PSYCHOMETRIKA—VOL. 88, NO. 2, 387–412
JUNE 2023
https://doi.org/10.1007/s11336-023-09905-w

A TEST TO DISTINGUISH MONOTONE HOMOGENEITY FROM MONOTONE MULTIFACTOR MODELS

JULES L. ELLIS

BEHAVIOURAL SCIENCE INSTITUTE, RADBOUD UNIVERSITY NIJMEGEN

KLAAS SIJTSMA

TILBURG UNIVERSITY

The goodness-of-fit of the *unidimensional* monotone latent variable model can be assessed using the empirical conditions of nonnegative correlations (Mokken in A theory and procedure of scale-analysis, Mouton, The Hague, 1971), manifest monotonicity (Junker in Ann Stat 21:1359–1378, 1993), multivariate total positivity of order 2 (Bartolucci and Forcina in Ann Stat 28:1206–1218, 2000), and nonnegative partial correlations (Ellis in Psychometrika 79:303–316, 2014). We show that *multidimensional* monotone factor models with independent factors also imply these empirical conditions; therefore, the conditions are insensitive to multidimensionality. Conditional association (Rosenbaum in Psychometrika 49(3):425–435, 1984) can detect multidimensionality, but tests of it (De Gooijer and Yuan in Comput Stat Data Anal 55:34–44, 2011) are usually not feasible for realistic numbers of items. The only existing feasible test procedures that can reveal multidimensionality are Rosenbaum's (Psychometrika 49(3):425–435, 1984) Case 2 and Case 5, which test the covariance of two items or two subtests conditionally on the *unweighted* sum of the other items. We improve this procedure by conditioning on a *weighted* sum of the other items. The weights are estimated in a training sample from a linear regression analysis. Simulations show that the Type I error rate is under control and that, for large samples, the power is higher if one dimension is more important than the other or if there is a third dimension. In small samples and with two equally important dimensions, using the unweighted sum yields greater power.

Key words: unidimensional measurement, multidimensional measurement, monotone latent variable model, monotone homogeneity model, conditional association.

For binary test data satisfying a monotone item response theory (IRT) model, we develop a statistical test procedure that can detect multidimensionality as opposed to unidimensionality. Investigating the dimensionality of a psychological test is an important step in test development and validation. Establishing unidimensionality can contribute to construct validity of the test because this renders the interpretation of test performance easier, comparable with measurement in other science areas. Multidimensional item sets are edited by removing or replacing items deviating from the target attribute or splitting the item set in subsets representing better interpretable test performance. A case in point is the development of Spearman's (1904) theory of general intelligence into the current multidimensional Cattell–Horn–Carroll (CHC) theory (Wasserman, 2019), based on psychometric analyses of numerous datasets.

Dimensionality analysis of an item set is usually done using factor analysis or IRT analysis (e.g., Sijtsma & Van der Ark, 2021). These approaches include parametric assumptions such as linearity and normality in factor analysis and logistic, normal-ogive, or step functions in IRT. These assumptions usually have little prior plausibility, which led several authors (Holland, 1981; Mokken, 1971; Rosenbaum, 1984; Stout, 1987) to study measurement using weaker assumptions,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-023-09905-w.

Correspondence should be made to Jules L. Ellis, Behavioural Science Institute, Radboud University Nijmegen, P.O.B. 9104, 6500 HE Nijmegen, The Netherlands. Email: jules.ellis@ru.nl

© 2023 The Author(s)

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

for example, replacing logistic and normal-ogive item response functions (IRFs) with monotone IRFs only subjected to order restrictions without choosing a parametric function.

The absence of restrictive parametric functions rendered the development of goodness-of-fit tests complex but was replaced with a focus on testable conditions that are the hallmark of an underlying quantitative variable. An example of a testable condition is that the inter-item correlations must be nonnegative (Mokken, 1971). The search for testable properties was also inspired by axiomatic measurement theory (Krantz et al., 1971) and probabilistic developments in it, such as the relation between simple scalability and strong stochastic transitivity in choice data (Tversky & Russo, 1969). This article also follows this approach.

We review two classes of monotone nonparametric IRT models and their testable conditions. Critical to this article, we argue that most conditions for which practical test procedures are available cannot distinguish multidimensional from unidimensional monotone IRT models. We target a specific set of covariance inequalities and demonstrate that we can use them to detect multidimensionality in cases that would previously remain undetected. We develop a practical test procedure and explore the Type I error rate and power using simulated data.

1. Models and Testable Conditions

1.1. Monotone Homogeneity and Monotone Factor Models

We discuss the definitions of three nonparametric IRT models that will be used throughout the article. The first model is monotone homogeneity (MH), which contends that the expected value of each observed binary item score variable increases with a single underlying variable, called the common factor, the latent variable, or the latent dimension. The second model, the monotone factor model (MFM), contends basically the same as the MH model with one or more independent factors to which the items are related in a simple structure. The third model is the higher-order monotone one-factor (HOMOF) model, which is like the MFM, but allows the factors to be correlated with the restriction that they depend on a single higher-order factor. The three models share the assumption of conditional (or local) independence or independent errors, which are similar assumptions. Thus, MH describes a general form of unidimensionality, MFM describes a general form of multidimensionality, and HOMOF is somewhere in between. Applied to intelligence, MH formally resembles Spearman's theory of a single general intelligence factor, MFM is like Thurstone's initial theory of multiple independent primary mental abilities, and HOMOF parallels the hierarchical factors of CHC, integrating the other two theories.

We assume the item scores are binary manifest variables, denoted $\mathbf{X} = (X_1, \ldots, X_J)$. Variable X_i represents the scores (1 = positive, 0 = negative) subjects obtained on the i-th item. Factors, latent variables, or dimensions are denoted $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_D)$. The rest of this section discusses the formal definitions of the three models, which we need to prove the theorems.

We adopt the following assumptions (Holland & Rosenbaum, 1986; Mokken, 1971; Rosenbaum, 1984):

MH1 (conditional independence): $\mathbf{X}$ is conditionally independent given $\boldsymbol{\Theta}$.
MH2 (monotonicity): $P(X_i = 1|\, \boldsymbol{\Theta})$ is an increasing function of $\boldsymbol{\Theta}$ for all $i = 1, \ldots, J$.
MH3 (unidimensionality): $D = 1$.

We use the term 'increasing' synonymous with 'monotone nondecreasing'. For readers who do not have the information ready, Appendix A provides formulations of the assumptions having greater precision. Following Holland and Rosenbaum (1986), we will say that ($\mathbf{X}$, $\boldsymbol{\Theta}$) is a monotone latent variable (MLV) model if MH1 and MH2 hold. Following Mokken (1971) Mokken and Lewis (1982) and Ellis and Junker (1997), we will say that $\mathbf{X}$ satisfies a *unidimensional MLV model* or *MH model* if there exists a variable Θ such MH1, MH2, and MH3 hold.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

Ellis (2015) studied a narrower class of monotone models. Slightly rephrasing Ellis, we will say that $\mathbf{X}$ satisfies an MFM if

$$\mathbf{X} = \phi\left(\psi\left(\boldsymbol{\lambda}\boldsymbol{\Theta}\right) + \boldsymbol{\varepsilon}\right),$$

where ϕ and ψ are component-wise increasing functions, $\boldsymbol{\Theta}$ is a multivariate random vector with independent components, $\boldsymbol{\varepsilon}$ is a multivariate random vector with components that are independent of each other and of $\boldsymbol{\Theta}$, the ε_is have log-concave densities (Appendix A), and $\boldsymbol{\lambda}$ is a nonnegative real matrix with simple structure (i.e., every manifest variable loads positive on one factor and zero on the other factors). As an example of an MFM for binary manifest variables, consider a case where the ε_is have standard normal distributions, ψ is the identity function, and the ϕ_is are step functions with $\phi_i\left(x\right) = 0$ if $x \leq \beta_i$ and $\phi_i\left(x\right) = 1$ if $x > \beta_i$ for some real number β_i. Then, $P\left(X_i = 1 \mid \boldsymbol{\Theta}\right) = \Phi\left(\sum_{d=1}^{D} \lambda_{id}\Theta_d - \beta_i\right)$, where $\Phi(.)$ is the standard normal distribution function. Hence, every multidimensional normal ogive IRT model with independent factors and nonnegative loadings with a simple structure is an MFM (also, see Takane & De Leeuw, 1987).

Ellis (2015) also studied a more general class of models, where the components of $\boldsymbol{\Theta}$ need not be independent but may be the result of a higher-order MFM factor with log-concave disturbances at each level. We call this class of models, with possibly many levels and one factor at the highest level, a HOMOF model. In this class of models, the factor loadings at the lowest level (i.e., $\boldsymbol{\lambda}$ in $\mathbf{X} = \phi\left(\psi\left(\boldsymbol{\lambda}\boldsymbol{\Theta}\right) + \boldsymbol{\varepsilon}\right)$) do not necessarily have simple structure.

