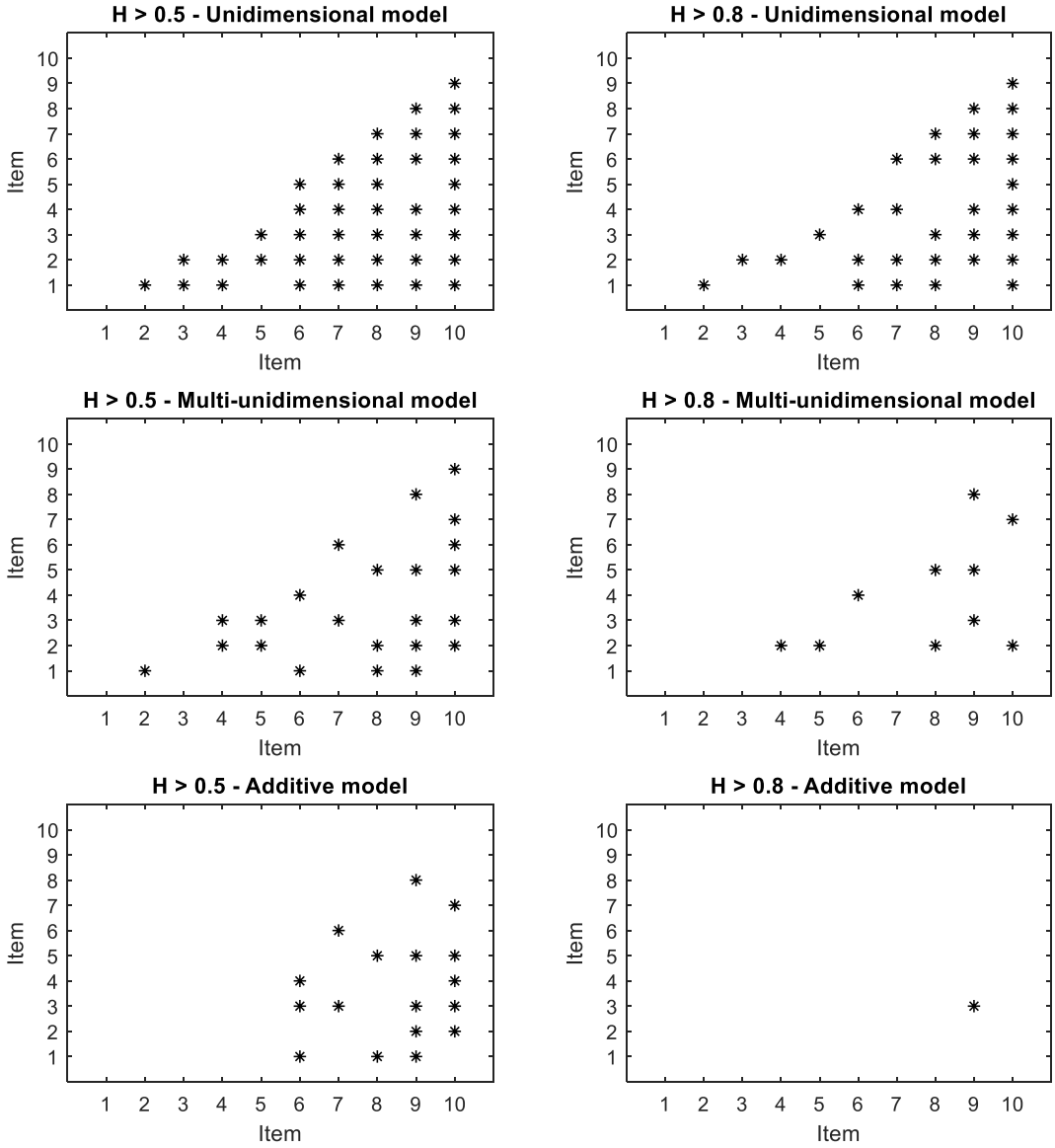

Figure 3.

Graphical representation of item pairs with a value of the MBC-H higher than 0.5 (on the left) and higher than 0.8 (on the right) for different models (data on tourism perceptions).