1.2. Testable Conditions of the Models

In this section, we review statistical inequalities that have been used to test whether MH holds for a given set of manifest variables. These inequalities can be expressed as covariances that are nonnegative. The general result implied by MH is conditional association (CA; Rosenbaum, 1984). Below, we discuss that CA is the hallmark of MH. Coincidentally, CA fares well with Spearman's (Spearman, 1904) idea that intelligence tests have positive correlations and together measure a single general intelligence factor, and Guttman's 'first law of intelligence', stating that any two intelligence items have a nonnegative correlation in any population that is "not artificially selected" (Guttman & Levy, 1991), thus suggesting the items should have nonnegative correlations in any subgroup defined by the other items. CA is hard to test fully because it involves many restrictions even for small item sets (De Gooijer & Yuan, 2011; Ligtvoet, 2022; Yuan & Clarke, 2001). Therefore, we will also discuss conditions that are easier to test, such as the condition that the expected item score increases with the rest score, called *marginal monotonicity* (MM) (Junker, 1993). These conditions can be viewed as incomplete tests of CA (Ligtvoet, 2022). We will now continue this section with the formal definitions.

Following Rosenbaum (1984), we say that $\mathbf{X}$ is CA if for every partition $\mathbf{X} = \left(\mathbf{Y}, \mathbf{Z}\right)$ and every function h, for all increasing functions ϕ and ψ,

$$\mathrm{Cov}\left(\phi\left(\mathbf{Y}\right), \psi\left(\mathbf{Y}\right) | h\left(\mathbf{Z}\right)\right) \geq 0.$$

Rosenbaum (1984) provides examples. Rosembaum's result that

$$\mathrm{MH} \Rightarrow \mathrm{CA}$$

is key to this article, in which we develop a practically feasible test for CA. Holland and Rosenbaum (1986) generalized Rosenbaum's (Rosenbaum, 1984) work to non-binary variables. Ellis and

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

Junker (1997; also Junker & Ellis, 1997) furthermore suggested that CA is sufficient to detect multidimensionality in a finite number of items. They studied infinite item sequences and used the condition of vanishing conditional dependence (VCD), which means that certain conditional covariances vanish as $J \to \infty$. They showed that CA and VCD are necessary and sufficient for a unidimensional monotone latent variable model in which the latent variable can be estimated consistently. Since VCD is defined only in the limit as $J \to \infty$, one would expect that if any condition can detect multidimensionality in a finite item set, it will be CA.

In addition to the practical infeasibility of CA due to large numbers of restrictions one must test even for small item sets, a complete test of CA is also impossible because of the sparseness of the data, because most item-score patterns never occur in the most commonly available sample sizes of 100 to 10,000 subjects. Maybe for this reason, many authors have studied weaker restrictions than CA, possibly believing they still capture the gist of CA. Straat et al. (2016) acknowledged this limitation and proposed an incomplete strategy. Ligtvoet (2022) gives an excellent review of weaker conditions, which he describes as "incomplete tests of conditional association".

An important condition discussed by Ligtvoet (2022) is *multivariate total positivity of order 2* (MTP_2). The ordinary formulation of this condition is given in Appendix A, but for the present purpose it suffices to note that $\mathbf{X}$ being MTP_2 implies that

$$\text{Cov}\left(\phi\left(\mathbf{Y}\right), \psi\left(\mathbf{Y}\right)|\mathbf{Z}\right) \geq 0.$$

(Note the omission of $h()$ here). Therefore, $\mathbf{X}$ being MTP_2 means that $\text{Cov}\left(\phi\left(\mathbf{Y}\right), \psi\left(\mathbf{Y}\right)|h(\mathbf{Z})\right) \geq 0$ holds for some, *but not all*, functions h (Ellis, 2015, p. 264–265). For binary variables, the difference between CA and MTP_2 thus lies in the kind of events on which one may condition: MTP_2 involves conditioning on the finest partition of subgroups that can be made with $\mathbf{Z}$, whereas CA also involves conditioning on combinations of such groups. Holland and Rosenbaum (1986) show that

$$\text{CA} \Rightarrow \text{MTP}_2.$$

Therefore, a test of MTP_2 can be viewed as an incomplete test of CA (Ligtvoet, 2022).

For tests consisting of realistic numbers of items, however, MTP_2 still involves a large number of restrictions (Bartolucci & Forcina, 2005, p. 35; Ligtvoet, 2022). Therefore, one may want to reduce the number of restrictions to be tested even further by considering properties derived from MTP_2, which include nonnegative partial correlations (NPC; Ellis, 2015; Brusco et al., 2015) and nonnegative covariances (NNC) (Mokken, 1971).

The important testable property of Manifest Monotonicity (MM) (Junker, 1993; Junker & Sijtsma, 2000) means that each item regression on the sum of the other items (known as the rest score) is increasing (Appendix A). Junker (1993, p. 1372), Junker and Sijtsma (2000) showed that

$$\text{MH} \Rightarrow \text{MM}.$$

Ligtvoet (2022) showed that CA $\Rightarrow$ MM, so MM may be viewed as another incomplete test of CA. MM is an important property because it is conceptually like the idea of a monotone IRF, and for the reader who is unfamiliar with these concepts it may be hard to see how one can have MM without MH. Ellis (2014) gave an example where MM holds while NPC fails, and therefore, MH must fail too.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

1.3. Limitations of Partial Tests of Conditional Association

Ligtvoet (2022) concluded that existing incomplete tests of CA perform poorly detecting violations of CA. In the data structures he generated, CA was often violated while MTP_2 and MM held. Thus, a researcher who tests MTP_2 and/or MM instead of CA misses the violation of CA. We agree but notice that MTP_2 and MM are not sensitive to violations of dimensionality, necessitating the discussion of the problem from a more theoretical perspective. This discussion is partially inspired by Van den Wollenberg's (Van den Wollenberg, 1982) proof that existing test statistics for the Rasch model were insensitive to violations of unidimensionality.

Ellis (2015; also proposition A1 of Appendix A) showed that for MTP_2 the problem is that,

$$\text{MFM} \Rightarrow \text{MTP}_2.$$

Consequently, any test of MH based on MTP_2 (Bartolucci & Forcina, 2000, 2005) logically cannot distinguish MH from MFM. That is, if the test suggests that MTP_2 holds, it is also possible that an MFM that violates MH generated the data. The same conclusion holds for any property derived from MTP_2, which includes NPC (Ellis, 2015; Brusco et al., 2015), and NNC (Mokken, 1971). Thus, testing MTP_2, NPC, and NNC cannot distinguish unidimensional and multidimensional monotone factor models.

MM is implied by CA but not by MTP_2, and therefore, we need to discuss it separately. For MM, the problem is that

$$\text{MFM} \Rightarrow \text{MM}.$$

This follows from Theorem 1 in Appendix B. Consequently, no test of MM (Douglas & Cohen, 2001; Junker & Sijtsma, 2000; Molenaar & Sijtsma, 2000; Tijmstra et al., 2013; Tijmstra & Bolsinova, 2019) can distinguish MH from MFM. That is, if the test suggests that MM holds, it is also possible that an MFM that violates MH generated the data. Thus, testing MM cannot distinguish unidimensional from multidimensional monotone factor models.

To summarize, based on data it is impossible to distinguish between MH and an MFM if one tests CA only partially with MTP_2, NNC, NPC, and MM. For example, assume that $J = 10$, the first five items satisfying the Rasch model with latent variable Θ_1, and the last five satisfying the Rasch model with latent variable Θ_2, and Θ_1 and Θ_2 are independent. This is an MFM, thus satisfying MM and MTP_2, which implies NNC and NPC. Hence, it would be impossible to reject MH if only these conditions are tested. This puts the MH model at a serious disadvantage in comparison with parametric IRT models such as the two-parameter logistic model, where one can easily identify this situation via the M_2-test (Maydeu-Olivares & Joe, 2006).

The argument given includes the somewhat artificial case of independent factors, which implies that some items have correlation zero. One might argue that such cases would be excluded with Mokken's (1971; Mokken & Lewis, 1982) criterion $H > .30$. However, Ellis (2015) also showed that

$$\text{HOMOF} \Rightarrow \text{MTP}_2.$$

Although HOMOF models involve a single higher-order factor, there will generally be multiple first-order factors that are positively correlated with any degree. Consequently, it is possible that such items also satisfy $H > .30$.

We conclude that to distinguish unidimensional MFMs from multidimensional MFMs, the testable conditions MTP_2, NPC, NNC, and MM are logically insufficient. Therefore, we will target

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

specific aspects of CA, which are covariance inequalities that will likely discriminate between MH and multidimensional MFMs. The next section discusses candidate covariance inequalities.

2. Conditioning on Added Regression Predictions (CARP) Inequalities

2.1. CARP Inequalities, Definition

Assume that all response probabilities of $\mathbf{X}$ are known; sample statistics will be discussed later. For a given pair (X_i, X_j), denote by $\mathbf{X}_{-ij}$ the variables of $\mathbf{X}$ except X_i and X_j. Our proposal is a generalization of Rosenbaum's (1984, p. 427) "Case 2" method to test the covariance of each item pair conditionally on the rest score of the pair. For these covariances, CA implies that

$$\mathrm{Cov}(X_i, X_j | \sum\nolimits_{k\neq i,j}^{J} X_k) \geq 0.$$

We call this the *conditioning on rest scores* (CRS) *inequality*, and we call a significance test for the CRS inequality a *CRS test*. A limitation of the CRS inequality for testing the dimensionality of a set of items is that the rest score used for conditioning is not adapted to the possible multidimensional structure if the item set does not satisfy MH. To obtain greater adaptation, we propose to use *weighted* rest scores. We use two linear regression analyses, where X_i and X_j serve as dependent variables and the other items are independent variables. Write $X_0 := 1$, and denote the regression coefficient of X_k in the prediction of X_i from $\mathbf{X}_{-ij}$ as $a_{k.ij}$, and denote the resulting predicted scores as $\hat{X}_{ij}$; that is,

$$\hat{X}_{ij} = \sum\nolimits_{k=0}^{J} a_{k.ij} X_k,$$

where the $a_{k.ij}$ are such that they minimize $\mathbb{E}\left(\left(X_i - \hat{X}_{ij}\right)^2\right)$, and with $a_{k.ij} = 0$ if $k = i$ or $k = j$. Similarly, $\hat{X}_{ji} = \sum_{k=0}^{J} a_{k.ji} X_k$ where the $a_{k.ji}$ are such that they minimize $\mathbb{E}\left(\left(X_j - \hat{X}_{ji}\right)^2\right)$, and with $a_{k.ji} = 0$ if $k = i$ or $k = j$. In other words, $\hat{X}_{ij}$ is the prediction of X_i if X_j is excluded as predictor, and $\hat{X}_{ji}$ is the prediction of X_j if X_i is excluded as predictor. As the basis for conditioning, we propose the variable

$$\hat{X}_{ij} + \hat{X}_{ji} = \sum\nolimits_{k=0}^{J} (a_{k.ij} + a_{k.ji}) X_k.$$

This is the sum of the predicted scores of X_i and X_j, where both X_i and X_j are excluded from the predictors. It may be noted that we are not assuming linearity or normality here; we are just using the least squares solution as a heuristic tool, without claiming that this produces a good model.

The variable $\hat{X}_{ij} + \hat{X}_{ji}$ can attain many different values, producing small conditioning groups. Therefore, we will use deciles or other quantiles of this function. This is like Rosenbaum's (1984, p. 428) "Case 5", which tests the covariance of two subtests conditionally on deciles of the rest score. If Q_m is an operator that divides any variable into m groups of approximately equal size, that is, $P[Q_m(X) = k] \approx m^{-1}$, then we obtain

$$\mathrm{Cov}[X_i, X_j | Q_m\left(\hat{X}_{ij} + \hat{X}_{ji}\right)] \geq 0.$$

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

We refer to covariance inequalities of this form (with or without grouping by Q_m) as *conditioning on added regression predictions* (CARP) *inequalities*. Similarly, we refer to the involved conditional covariances as *CARP covariances*, and to the corresponding correlations as *CARP correlations*. We call the property that $\mathbf{X}$ satisfies all CARP inequalities, simply *CARP*.

CARP is the special case of CA with (if we use the notation used in the definition of CA) $\mathbf{Y} = (X_i, X_j)$, $\phi(\mathbf{Y}) = X_i$, $\psi(\mathbf{Y}) = X_j$, $\mathbf{Z} = \mathbf{X}_{-ij}$, and $h(\mathbf{Z}) = \sum_{k=0}^{J} (a_{k.ij} + a_{k.ji}) X_k$. The latter weighted sum is a function of $\mathbf{X}_{-ij}$ because X_i has weight $a_{i.ij} + a_{i.ji} = 0$ and X_j has weight $a_{j.ij} + a_{j.ji} = 0$. We assume in this section that all response probabilities of $\mathbf{X}$ are known, and therefore, the regression coefficients $a_{k.ij}$ are parameters, not sample statistics.

Hence, MH implies CARP. Furthermore, MFM does not imply CARP, which is demonstrated in later simulations, and which can also be seen in theoretical computations of some special cases. Therefore, testing CARP inequalities can reveal some violations of MH that testing MTP_2 or MM cannot reveal.

Let us now briefly explain why CARP inequalities may be useful in the assessment of multidimensionality. Suppose that X_i and X_j load on different independent latent variables, say, Θ_1 and Θ_2, and that the other items load on either Θ_1 or Θ_2. After a suitable transformation of Θ_1 and Θ_2, we may say that $\hat{X}_{ij}$ estimates Θ_1 and $\hat{X}_{ji}$ estimates Θ_2 (set $\Theta_1 := \mathbb{E}\left(\hat{X}_{ij} \mid \Theta_1\right)$ and $\Theta_2 := \mathbb{E}\left(\hat{X}_{ji} \mid \Theta_2\right)$), so conditioning on $\hat{X}_{ij} + \hat{X}_{ji}$ tends to create groups with $\Theta_1 + \Theta_2$ approximately equal, which induces a negative correlation between Θ_1 and Θ_2 in these groups (in groups where $\Theta_1 + \Theta_2$ is constant, Θ_2 is a decreasing function of Θ_1), which in turn tends to create a negative correlation between X_i and X_j. Theorem 2 of the Appendix states more formally that in this situation, the mean conditional covariance given the unweighted rest scores will be negative or zero, and Theorem 3 of the Appendix states that this will also be true for the mean conditional covariance given the weighted rest score $\hat{X}_{ij} + \hat{X}_{ji}$ provided that $\mathbb{E}\left(X_i \mid \hat{X}_{ij}\right)$ and $\mathbb{E}\left(X_j \mid \hat{X}_{ji}\right)$ are both increasing (i.e., the items have MM with respect to the partial weighted sum score of their respective subtest). In the standard two-dimensional case (defined in section 5.1) with ten items, we computed these correlations using numerical integration, and the outcomes supported our expectation that such correlations tend to be negative or zero. The simulations to assess the power of the CARP tests, reported later in this study, also support this result.

When we developed the test, we initially created a slightly different method, which can produce smaller correlations than the CARP correlations (the computation of the next example can be found in the Supplementary Material). For example, take two uncorrelated standard normal dimensions each with five Rasch items, all having $\beta_i = 0$. Then, using numerical integration, one can obtain a correlation of -.204 in the union of the two most extreme vigintile groups of $\hat{X}_{ij} - \hat{X}_{ji}$; that is, $\varphi\left(X_i, X_j | [Q_{20}\left(\hat{X}_{ij} - \hat{X}_{ji}\right) = 1] \cup [Q_{20}\left(\hat{X}_{ij} - \hat{X}_{ji}\right) = 20]\right) = -.204$. This is not aCARP correlation, because we condition on $\hat{X}_{ij} - \hat{X}_{ji}$ instead of $\hat{X}_{ij} + \hat{X}_{ji}$. Although this correlation is smaller than the CARP correlations we obtained, a statistical test based on this conditional correlation of $-.204$ turns out to be less powerful because the 90% observations with $2 \leq Q_{20}\left(\hat{X}_{ij} - \hat{X}_{ji}\right) \leq 19$ are discarded. Simulations showed that a CARP test has greater power in this case. The next section focusses on a test statistic that can be used to test CARP.

Our approach is almost the opposite of the DETECT and DIMTEST procedures for investigating an item set's dimensionality (Stout et al., 1996; Zhang & Stout, 1999a, 1999b). DETECT and DIMTEST look for large conditional covariances, averaged over item pairs, as a sign that unidimensionality is violated. Unlike CARP, these approaches do not use rigorously established inequalities for the conditional covariances, but rather assume that they are approximately equal to certain theoretical conditional covariances given Θ.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

3. A Statistical Test of CARP for a Single Focal Pair

We develop a significance test that we can use to check whether the CARP inequality holds for a single pair (X_i, X_j), called the *focal* pair. We discuss computation in six steps. The algorithm will become available in the R-package `mokken` (Van der Ark, 2007, 2012). Analyzing 100,000 samples with $N = 10000$ and $J = 10$ took less than 6 min in total.

Step 1: Select a focal item pair. We propose four strategies:

1. If the researcher suspects different items measure different attributes, pick one item representative of one attribute and another item representing another attribute. For example, some arithmetic items including item i may also measure a verbal attribute and others including item j a nonverbal attribute. Pick item i and item j.
2. If data are available from previous research, an explorative analysis may be done using factor analysis or a parametric multidimensional IRT model. If different dimensions appear, again pick two items each representing another attribute. For example, in a factor solution, not necessarily from a well-fitting model, items can be selected that load high on one factor and close to zero on another factor, thus providing a heuristic tool.
3. The CARP procedure involves splitting the sample into training and test samples. The training sample can be used to select the focal pair in the same way as in Strategy 2.
4. Let (X_i, X_j) run over all possible pairs (X_1, X_2), (X_1, X_3), ..., (X_2, X_3), (X_2, X_4), ... and apply the test to each pair. A later section discusses methods to combine multiple item pairs.

Step 2: Select a training sample. Split the total sample of N subjects randomly into a training sample of L subjects and a test sample of M subjects ($N = L + M$). The proportion of subjects in the training sample is $\ell = L/N$. We use the training sample to estimate the regression coefficients and use these in the test sample to compute the test statistic. Based on simulation work reported later, for small samples ($N \leq 500$), we recommend $\ell = .5$, and for larger samples $\ell = .2$ or $\ell = .3$.

Step 3: Estimate the regression coefficients. Linear regression analysis on the training sample yields estimates of the coefficients $a_{k.ij}$ and $a_{k.ji}$, denoted $\hat{a}_{k.ij}$ and $\hat{a}_{k.ji}$ (with $\hat{a}_{i.ij} = \hat{a}_{j.ij} = 0$).