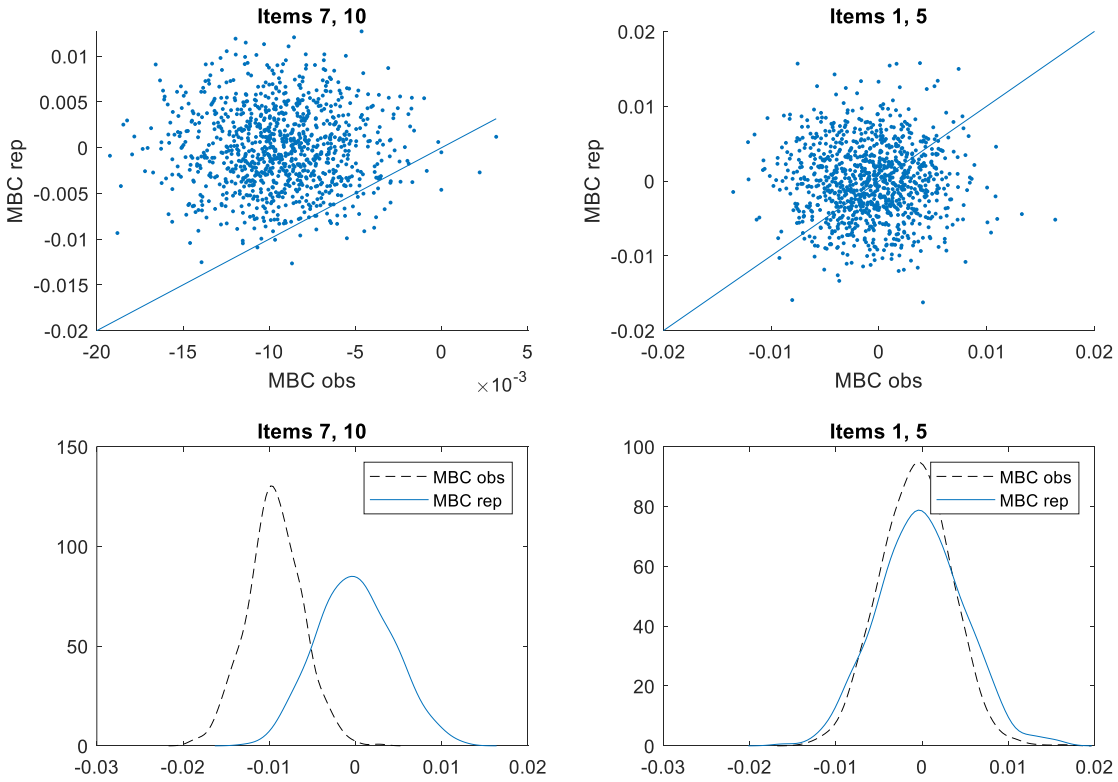

Figure 4.

Scatterplots and kernel densities of the realized (MBC obs) and predictive (MBC rep) discrepancies of MBC for the item pair 7,10 (on the left) and the item pair 1,5 (on the right).

Table 1.

Descriptive statistics for the MBC-RE and values of the GDDM-RE under different simulation conditions when the generating model and the estimated model are the same.

	ρ	MBC-RE							Median MBC-RE			GDDM-RE
		Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Within	Within2	Between	
Unidimensional		0.697	0.216	0.546	0.647	0.857	0.349	1.211	-	-	-	0.512
Multi-unidimensional	0.0	0.388	0.133	0.316	0.398	0.462	0.110	0.778	0.344	0.408	0.409	0.172
	0.3	0.392	0.141	0.289	0.391	0.474	0.158	0.794	0.271	0.274	0.406	0.182
	0.6	0.466	0.119	0.357	0.484	0.549	0.253	0.746	0.403	0.350	0.531	0.331
	0.9	0.799	0.146	0.689	0.800	0.874	0.465	1.179	0.820	0.881	0.710	6.437
Additive	0.0	0.420	0.117	0.351	0.393	0.466	0.191	0.830	0.359	0.362	0.448	0.355
	0.3	0.379	0.078	0.328	0.375	0.449	0.179	0.507	0.411	0.373	0.378	0.241
	0.6	0.434	0.113	0.348	0.434	0.482	0.252	0.805	0.430	0.358	0.448	0.256
	0.9	0.512	0.102	0.446	0.499	0.555	0.292	0.757	0.489	0.447	0.511	0.319

Note. Sd is the standard deviation, Q_1, Q_2, and Q_3 are the first, second, and third quartiles, respectively.

Table 2.
Descriptive statistics for the MBC-H and values of the GDDM-H under different simulation conditions when the generating model and the estimated model are the same.

	ρ	MBC-H								Median MBC-H			GDDM-H
		Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Prop ≥0.5	Within 1	Within 2	Between	
Unidimensional		0.450	0.082	0.397	0.435	0.524	0.295	0.619	0.311	-	-	-	0.376
Multi-unidimensional	0.0	0.321	0.072	0.287	0.330	0.366	0.158	0.487	0.000	0.311	0.338	0.332	0.203
	0.3	0.322	0.073	0.270	0.326	0.359	0.186	0.508	0.022	0.263	0.257	0.326	0.215
	0.6	0.353	0.052	0.306	0.356	0.396	0.253	0.440	0.000	0.319	0.302	0.390	0.283
	0.9	0.457	0.044	0.429	0.454	0.488	0.346	0.547	0.133	0.465	0.482	0.450	0.519
Additive	0.0	0.327	0.045	0.303	0.326	0.352	0.215	0.441	0.000	0.294	0.302	0.339	0.292
	0.3	0.319	0.042	0.289	0.320	0.352	0.207	0.388	0.000	0.344	0.300	0.320	0.248
	0.6	0.341	0.053	0.299	0.343	0.369	0.247	0.508	0.022	0.340	0.305	0.345	0.252
	0.9	0.378	0.049	0.346	0.368	0.403	0.276	0.502	0.022	0.359	0.352	0.373	0.267

Note. Prop ≥0.5 is the proportion of item pairs with MBC-H equal or higher than 0.5.

Table 3.

Proportion of extreme PPP-values for MH and MBC and PPP-value of GDDM under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model	ρ	Proportion of extreme PPP-values								GDDM PPP-value
		MH	MBC	MH within1	MH within2	MH between	MBC within1	MBC within2	MBC between	
Multi-unidimensional	0.0	0.800	0.222	1.000	0.100	1.000	1.000	0.000	0.000	0.000
	0.3	1.000	0.689	1.000	1.000	1.000	0.200	0.400	1.000	0.000
	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
	0.9	0.022	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.009
Additive	0.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
	0.6	0.689	0.644	0.800	0.600	0.680	0.800	0.600	0.600	0.000
	0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030

Table 4.

Descriptive statistics for the MBC-RE and values of the GDDM-RE under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model		MBC-RE							Median MBC-RE			GDDM-RE
	ρ	Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Within 1	Within 2	Between	
Multi-unidimensional	0.0	Inf	NaN	0.598	0.723	0.844	0.099	Inf	212.518	0.405	0.715	Inf
	0.3	Inf	NaN	9.263	12.948	Inf	4.279	Inf	Inf	Inf	9.346	Inf
	0.6	17.108	15.254	7.625	11.622	17.257	3.932	67.999	41.989	9.692	10.811	Inf
	0.9	1.252	0.426	1.014	1.180	1.400	0.717	3.445	1.073	1.426	1.118	27.548
Additive	0.0	Inf	NaN	41.756	87.047	Inf	13.176	Inf	Inf	Inf	54.548	Inf
	0.3	Inf	NaN	29.124	37.307	61.076	14.628	Inf	67.601	65.666	32.364	Inf
	0.6	5.928	5.303	2.610	4.573	6.659	0.827	24.931	4.599	6.051	4.153	Inf
	0.9	1.118	0.317	0.954	1.050	1.158	0.505	2.098	1.424	1.072	0.995	11.488

Table 5.