Step 4: Estimate quantiles of the predicted scores. Using only the training sample, compute the estimated predicted scores

$$\hat{\hat{X}}_{ij} = \sum\nolimits_{k=0}^{J} \hat{a}_{k.ij} X_k,$$

and similarly, for $\hat{\hat{X}}_{ji}$. Next, determine the empirical distribution function of $\hat{\hat{X}}_{ij} + \hat{\hat{X}}_{ji}$ in the training sample. The distribution is used to define m quantiles. Here, we propose using deciles ($m = 10$). Thus, the outcome of Step 4 is a list of real numbers $\hat{q}_{1\,ij} < \hat{q}_{2\,ij} < \ldots < \hat{q}_{9\,ij}$ such that

$$P\left(\hat{q}_{(s-1)\,ij} < \hat{\hat{X}}_{ij} + \hat{\hat{X}}_{ji} \leq \hat{q}_{s\,ij}\right) \approx 0.1$$

for $s = 1, \ldots, 10$, where we write $\hat{q}_{0\,ij} = -\infty$ and $\hat{q}_{10\,ij} = \infty$. The precise algorithm used in the simulations is provided in the Supplementary Material.

Step 5: Compute the conditioning variable in the test sample. Using the estimated regression coefficients $\hat{a}_{k.ij}$ and $\hat{a}_{k.ji}$, and the estimated quantile separators $\hat{q}_{s\,ij}$ estimated in the training

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

sample, we extend the computation of $\hat{\hat{X}}_{ij} + \hat{\hat{X}}_{ji}$ to the test sample. Next, we compute the conditioning variable C_{ij} in the test data by

$$\hat{q}_{(s-1)\,ij} < \hat{\hat{X}}_{ij} + \hat{\hat{X}}_{ji} \leq \hat{q}_{s\,ij} \Longleftrightarrow C_{ij} = s$$

Step 6: Compute the one-sided Mantel–Haenszel Z. Using the test sample, test the null hypothesis that $Cov(X_i, X_j|C_{ij} = s) \geq 0$ for $s = 1, 2, \ldots, m$ by means of a one-sided version of the Mantel–Haenszel statistic. We will use the test proposed by Rosenbaum (1984, p. 429; see Kuritz et al., 1988, for a discussion of different versions of the Mantel–Haenszel method). The following description is copied almost verbatim from Rosenbaum: Denote the number of subjects in the test sample having $X_i = a$, $X_j = b$, and $C_{ij} = s$ as n_{abs} for $a, b = 0, 1; s = 1, 2, \ldots, m$ and denote the marginal totals as $n_{+bs} = n_{0bs} + n_{1bs}$, $n_{a+s} = n_{a0s} + n_{a1s}$, $n_{ab+} = \sum_s n_{abs}$, etc. Compute

$$e_+ = \sum\nolimits_{s=1}^{m} \frac{n_{1+s} n_{+1s}}{n_{++s}},$$
$$v_+ = \sum\nolimits_{s=1}^{m} \frac{n_{1+s} n_{0+s} n_{+1s} n_{+0s}}{n_{++s}^2 (n_{++s} - 1)},$$

and then the test statistic

$$Z_{ij} = \frac{n_{11+} - e_+ + 0.5}{\sqrt{v_+}}.$$

The p-value is computed as $p_{ij} = \Phi^{-1}(Z_{ij})$, where $\Phi^{-1}(.)$ is the inverse of the standard normal distribution function. See Rosenbaum (1984, p. 429) for more details and the rationale of Z_{ij}. The sample covariance of X_i and X_j in the layer with $C_{ij} = s$ is given by

$$\widehat{cov}(X_i, X_j|C_{ij} = s) = \frac{n_{11s}}{n_{++s}} - \frac{n_{1+s}}{n_{++s}} \frac{n_{+1s}}{n_{++s}}$$

and therefore,

$$n_{11+} - e_+ = \sum\nolimits_{s=1}^{m} n_{++s} \widehat{cov}(X_i, X_j|C_{ij} = s).$$

The numerator of Z_{ij} is therefore a weighted sum of the conditional covariances of X_i and X_j, given the grouped weighted rest scores, with a continuity correction. Rosenbaum noticed that the quantities e_+ and v_+ are the expectation and variance of n_{11+} in the least favorable case of the null hypothesis, which is the case where $Cov(X_i, X_j|C_{ij} = s) = 0$ for $s = 1, 2, \ldots, m$. If the null hypothesis is true, then $Cov(X_i, X_j|C_{ij} = s) \geq 0$ and $\mathbb{E}(n_{11+} - e_+) \geq 0$, so that Z_{ij} has an asymptotic normal distribution with $\mathbb{E}(Z_{ij}) \geq 0$.

The optimal number (m) of quantile groups in Step 4 is rather arbitrary in the sense that there is no definitive number. Rosenbaum (1984) suggested deciles ($m = 10$) in his "Case 5", but also used the raw rest score, which has $J - 2$ levels. We did simulations with both linear and logistic regression, with $m = 10$, $m = J - 2$, and $m = \sqrt{N}$. The differences in power between these versions were small, but linear regression with $m = 10$ had slightly higher power than the other options. Therefore, we use linear regression with $m = 10$ in all simulations below.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

4. Asymptotic Type 1 Error Rate

We provide a formal proof that the Type 1 error rate is under control as $N \to \infty$. Note that in Step 6 we suggested a one-sided version of a Mantel–Haenszel test, but that multiple versions of the Mantel–Haenszel test exist (Kuritz et al., 1988), and more versions can be developed in the future. We want a result that is valid for all these versions, and rather than delving into the details of each possible version, we will make the general assumption that in Step 6 one applies a test with the following property: If the test is applied to data of a $2 \times 2 \times K$ table to test the null hypothesis that the covariance is nonnegative in each of the K layers, then the asymptotic Type 1 error rate is under control in the sense that the p-values stochastically dominate a standard uniform random variable as the sample size grows to infinity. Now, the question is whether that remains true in our case, where the layers are partially based on the regression estimated from a training sample rather than on a fixed variable in the test sample.

Proposition 1. *If subjects are drawn randomly and independently and if the test sample grows infinitely large while the training sample remains fixed, then the asymptotic Type 1 error rate of the CARP test is under control.*

Proof. Denote the data of the L subjects of the training sample $\mathbf{X}^{(1)} = (\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_L^{(1)})$ and the data of the M subjects of the test sample $\mathbf{X}^{(2)} = (\mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_M^{(2)})$. Subjects are drawn randomly and independently, therefore we consider the random vectors $\mathbf{X}_n^{(t)}$ as independent and identically distributed (iid) copies of $\mathbf{X}$. The k-th item score in $\mathbf{X}_n^{(t)}$ is denoted $X_{nk}^{(t)}$. Define $g(.)$ to be the function such that, for any vectors $\boldsymbol{\alpha} \in \mathbb{R}^{2(J+1)}$ and $\boldsymbol{\beta} \in \mathbb{R}^{m-1}$ with $\beta_1 < \beta_2 < \ldots < \beta_{m-1}$, if we write $\beta_0 = -\infty$ and $\beta_m = \infty$, then

$$\beta_{s-1} < \sum_{k=0}^{J} (\alpha_k + \alpha_{k+J+1}) X_k \leq \beta_s \Longleftrightarrow g(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = s,$$

for $s = 1, 2, \ldots, m$. This definition parallels the definition of the conditioning variable C_{ij}, given in the description of Step 5. We denote the vectors of estimated regression coefficients and quantile separators $\hat{\mathbf{a}}_{ij}\left(\mathbf{X}^{(1)}\right)$ and $\hat{\mathbf{q}}_{ij}\left(\mathbf{X}^{(1)}\right)$, respectively, so that we can write the conditioning variable for the n-th subject in the test sample as

$$C_{ijn} = g\left(\mathbf{X}_n^{(2)}, \hat{\mathbf{a}}_{ij}\left(\mathbf{X}^{(1)}\right), \hat{\mathbf{q}}_{ij}\left(\mathbf{X}^{(1)}\right)\right).$$

Now, consider the conditional covariance of the form $Cov(X_i, X_j | C_{ij} = s)$. For the n-th subject in the test sample, the corresponding covariance is $Cov\left(X_{ni}^{(2)}, X_{nj}^{(2)} | C_{ijn} = s\right)$. Consider the latter covariance conditionally on $\mathbf{X}^{(1)}$. Given $\mathbf{X}^{(1)}$, $\mathbf{X}_n^{(2)}$ is conditionally associated (because $\mathbf{X}$ is conditionally associated and $\mathbf{X}_n^{(2)}$ is a copy of $\mathbf{X}$ that is independent of $\mathbf{X}^{(1)}$). Furthermore, given $\mathbf{X}^{(1)}$, C_{ijn} depends only on $\mathbf{X}_n^{(2)}$ with the two variables $X_{ni}^{(2)}$ and $X_{nj}^{(2)}$ excluded (since we required $\hat{a}_{i.ij} = \hat{a}_{j.ij} = 0$), and therefore, $Cov\left(X_{ni}^{(2)}, X_{nj}^{(2)} | C_{ijn} = s, \mathbf{X}^{(1)}\right)$ is implied to be nonnegative by conditional association of $\mathbf{X}_n^{(2)} | \mathbf{X}^{(1)}$. Nonnegativity holds for $n = 1, 2, \ldots, M$. It can be concluded that the data of the test sample can be considered as M independent draws from a population with $Cov(X_i, X_j | C_{ij}, \mathbf{X}^{(1)}) \geq 0$. Therefore, the asymptotic distribution of $p_{ij} | \mathbf{X}^{(1)}$ dominates the uniform (0, 1) distribution in the sense that for any $\alpha \in (0, 1)$, $\limsup_{M \to \infty} P(p_{ij} < \alpha | \mathbf{X}^{(1)}) \leq \alpha$. The decision rule "reject the null hypothesis if $p_{ij} < \alpha$" will thus lead to asymptotic

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

Type 1 error rate $\limsup_{M\to\infty} P(p_{ij} < \alpha) = \limsup_{M\to\infty} \mathbb{E}(P(p_{ij} < \alpha|\mathbf{X}^{(1)}))$, and with the reverse Fatou lemma we have that this is $\leq \mathbb{E} \limsup_{M\to\infty} P(p_{ij} < \alpha|\mathbf{X}^{(1)}) \leq \alpha$ □

Proposition 1 holds no matter how poor the estimates $\hat{\mathbf{a}}$, $\hat{\mathbf{q}}$ are or how much off-target the heuristic tool is. All that is needed to control the Type 1 error rate is that the subjects are drawn iid, that the estimates $\hat{\mathbf{a}}$, $\hat{\mathbf{q}}$ are based on the training sample, that the training sample is independent of the test sample, and that the weights of the focal variables are fixed to zero in $\hat{\mathbf{a}}$.