Descriptive statistics for the MBC-H and values of the GDDM-H under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model		MBC-H								Median MBC-H			GDDM-H
	ρ	Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Prop ≥0.5	Within 1	Within 2	Between	
Multi-unidimensional	0.0	0.540	0.261	0.415	0.454	0.485	0.150	1.000	0.222	1.000	0.339	0.451	1.000
	0.3	0.855	0.057	0.795	0.866	0.907	0.739	0.933	1.000	0.794	0.792	0.905	1.000
	0.6	0.927	0.055	0.894	0.944	0.962	0.797	0.998	1.000	0.988	0.869	0.943	1.000
	0.9	0.557	0.050	0.533	0.553	0.586	0.428	0.720	0.933	0.553	0.602	0.541	0.953
Additive	0.0	0.983	0.028	0.985	0.995	0.999	0.888	1.000	1.000	0.999	0.937	0.997	1.000
	0.3	0.989	0.011	0.986	0.992	0.996	0.941	1.000	1.000	0.993	0.990	0.992	1.000
	0.6	0.806	0.120	0.728	0.842	0.900	0.473	0.978	0.978	0.837	0.844	0.842	1.000
	0.9	0.535	0.050	0.505	0.536	0.552	0.365	0.649	0.822	0.593	0.536	0.524	0.884

Table 6.

Proportion of extreme PPP-values for MH and MBC and PPP-value of GDDM for the unidimensional, multi-unidimensional and additive models (data on tourism perceptions).

	Proportion of extreme PPP-values								GDDM PPP-value
	MH	MBC	MH within1	MH within2	MH between	MBC within1	MBC within2	MBC between	
Unidimensional	0.822	0.778	0.700	0.900	0.840	0.500	1.000	0.800	0.000
Multi-unidimensional	0.311	0.289	0.200	0.400	0.320	0.300	0.200	0.320	0.000
Additive	0.156	0.089	0.000	0.200	0.200	0.000	0.200	0.080	0.000

Appendix A

Simulation results for n=2000, k=10, $k_1=k_2=5$.

Table A1.

	ρ	MBC-RE							Median MBC-RE			GDDM-RE
		Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Within	Within2	Between	
Unidimensional		0.691	0.231	0.544	0.639	0.835	0.303	1.376	-	-	-	0.411
Multi-unidimensional	0.0	0.386	0.152	0.265	0.385	0.479	0.150	0.801	0.282	0.255	0.396	0.187
	0.3	0.415	0.135	0.324	0.400	0.511	0.165	0.796	0.343	0.358	0.454	0.273
	0.6	0.468	0.155	0.358	0.466	0.571	0.154	0.830	0.374	0.469	0.497	0.242
	0.9	1.151	0.803	0.863	1.023	1.132	0.665	6.142	0.956	1.196	0.988	58.819
Additive	0.0	0.367	0.121	0.288	0.343	0.427	0.127	0.788	0.335	0.308	0.382	0.218
	0.3	0.400	0.096	0.346	0.414	0.460	0.173	0.627	0.264	0.360	0.439	0.374
	0.6	0.416	0.128	0.335	0.406	0.503	0.197	0.880	0.300	0.340	0.444	0.283
	0.9	0.487	0.088	0.422	0.478	0.535	0.306	0.723	0.422	0.428	0.504	0.325

Descriptive statistics for the MBC-RE and values of the GDDM-RE under different simulation conditions when the generating model and the estimated model are the same.

Note. Sd is the standard deviation, Q_1, Q_2, and Q_3 are the first, second, and third quartiles, respectively.

Table A2.
Descriptive statistics for the MBC-H and values of the GDDM-H under different simulation conditions when the generating model and the estimated model are the same.