Proposition 1 assumes that the size of the training sample remains fixed while the size of the test sample increases. If, however, L and M increase simultaneously, we presumably also need that $\hat{\mathbf{a}}$ and $\hat{\mathbf{q}}$ converge as $L \to \infty$, because C_{ijn} depends on $\mathbf{X}^{(1)}$ and therefore, on L. If $\hat{\mathbf{a}}$ and $\hat{\mathbf{q}}$ converge, then the C_{ijn}s stabilize, and we expect that the proof can be modified to establish that the Type 1 error rate is under control in this situation as well. However, we see no point in dwelling on cases with $L \to \infty$, because increasing the training sample has almost no benefits once the standard errors of $\hat{\mathbf{a}}$ and $\hat{\mathbf{q}}$ become very small. For all practical purposes we can therefore add to our procedure the prescription to cap L when the estimated standard errors of $\hat{\mathbf{a}}$ and $\hat{\mathbf{q}}$ are below a certain small threshold. This happens almost surely for large L if these estimates are obtained by linear regression and the empirical distribution function, as we discussed in the previous section. Then, the proof suffices to establish asymptotic Type 1 error rate control.

We explored whether the test can be modified such that the training sample and the test sample both include the whole sample, but simulations showed that this modification causes the Type I error rate to exceed the nominal significance level in some cases. Hence, we recommend cross-validation.

5. Simulation Studies

5.1. Method

5.1.1. General Set-Up We used J items and a logistic model, $P(X_i = 1|\Theta_1, \Theta_2, \Theta_3) = (1 + \exp(-(\alpha_{i1}\Theta_1 + \alpha_{i2}\Theta_2 + \alpha_{i3}\Theta_3 + \beta_i)))^{-1}$, where $(\Theta_1, \Theta_2, \Theta_3)$ has a trivariate standard normal distribution with correlations 0. Denote the number of items that load on dimensions 1, 2, and 3 as J_1, J_2, and J_3, respectively, so that $J_1 + J_2 + J_3 = J$. We distinguish the *standard two-dimensional case* as a special case with $\alpha_{id} = 1$ if item i loads on dimension d, and $\alpha_{id} = 0$ otherwise, $\beta_i = 0$, and $J_1 = J_2$ and $J_3 = 0$. We call this the 'standard', but it represents a failure if the goal was to create a unidimensional test that satisfies MH.

5.1.2. Optimum Size of Training Sample Training samples in cross-validation often contain at least 70% of the observations of the whole sample. We did some simulations to find out whether we must maintain that percentage here. We used the standard two-dimensional case with $J \in \{12, 24\}$ and $N \in \{500, 1000, 2000, 5000\}$, and $\ell \in \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$, using 1000 samples per (J, N, ℓ) cell. For each combination of J and N, we fitted the power of the CARP method found in the simulation by a quadratic regression on ℓ. From the estimated regression coefficients, we computed the value of ℓ for which the quadratic curve has its maximum. Table 1 shows that for small samples ($N = 500$), the estimated optimum was close to $\ell = .5$, and for large samples the optimum was rather $\ell = .2$ or $\ell = .3$.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

TABLE 1.
Estimated optimum values of training proportion ℓ for varying sample size N and test length J.

	J	
N	12	24
500	.45	.48
1000	.38	.43
2000	.31	.36
5000	.19	.31

6. Results

6.1. Type I Error Rate

For large samples, the Type I error rate is under control because of the asymptotic properties of the Mantel–Haenszel test. It suffices to study the error rate for small samples, and we focused first on $N = 500$ with $\ell = .3$ for this purpose. In unidimensional cases, we chose $\alpha_{i2} = \alpha_{i3} = 0$ for $i = 1, \ldots, J$. We focused on cases with $\alpha_{i1} = \alpha_{i2} = \alpha_{i3} = 0$, which we refer to as zero-dimensional. (One can also describe this case as J-dimensional, but the number of common dimensions would still be 0.) These cases are interesting because all CARP covariances are zero, whereas they are positive in the unidimensional case with positive loadings. Consequently, the rejection rates were generally higher in zero-dimensional cases than in other unidimensional cases. Parameters that were not fixed to 0 were randomly drawn from the following distributions: $\beta_i \sim Uniform(-1.5,\ 1.5)$ and $\alpha_{i1} \sim Uniform(0.5,\ 2.5)$. We studied the effect of the number of items, J, varying between 10 and 50. For each J, we simulated 100 parameter sets S, each consisting of $(\alpha_{i1},\ \beta_i), i = 1, \ldots, J$. Next, for each of the 100 parameter sets we simulated 1000 samples of N subjects responding to the J items and applied the CARP test procedure to this sample with nominal significance level $\alpha = .05$. Thus, for each J we have 100 parameter sets, and for each parameter set, we obtained a rejection rate based on 1000 samples.

Table 2 shows the quartiles of the rejection rates with $\ell = .3$ for some selected values of J. The maximum rejection rate over all 4100 zero-dimensional cases (41 values of J, each with 100 cases of 1000 samples) was .065, which is not significantly larger than .05 according to a binomial test with multiple testing correction (for a single case of 1000 samples, the p-value would be $1 - pbinom\,(65, 1000, .05) = .0149$, but for the maximum of 100 cases the p-value is $1 - pbinom\,(65, 1000, .05)^{4100} = 1$). The mean rejection rate was .038. Figure 1 shows the cumulative percentages of the rejection rates along with the expected cumulative percentages derived from a binomial distribution with success probability .05. The expected distribution clearly dominates the distribution of rejection rates. Therefore, we conclude that the Type I error rate of the CARP test is under control in these cases.

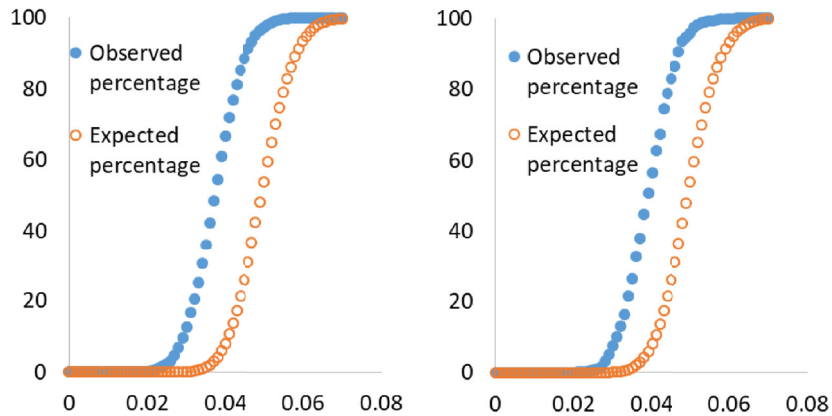
We also did simulations where ℓ, N, and J were randomly drawn from a uniform distribution, with $\ell \in (.2, .5)$, $N \in [500, 1000]$, and $J \in [10, 50]$. We simulated zero-dimensional parameter cases with 1000 samples each. The maximum rejection rate was .062, which is not significant according to a binomial test with multiple testing correction ($p = 1 - pbinom\,(62, 1000, 0.05)^{1000} = 1$). The mean rejection rate was .040. Figure 1 shows the cumulative frequencies of the rejection rates along with the expected cumulative frequencies derived from a binomial distribution with success probability .05. The binomial distribution clearly dominates the distribution of the rejection rates. Therefore, we conclude that the Type I error rate is

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

TABLE 2.
Type I error rates in zero-dimensional cases.

	J				
Quartile	10	20	30	40	50
0 (minimum)	.018	.024	.019	.025	.022
1	.032	.034	.034	.032	.033
2	.038	.038	.038	.037	.038
3	.041	.040	.043	.041	.041
4 (maximum)	.052	.052	.055	.051	.051

Each column is based on 100 cases with 1000 samples each, with $\ell = .3$ and $N = 500$.



Note. Left: Based on 4,100 cases with $N = 500$ and $\ell = .3$ and $J = 10,\ldots,50$. Right: Based on 1,000 cases with $N \in [500, 1{,}000]$, $\ell \in (.2, .5)$ and $J \in [10,50]$. In all cases $\alpha = .05$.

FIGURE 1.
Cumulative Percentages of Type I Error Rates in Zero-Dimensional Cases.

under control. The simulations with random ℓ, N, and J were also conducted for unidimensional cases with $\alpha_{i1} \sim Uniform(0.5, 2.5)$. These rejection rates were all well below .05.

Finally, we did some simulations with $N \gg 500$ and small J. Both zero-dimensional and unidimensional cases were simulated with $\ell \in (.2, .5)$, $N \in \{5000, 10000, 20000\}$, and $J \in \{5, 10\}$ with 10 parameter cases per cell and 1000 samples per parameter case. The rejection rates were again around .05 in the zero-dimensional cases, and close to 0 in the unidimensional cases.

6.2. *Power: Single Focal Pair, Effect of Dimensionality, and Item Distribution*

We chose $\alpha_{id} = 1$ if item i loads on dimension d, and $\alpha_{id} = 0$ otherwise. We used $\beta_i = 0$. Consider the standard two-dimensional case where $J_1 = J_2$ and $J_3 = 0$. For large N, if linear regression is used, $\hat{X}_{ij} + \hat{X}_{ji}$ converges to a linear transform of the rest score $(\sum_k X_k) - X_i - X_j$. Therefore, the power of the CARP test will approach that of a CRS test. However, for finite N, the power of the CARP test will remain below that of the CRS test, because part of the sample is used for training and not for testing, and because $\hat{X}_{ij} + \hat{X}_{ji}$ is not exactly equal to the rest score yet. This was indeed what we found in simulations. Therefore, one may consider the standard

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

TABLE 3.
Rejection Rates of CARP and CRS Tests in Two-Dimensional Cases with $N = 5000$.

J	J_1	J_2	J_3	CARP	CRS
12	2	10	0	**.363**	.231
12	3	9	0	**.601**	.533
12	4	8	0	.727	**.797**
12	5	7	0	.792	**.929**
12	6	6	0	.791	**.949**
24	2	22	0	**.320**	.118
24	3	21	0	**.627**	.234
24	4	20	0	**.753**	.450
24	5	19	0	**.836**	.680
24	6	18	0	**.889**	.839
24	7	17	0	.892	**.913**
24	8	16	0	.937	**.970**
24	9	15	0	.929	**.989**
24	10	14	0	.950	**.995**
24	11	13	0	.945	**.995**
24	12	12	0	.954	**.998**

Values in bold are the largest power in the row
Each row is based on 1000 samples.

two-dimensional case as the ideal case for Rosenbaum's CRS test. Next, we will compare the power of the CARP test and the CRS test in various deviations from this standard case. The first kind of deviations is that $J_1 \neq J_2$. The second kind of deviation is that $J_3 > 0$, introducing items that load on a third dimension that we assume uncorrelated with the other two dimensions. These simulations were conducted with all J between 9 and 39 that are multiples of 3, but we report results in detail only for $J = 12$ and $J = 24$.

Table 3 shows the CARP test's power with $N = 5000$, and $\ell = .2$ for cases with $J = 12$ or $J = 24$, and $J_3 = 0$. The CARP test had significantly greater power than the CRS test in the seven cases with $J_1/J < .27$ or $J_1/J > .73$ and in some of these cases, the power of the CARP test was considerably larger. In the other nine cases, the CRS test had more power, but the power of the CARP test was rather close to it. In general, the CARP test had greater power. The results for other values of J were similar: For $J_1/J < .27$ or $J_1/J > .73$, the CARP test had significantly greater power than the CRS test. For J_1/J between .30 and .70, the CARP test had significantly smaller power than the CRS test. For values of J_1/J between .27 and .30, or between .70 and .73, the difference in power between the CARP and CRS tests was usually not significant.

Table 4 shows the power with $N = 5000$, and $\ell = .2$ for cases with $J = 12$ or $J = 24$, $J_1 = J_2$, and $J_3 > 0$. The CARP test had greater power than the CRS test in the ten cases with $J_1 = J_2 \leq J_3$, and in some of these cases the power of the CARP test was considerably greater. In the other four cases with $J_1 = J_2 > J_3$, the CRS test had greater power, but the power of the CARP test was still substantial. In general, the CARP test had greater power. Results for other values of J were similar. For small N, the CARP test lost power compared to the CRS test, because the training sample was excluded from the test. Thus, the results for smaller N were more favorable for the CRS test.

6.3. Power: Single Focal Pair, Effect of Item Parameters and Sample Size

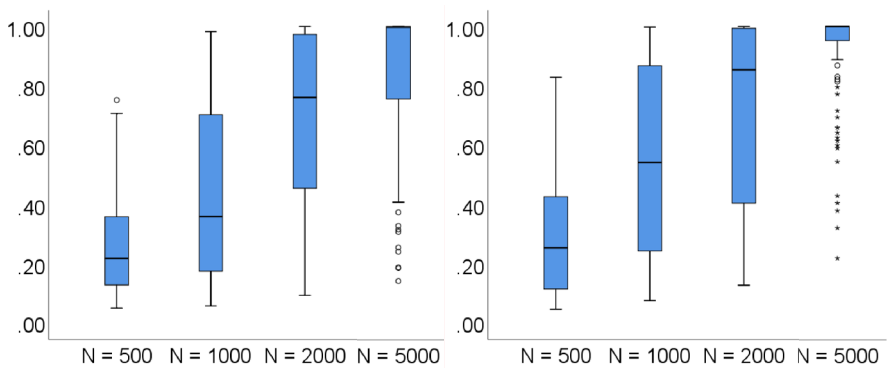
We studied two-dimensional tests with $J_1 = J_2$ and $J_3 = 0$, with $N = 500,\ 1000,\ 2000,\ 5000$, $\ell = .3$, and $J = 12$ or $J = 24$. We chose $\alpha_{i2} = 0$ for $i = 1, \ldots, J_1$ and $\alpha_{i1} = 0$ for

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

TABLE 4.
Rejection rates of CARP and CRS tests in three-dimensional cases with $N = 5{,}000$.

J	J_1	J_2	J_3	CARP	CRS
12	2	2	8	**.154**	.064
12	3	3	6	**.373**	.148
12	4	4	4	**.541**	.398
12	5	5	2	.684	**.767**
24	2	2	20	**.130**	.061
24	3	3	18	**.285**	.060
24	4	4	16	**.476**	.092
24	5	5	14	**.622**	.171
24	6	6	12	**.732**	.240
24	7	7	10	**.801**	.493
24	8	8	8	**.855**	.729
24	9	9	6	.899	**.915**
24	10	10	4	.914	**.982**
24	11	11	2	.945	**.995**

Values in bold are the largest power in the row
Each row is based on 1000 samples.



Note. Left: $J = 12$; Right: $J = 24$. Each boxplot is based on 100 cases, each case has 1000 samples.

FIGURE 2.
Rejection rates in two-dimensional cases with different item parameters.

$i = J_1 + 1, \ldots, J$. Parameters that were not fixed to 0 were randomly drawn from the following distributions: $\beta_i \sim Uniform(-1.5,\ 1.5)$ and $\alpha_{i1},\ \alpha_{i2} \sim Uniform(0.5,\ 2.5)$. For $J = 12,\ 24$, we simulated 100 parameter sets S, each consisting of $(\alpha_{i1},\ \alpha_{i2},\ \beta_i), i = 1, \ldots, J$. Next, for each of the 100 parameter sets we simulated 1000 samples of N subjects responding to the J items and applied the CARP test procedure to this sample with nominal significance level $\alpha = .05$. Figure 2 shows modified boxplots of the rejection rates. As expected, the power increased with N, but variation was large due to the parameter sets. For $N = 5000$, most parameter sets would have had power greater than .80.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

7. Additional Results for the CARP Tests

7.1. Aggregation of CARP Tests across Item Pairs

We discuss how one can combine tests of multiple item pairs without prior selection. If one applies a CARP test to all item pairs of a scale, this produces a sequence of $J(J-1)/2$ p-values. To keep the family-wise Type I error rate (FWER) under control, a multiple testing correction may be in order. Alternatively, one could choose to control the false discovery rate (FDR), which generally leads to tests with higher power (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). However, applying either correction method to all $J(J-1)/2$ p-values results in unnecessary loss of power if many of the involved covariances are positive. Instead, one may consider only the item pairs that have a negative conditional covariance in the training sample and test their conditional covariances in the test sample. We collect the set of item pairs $(i, j), i > j$ with a negative conditional covariance in the training sample in set $\mathcal{S}$, and let $S := |\mathcal{S}|$ denote its size. Since this statistic is independent of the data in the test sample, one may apply the Bonferroni correction with this number; that is, reject the null hypothesis for pair (i, j) iff $(i, j) \in \mathcal{S}$ and $p_{ij} \leq \alpha/S$. To check that this correction with S controls the FWER, assume that under the null hypothesis the distribution of each p_{ij} dominates the $Uniform(0, 1)$ distribution in the sense that $\mathbb{P}\left(p_{ij} \leq x\right) \leq x$ for all $x \in \mathbb{R}$ where $\mathbb{P}$ denotes the probability measure of the p_{ij}. This condition is called supra-uniformity by Ellis et al. (2020). The condition is satisfied if p_{ij} dominates the $Uniform(0, 1)$ distribution in likelihood ratio order (Whitt, 1980), and it is equivalent to p_{ij} having a density that is increasing on the interval $(0, 1)$. This is true if the test is conducted with $p_{ij} = \Phi^{-1}\left(Z_{ij}\right)$ with $Z_{ij} \sim \mathcal{N}(\mu_{ij}, 1)$ and $\mu_{ij} \geq 0$. Let $\mathcal{R}_{ij}$ denote the event that the null hypothesis is rejected for pair (i, j), then, for a fixed pair (i, j) we have $\mathbb{P}\left(\mathcal{R}_{ij}|(i, j) \in \mathcal{S}\right) \leq \alpha/S$ and $\mathbb{P}\left(\mathcal{R}_{ij}|\ (i, j) \notin \mathcal{S}\right) = 0$; therefore, the FWER is

$$\mathbb{P}\left(\bigcup_{i,j} \mathcal{R}_{ij}\right) = \mathbb{E}\left(\mathbb{P}\left(\bigcup_{i,j} \mathcal{R}_{ij}|\mathcal{S}\right)\right) \leq \mathbb{E}\left(\Sigma_{i,j}\mathbb{P}\left(\mathcal{R}_{ij}|\mathcal{S}\right)\right) \leq \mathbb{E}\left(S\frac{\alpha}{S}\right) = \alpha.$$

(In $\mathbb{P}\left(\bigcup_{i,j} \mathcal{R}_{ij}|\mathcal{S}\right)$ and $\mathbb{P}\left(\mathcal{R}_{ij}|\mathcal{S}\right)$, $\mathcal{S}$ is treated as a random variable with outcomes that enumerate the power set of $\{1, \ldots, J\}^2$.)