	ρ	MBC-H								Median MBC-H			GDDM-H
		Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Prop ≥0.5	Within 1	Within 2	Between	
Unidimensional		0.447	0.083	0.384	0.430	0.508	0.283	0.643	0.289	-	-	-	0.339
Multi-unidimensional	0.0	0.315	0.077	0.256	0.313	0.366	0.180	0.499	0.000	0.267	0.253	0.336	0.218
	0.3	0.329	0.063	0.294	0.323	0.374	0.197	0.488	0.000	0.295	0.306	0.351	0.256
	0.6	0.353	0.071	0.314	0.361	0.402	0.187	0.485	0.000	0.315	0.356	0.372	0.255
	0.9	0.488	0.052	0.452	0.479	0.534	0.400	0.574	0.422	0.471	0.533	0.477	0.571
Additive	0.0	0.308	0.061	0.272	0.300	0.338	0.168	0.501	0.022	0.293	0.289	0.310	0.227
	0.3	0.322	0.046	0.293	0.333	0.356	0.207	0.398	0.000	0.253	0.295	0.342	0.298
	0.6	0.333	0.058	0.300	0.336	0.364	0.215	0.523	0.022	0.279	0.308	0.351	0.257
	0.9	0.365	0.036	0.340	0.366	0.388	0.291	0.457	0.000	0.339	0.327	0.370	0.285

Note. Prop ≥0.5 is the proportion of item pairs with MBC-H equal or higher than 0.5.

Table A3.

Proportion of extreme PPP-values for MH and MBC and PPP-value of GDDM under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model	ρ	Proportion of extreme PPP-values								GDDM PPP-value
		MH	MBC	MH within1	MH within2	MH between	MBC within1	MBC within2	MBC between	
Multi-unidimensional	0.0	0.844	0.000	0.300	1.000	1.000	0.000	0.000	0.000	0.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
	0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
	0.9	0.067	0.044	0.100	0.200	0.000	0.000	0.200	0.000	0.000
Additive	0.0	1.000	0.911	1.000	1.000	1.000	0.600	1.000	1.000	0.000
	0.3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
	0.6	0.956	0.778	0.900	1.000	0.960	0.900	0.800	0.720	0.000
	0.9	0.022	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.002

Table A4.

Descriptive statistics for the MBC-RE and values of the GDDM-RE under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model		MBC-RE							Median MBC-RE			GDDM-RE
	ρ	Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Within 1	Within 2	Between	
Multi-unidimensional	0.0	Inf	NaN	0.820	0.956	Inf	0.643	Inf	Inf	Inf	0.833	Inf
	0.3	Inf	NaN	41.781	66.299	Inf	27.986	Inf	Inf	Inf	42.173	Inf
	0.6	Inf	NaN	14.950	39.857	63.384	3.828	Inf	46.939	36.273	19.279	Inf
	0.9	2.034	2.251	1.360	1.621	1.939	1.078	16.220	1.527	2.053	1.463	Inf
Additive	0.0	Inf	NaN	13.109	32.261	104.893	2.495	Inf	6.362	Inf	32.261	Inf
	0.3	Inf	NaN	78.908	109.278	151.209	3.742	Inf	Inf	87.900	99.684	Inf
	0.6	13.219	18.471	4.885	7.397	10.789	2.264	104.224	22.126	6.084	7.037	Inf
	0.9	1.394	0.417	1.106	1.352	1.631	0.711	2.928	1.467	1.505	1.162	41.337

Table A5.

Descriptive statistics for the MBC-H and values of the GDDM-H under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model		MBC-H								Median MBC-H			GDDM-H
	ρ	Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Prop ≥0.5	Within 1	Within 2	Between	
Multi-unidimensional	0.0	0.550	0.178	0.430	0.490	0.551	0.359	0.909	0.378	0.411	0.860	0.460	1.000
	0.3	0.951	0.053	0.913	0.979	0.990	0.810	0.996	1.000	0.876	0.928	0.990	1.000
	0.6	0.972	0.036	0.959	0.986	0.998	0.844	1.000	1.000	0.983	0.987	0.982	1.000
	0.9	0.626	0.077	0.579	0.620	0.654	0.516	0.945	1.000	0.589	0.663	0.616	1.000
Additive	0.0	0.949	0.087	0.947	0.989	1.000	0.679	1.000	1.000	0.833	1.000	0.989	1.000
	0.3	0.973	0.060	0.998	1.000	1.000	0.794	1.000	1.000	1.000	0.997	1.000	1.000
	0.6	0.885	0.080	0.834	0.904	0.941	0.716	1.000	1.000	0.973	0.851	0.900	1.000
	0.9	0.578	0.058	0.537	0.575	0.624	0.445	0.738	0.911	0.571	0.614	0.558	0.984

Appendix B

Simulation results for n=1000, k=20, k_1=k_2=10.