Alternatively, one could try to compound the Z-statistics in the formula

$$Z_{total} = \frac{\sum_{(i,j)\in\mathcal{T}} Z_{ij}}{\sqrt{v}},$$

where $\mathcal{T} \subseteq \mathcal{S}$ is some further restricted subset of item pairs, and v is an estimate of the variance of the numerator. The Z_{ij}s are correlated, and to obtain v one should somehow estimate their average correlation; see Efron (2010), who discusses methods for this purpose. An advantage is that compounding can increase the power, as was concluded by Straat et al. (2016) for the CA tests of Rosenbaum (1984). We encourage future researchers to develop improved compounding rules.

7.2. Comparison with Other Methods

So far, we compared CARP mathematically with other incomplete tests of CA, such as testing MTP_2, and we compared it in simulations with Rosenbaum's (1984) CRS test, which also assesses CA. Next, we briefly touch upon the topics of parametric goodness-of-fit tests and

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

discuss the difference of the CARP method and existing approaches to dimensionality assessment in nonparametric IRT. We notice that the CARP test has been developed for a single focal pair, so that aggregation over multiple item pairs is a topic for future research. The alternatives that we discuss typically aggregate over all item pairs. This is true for Chi-square and RMSEA in factor analysis, the statistics Q_2 (Van den Wollenberg, 1982), R_2 (Glas, 1988), and M_2 (Maydeu-Olivares & Joe, 2006) in logistic IRT models, and DETECT (Zhang & Stout, 1999b) in nonparametric IRT. Our discussion can therefore not include a quantitative comparison with respect to statistical power.

7.2.1. Comparison with Parametric Goodness-of-Fit Tests Two reviewers suggested to compare the CARP procedure with a method where a parametric multidimensional model is fitted first and then a parametric unidimensional model is compared with it, using a goodness-of-fit test such as a likelihood-ratio test. We will call this method the *parametric goodness-of-fit comparison* (PGC). Examples of this are (1) testing multiple linear factor models and compare their chi-square statistics, assuming normal distributions, or compare their Chi-squares, RMSEAs or eigenvalues. We mention this possibility because of its popularity in psychology; (2) similar but using logistic IRT models instead of linear factor models (e.g., Bartolucci, 2007; Christensen et al. 2002); (3) testing different latent class models (e.g., Bartolucci et al. 2017; Ligtvoet & Vermunt 2012; Van Onna, 2002; Vermunt, 2001). The idea is that a latent class model can approximate nonparametric multidimensional and unidimensional models if the number of latent classes is large enough; and (4) testing monotonic polynomial models (Falk & Cai, 2016). These models can be used to approximate multidimensional and unidimensional models if the number of polynomial terms is large enough, and therefore a similar strategy can be used.

We find this approach interesting, but we are not yet convinced that in the long run it is more helpful than our approach, which is focussed on critical data patterns such as negative covariances rather than comparing goodness-of-fit statistics. Our restraints concerning the PGC methods are the following. First, it is generally difficult to know how many dimensions the multidimensional model should have, and how this influences the decision on the unidimensional model. Second, it is unclear to which extent the auxiliary assumptions (linearity, logistic response function, normality, number of latent classes, number of polynomial terms) influence the goodness-of-fit of the unidimensional model. Third, if a goodness-of-fit test indicates that the unidimensional model is wrong, it might not be clear which items are causing the problem. For some models, item-fit statistics have been proposed (e.g., Sijtsma & Van der Ark, 2021) that must be used in combination with statistics assessing the fit of sets of items. Another variant is that an alternative unidimensional model must be chosen, but then the large array of possibilities provides a new choice problem (which model is the most obvious choice?) and corresponding analysis problem (how to avoid endless trial and error?). We conclude that the application of methodologies other than the one we study in this article comes with complexities hindering their straightforward use as well as a simple comparison with our CARP methodology.

7.2.2. Comparison with Item Selection Procedures in Nonparametric IRT In the context of nonparametric IRT, several procedures have been proposed to assess the dimensionality of an item set. The automated item selection procedure (AISP; Mokken, 1971; Sijtsma & Molenaar, 2002) uses a bottom-up algorithm to select items in unidimensional subsets based on a definition of a scale that uses non-negative inter-item covariances and positive scalability coefficients. Straat et al. (2013) proposed a genetic algorithm to replace and remedy some of the peculiarities of the AISP. The goal of both procedures is to have as many items possible in the first scale, as many from the remaining items—if available—in the second scale, and so on. Zhang & Stout (1999b, p. 239) defined the "bias-corrected estimator for the theoretical DETECT index" as a weighted average of covariances of the form $Cov\left(X_i, X_j \mid T\right)$, with sum score $T = \sum_i X_i$, and $Cov\left(X_i, X_j \mid R_{ij}\right)$,

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

with rest score $R_{ij} = \sum_{k \neq i,j} X_k$, where the DETECT weights are such that the pair (i, j) contributes if and only if both items are in the same cluster. Next, Zhang and Stout try to find the partition that maximizes this index using a heuristic procedure. Roussos, Stout, and Marden (1998) proposed an agglomerative hierarchical cluster analysis for finding subsets of items, using the software package HCA/CCPROX. The procedure provided the choice between different statistics, including covariances conditional on rest scores not including items known to be in already formed clusters, for assessing the relationship between items, and different agglomerative hierarchical clustering methods. They did not use a formal criterion for identifying a final solution but rather left this to the researcher to decide, for example, based on theoretical expectations of the item set's dimensionality. The DIMTEST procedure assesses the hypothesized unidimensionality of a user-specified item set (Nandakumar & Stout, 1993; Stout, 1987). Thus, unlike the other procedures, DIMTEST is confirmatory and cannot directly be used to partition items in different clusters in an exploratory analysis. Several variations on the original procedure have been proposed; see Stout et al. (2001) and Kieftenbeld & Nandakumar (2015). Van Abswoude, Van der Ark, and Sijtsma (2004) systematically compared the methods.

The CARP procedure is different from these and other item selection procedures proposed in the nonparametric IRT context (e.g., Brusco, Köhn, & Steinley, 2015). It shares with several of these procedures a certain open-endedness caused by the complexities typical of a fine-grained analysis of the data involving many item pairs or item subsets and subdivisions of the sample into score groups, dealing with finite sample sizes and empty or near-empty cells in contingency tables, and combining many detailed results into one useful conclusion about the dimensionality of an item set. Because so many arbitrary researcher decisions are needed to obtain a result, not only for the CARP procedure but also for other procedures many precautions are needed to be able to compare them thoroughly. This is a project requiring a separate study.

8. Discussion

We developed the CARP test, which often distinguishes data generated by a two-dimensional model from data generated by a unidimensional monotone model, even if the data are MTP_2 and have MM. The test uses CA and can be viewed as a generalization of Rosenbaum's (1984) proposal to test the covariance of each item pair conditionally on their unweighted rest score (the CRS test). The CARP test conditions on a weighted rest score, where the weights are based on regression analyses in a training sample consisting of 20% to 50% of the total sample. Each of the items in a focal pair (i, j) is used as dependent variable in a linear regression analysis that predicts them from the remaining items. The sum of the two predicted scores is computed in the test sample and is used as the weighted rest score. The weighted rest score divides the test sample into deciles and a directional Mantel–Haenszel test tests whether the covariance of (i, j) is nonnegative in each decile group.

Data generated by means of unidimensional logistic models showed that the Type I error rate is under control, even if the overall inter-item correlations are 0. Simulations with two-dimensional logistic models showed the power of the CARP test exceeds the power of the CRS test if one dimension has three times more items than the other dimension. Simulations with three-dimensional logistic models showed the power of the CARP test exceeds the power of the CRS test if the third dimension has at least a third of the items. In the extreme two-dimensional case, where both dimensions have the same number of items with equal loadings and difficulty parameters for all items, the CARP test converges to the CRS test as the sample size increases. Thus, in comparison with Rosenbaum's (1984) CRS test, our CARP test gains power in a variety of multidimensional cases at the cost of losing some power in extreme two-dimensional cases with

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

equally important dimensions. Because tests are usually aimed at being unidimensional, most of the items indeed resulting in targeting this dimension, the results for the CARP test are positive.