Table B1.

	ρ	MBC-RE							Median MBC-RE			GDDM-RE
		Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Within	Within2	Between	
Unidimensional		0.998	0.188	0.856	1.011	1.152	0.528	1.412	-	-	-	0.768
Multi-unidimensional	0.0	0.697	0.144	0.605	0.674	0.764	0.425	1.269	0.643	0.725	0.680	0.547
	0.3	0.719	0.133	0.632	0.705	0.802	0.421	1.124	0.684	0.694	0.716	0.513
	0.6	0.725	0.166	0.603	0.721	0.837	0.326	1.241	0.663	0.743	0.723	0.426
	0.9	0.885	0.236	0.709	0.859	1.027	0.395	1.848	0.821	0.909	0.856	7.996
Additive	0.0	0.585	0.130	0.511	0.567	0.636	0.311	1.064	0.583	0.550	0.565	0.356
	0.3	0.663	0.124	0.582	0.656	0.729	0.377	1.142	0.663	0.653	0.655	0.494
	0.6	0.737	0.121	0.641	0.736	0.809	0.491	1.077	0.692	0.741	0.740	0.553
	0.9	0.789	0.131	0.697	0.777	0.880	0.471	1.145	0.727	0.768	0.787	0.614

Descriptive statistics for the MBC-RE and values of the GDDM-RE under different simulation conditions when the generating model and the estimated model are the same.

Note. Sd is the standard deviation, Q_1, Q_2, and Q_3 are the first, second, and third quartiles, respectively.

Table B2.
Descriptive statistics for the MBC-H and values of the GDDM-H under different simulation conditions when the generating model and the estimated model are the same.

	ρ	MBC-H								Median MBC-H			GDDM-H
		Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Prop ≥0.5	Within 1	Within 2	Between	
Unidimensional		0.552	0.057	0.515	0.562	0.598	0.400	0.653	0.847	-	-	-	0.479
Multi-unidimensional	0.0	0.453	0.051	0.421	0.446	0.477	0.343	0.626	0.147	0.436	0.461	0.447	0.394
	0.3	0.461	0.048	0.432	0.459	0.492	0.349	0.595	0.211	0.440	0.455	0.463	0.379
	0.6	0.463	0.060	0.424	0.463	0.505	0.282	0.620	0.279	0.450	0.468	0.464	0.354
	0.9	0.512	0.063	0.471	0.511	0.563	0.347	0.652	0.568	0.485	0.530	0.512	0.359
Additive	0.0	0.407	0.052	0.377	0.404	0.431	0.283	0.584	0.047	0.413	0.399	0.405	0.306
	0.3	0.437	0.045	0.412	0.435	0.462	0.313	0.584	0.079	0.437	0.422	0.435	0.368
	0.6	0.465	0.041	0.432	0.465	0.492	0.378	0.575	0.184	0.451	0.460	0.469	0.397
	0.9	0.485	0.043	0.455	0.483	0.517	0.371	0.597	0.342	0.474	0.482	0.486	0.409

Note. Prop ≥0.5 is the proportion of item pairs with MBC-H equal or higher than 0.5.

Table B3.

Proportion of extreme PPP-values for MH and MBC and PPP-value of GDDM under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model	ρ	Proportion of extreme PPP-values								GDDM PPP-value
		MH	MBC	MH within1	MH within2	MH between	MBC within1	MBC within2	MBC between	
Multi-unidimensional	0.0	0.968	0.000	0.867	1.000	1.000	0.000	0.000	0.000	0.000
	0.3	1.000	0.758	1.000	1.000	1.000	1.000	0.022	0.980	0.000
	0.6	0.853	0.816	0.844	0.911	0.830	0.844	0.756	0.830	0.000
	0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Additive	0.0	1.000	0.974	1.000	1.000	1.000	0.911	1.000	0.990	0.000
	0.3	0.979	0.963	1.000	1.000	0.960	0.911	1.000	0.970	0.000
	0.6	0.689	0.663	0.622	0.689	0.720	0.778	0.444	0.710	0.000
	0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001

Table B4.

Descriptive statistics for the MBC-RE and values of the GDDM-RE under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model		MBC-RE							Median MBC-RE			GDDM-RE
	ρ	Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Within1	Within2	Between	
Multi-unidimensional	0.0	Inf	NaN	1.158	1.374	Inf	0.958	Inf	Inf	Inf	1.167	Inf
	0.3	Inf	NaN	5.167	12.241	27.790	1.234	Inf	228.158	2.816	11.972	Inf
	0.6	11.739	15.137	4.978	7.273	12.079	1.778	124.599	12.333	7.188	6.663	Inf
	0.9	1.485	0.371	1.285	1.421	1.567	1.071	3.690	1.296	1.595	1.418	Inf
Additive	0.0	Inf	NaN	18.331	39.912	80.801	2.758	Inf	23.443	109.342	38.048	Inf
	0.3	Inf	NaN	14.588	33.294	56.490	2.155	Inf	18.207	60.132	33.294	Inf
	0.6	7.927	8.630	3.530	5.603	9.379	1.496	74.654	7.119	3.422	5.603	Inf
	0.9	1.388	0.268	1.230	1.349	1.508	0.800	2.907	1.445	1.266	1.359	78.615

Table B5.

Descriptive statistics for the MBC-H and values of the GDDM-H under different simulation conditions when the generating model is multidimensional (multi-unidimensional or additive) and the estimated model is unidimensional.

Generating model		MBC-H								Median MBC-H			GDDM-H
	ρ	Mean	Sd	Q_1	Q_2	Q_3	Min	Max	Prop ≥0.5	Within 1	Within 2	Between	
Multi-unidimensional	0.0	0.642	0.143	0.552	0.573	0.662	0.478	0.920	0.974	0.549	0.890	0.564	1.000
	0.3	0.893	0.124	0.862	0.944	0.984	0.597	1.000	1.000	1.000	0.679	0.944	1.000
	0.6	0.900	0.066	0.860	0.908	0.953	0.724	1.000	1.000	0.956	0.874	0.909	1.000
	0.9	0.628	0.038	0.601	0.625	0.646	0.555	0.784	1.000	0.605	0.657	0.626	1.000
Additive	0.0	0.974	0.044	0.965	0.995	0.999	0.782	1.000	1.000	0.956	0.999	0.995	1.000
	0.3	0.969	0.052	0.967	0.994	0.998	0.709	1.000	1.000	0.967	0.998	0.994	1.000
	0.6	0.866	0.087	0.806	0.885	0.934	0.628	0.999	1.000	0.899	0.798	0.890	1.000
	0.9	0.618	0.039	0.596	0.617	0.641	0.486	0.755	0.989	0.633	0.605	0.618	0.993

Appendix C

Generating and estimated model parameters for multi-unidimensional data ($\rho=0$) analyzed with the unidimensional model in the simulation study with n=1000, k=10.

Table C1.

Generating item discrimination (α) and difficulty (δ) parameters for the multi-unidimensional data with $\rho=0$.

Item	Subtest	α_v	δ
1	v=1	1.044	-1.116
2		1.249	-1.320
3		1.531	1.613
4		1.988	1.526
5		1.133	-1.475
6	v=2	1.989	1.056
7		1.758	-1.989
8		1.730	1.881
9		1.636	1.115
10		1.561	-1.427

Table C2.

Estimated item discrimination (α) and difficulty (δ) parameters with the unidimensional model when data are generated from a multi-unidimensional model with $\rho=0$.

Item	α	Sd(α)	δ	Sd(δ)
1	0.049	0.034	-0.819	0.045
2	0.022	0.020	-0.908	0.046
3	0.054	0.038	0.869	0.046
4	0.048	0.034	0.659	0.042
5	0.105	0.050	-1.020	0.048
6	1.902	0.196	0.937	0.107
7	2.332	0.328	-2.588	0.295
8	1.733	0.204	1.960	0.194
9	1.871	0.208	1.105	0.128
10	1.142	0.111	-1.243	0.089