We explored for multiple focal items that compounding their test statistics can increase the power. The CARP method looks promising, but as with any newly developed method it also raises questions for future research. First, what are the optimal values of ℓ (the size of the training sample) and m (the number of groups in conditioning), and how do these values depend on N and J? Second, in the cross-validation, rather than drawing one training sample one might repeat drawing and then aggregate the results over draws, thus reducing the variability of the outcomes. Which aggregation rules are suitable? Third, how can one compound test results for multiple item pairs? Fourth, a more elaborate study of the dependence of the power on the number of items, the number of dimensions, the shape of the response function (logistic or other), and the item parameters could be done. Fifth, the CARP inequalities also hold for polytomous items. Which test procedures are most useful? Rosenbaum (1984, p. 429) provides suggestions. Sixth, how does the power profile of the CARP test compare to the semiparametric methods of Bartolucci (2007) and Falk and Cai (2016)?.

In our analysis, we assumed a priori that conditional independence holds, which is consistent with the fact that for a finite number of binary items, without other restrictions, conditional independence is a "vacuous assumption" (Holland & Rosenbaum, 1986, p. 1525). Moreover, assuming monotonicity, we developed the CARP test as a test of unidimensionality versus multidimensionality. However, if the CARP test points to a violation of MH, this cannot be attributed to a single assumption. An alternative model may thus assume local dependence or correlated errors instead of multidimensionality.

The CARP method can be a useful addition to the existing methods for testing MH and detecting multidimensionality in monotone models. It may help answer a fundamental empirical question without relying on features of parametric models that are irrelevant to the research question. We already mentioned the present CHC intelligence representation using multiple factors (Wasserman, 2019) that is based on parametric—mostly linear—models. This choice is mathematically convenient but may be irrelevant for distinguishing the factors and damaging when it dominates the data analysis. A significant negative covariance obtained in a CARP test would demonstrate that the distinction between intelligence factors is not an artifact of the parametric assumptions, and it would rule out every unidimensional monotone model for intelligence. This is another topic for future research.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix A

In this appendix, we state the precise definitions of the various models and conditions that are relevant here. Consider a vector of binary manifest variables, $\mathbf{X} = (X_1, \ldots, X_J)$. Variable X_i represents the scores (1 = correct, 0 = incorrect) subjects obtained on the i-th item. Suppose that

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

in the probability space of $\mathbf{X}$ there is some random vector, $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_D)$, which represents the latent variables. We will use the following conditions (Holland & Rosenbaum, 1986; Mokken, 1971; Rosenbaum, 1984), adapted to binary manifest variables:

MH1. (conditional independence). $\mathbf{X}$ is *conditionally independent* given $\boldsymbol{\Theta}$ if $P(\mathbf{X} = \mathbf{x} | \boldsymbol{\Theta}) = \prod_{i=1}^{J} P(X_i = x_i | \boldsymbol{\Theta})$ for all $\mathbf{x} \in \{0, 1\}^J$.
MH2. (monotonicity). $\mathbf{X}$ is *monotone* with $\boldsymbol{\Theta}$ if $P(X_i = 1 | \boldsymbol{\Theta})$ is monotonically increasing in each coordinate of $\boldsymbol{\Theta}$ for all $i = 1, \ldots, J$.
MH3. (unidimensionality). $\boldsymbol{\Theta}$ is *unidimensional* if $D = 1$.

Definition 1. *(monotone latent variable model;* Holland & Rosenbaum, 1986). $(\mathbf{X}, \boldsymbol{\Theta})$ is a *monotone latent variable* (MLV) model if $\mathbf{X}$ is conditionally independent given $\boldsymbol{\Theta}$ and $\mathbf{X}$ is monotone with $\boldsymbol{\Theta}$. If, additionally, $\boldsymbol{\Theta}$ is unidimensional, then $(\mathbf{X}, \boldsymbol{\Theta})$ is a unidimensional monotone latent variable model.

Mokken (1971), Mokken and Lewis (1982) introduced the unidimensional monotone latent variable model for binary items as the monotone homogeneity (MH) model. Ellis and Junker (1997) reformulated Definition 1 such that it is a property of $\mathbf{X}$ rather than $(\mathbf{X}, \boldsymbol{\Theta})$:

Definition 2. *(monotone homogeneity).* $\boldsymbol{X}$ satisfies a *unidimensional monotone latent variable model* or *monotone homogeneity* (MH) if there exists a unidimensional variable Θ such that $\mathbf{X}$ is conditionally independent given Θ and $\mathbf{X}$ is monotone with Θ.

Ellis (2015) studied a narrower formulation of monotone models that will be called 'monotone factor models' here. His assumptions can be rephrased as follows:

MF1 $X_i = \phi_i(\eta_i + \varepsilon_i)$ for $i = 1, \ldots, J$, where each ϕ_i is an increasing function (henceforth, called a response function), and η_i and ε_i are latent variables. We write $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)$.
MF2 $\boldsymbol{\eta} = \psi(\boldsymbol{\lambda}\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is a multivariate vector with independent components, $\boldsymbol{\lambda}$ is a real matrix, and ψ is an increasing function.
MF3 The latent variables $\varepsilon_1, \varepsilon_2, \ldots$ are independent from each other and independent of $\boldsymbol{\Theta}$.
MF4 The ε_i have densities that are Pólya frequency functions of order 2 (PF2; Efron, 1965).
MF5 $\boldsymbol{\lambda}$ has a simple structure where every manifest variable loads positive on one factor and zero on the other factors.

Assumption MF1 means that the item responses are a dichotomization of underlying item-specific latent variables $\eta_i + \varepsilon_i$ that one may view as latent responses. This has been discussed earlier in the context of the normal ogive model (Takane & de Leeuw, 1987), but normality is not assumed here. Assumption MF2 specifies that the common parts η_i of the latent responses have a monotone relationship with the same underlying, more fundamental factors in $\boldsymbol{\Theta}$. These underlying factors should be independent. MF3 states that the unique factors or error variables ε_i are independent, which is comparable to conditional independence. Assumption MF4 is equivalent to the assumption that the distributions of ε_i are log-concave (e.g., Saumard & Wellner, proposition 2.3) or strongly unimodal (Walther, 2009, p. 320). This assumption is satisfied in many models, as this includes normal, uniform, gamma, beta, and logistic densities. Assumption MF5 requires a simple structure of the factor loadings.

Definition 3. *(monotone factor model).* $\mathbf{X}$ satisfies a *monotone factor model* (MFM) if there are $\boldsymbol{\lambda}$, $\boldsymbol{\Theta}$, and $\varepsilon_1, \varepsilon_2, \ldots$ such that MF1 – MF5 hold.

In other words, an MFM has factors that are independent, nonnegative loadings with a simple structure, strongly unimodal latent errors, and increasing response functions.

We will now define the concept of *multivariate positivity of order 2* (MTP_2) and related concepts. Let $\chi := \times_{i=1}^{m} \chi_m$ be a product lattice in R^m. Let the lattice operators be defined as, for

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

all $\mathbf{x}, \mathbf{y} \in \chi$:

$$\mathbf{x} \vee \mathbf{y} := (\max\{x_1, y_1\}, \ldots, \max\{x_m, y_m\})$$
$$\mathbf{x} \wedge \mathbf{y} := (\min\{x_1, y_1\}, \ldots, \min\{x_m, y_m\})$$

Definition 4. (MTP_2; Karlin and Rinott (1980)). A random vector $\mathbf{X}$ on product lattice χ is *MTP*$_2$ if it has density f, and for all $\mathbf{x}, \mathbf{y} \in \chi$,

$$f(\mathbf{x} \vee \mathbf{y}) f(\mathbf{x} \wedge \mathbf{y}) \geq f(\mathbf{x}) f(\mathbf{y})$$

MTP_2 generalizes the idea of a positive correlation and is also known as supermodularity, the FKG condition (Denuit et al. 2005) or affiliation (Milgrom & Weber, 1982). Denote the correlation between variables X and Y by ρ_{XY}.

Definition 5. *(nonnegative partial correlations, NPC;* Ellis (2014)*).* $\mathbf{X}$ has *nonnegative partial correlations* (NPC) if for every triplet (X, Y, Z) of variables in $\mathbf{X}$, $\rho_{XY} \geq \rho_{XZ}\rho_{ZY}$.

Definition 6. *(nonnegative covariances, NNC;* Mokken (1971)*).* $\mathbf{X}$ has *nonnegative covariance* (NNC) if $\text{Cov}(X_i, X_j) \geq 0$ for every pair (X_i, X_j) of variables in $\mathbf{X}$.

For any item i, we define the rest score as $X_{-i} = \left(\sum_{j=1}^{J} X_j\right) - X_i$; that is, the sum score with the score of item i omitted.

Definition 7. *(manifest monotonicity, MM;* Junker (1993)*).* $\mathbf{X}$ has *manifest monotonicity* (MM) if $\mathbb{E}(X_i|X_{-i})$ is increasing in X_{-i} for each $i = 1, \ldots, J$.

The following proposition is mostly implied by Corollary 3 of Ellis (2015), except that we do not require $\boldsymbol{\Theta}$ to be PF2 here.

Proposition A1. *If* $\boldsymbol{X}$ *satisfies a MFM, then* $\boldsymbol{X}$ *is MTP*$_2$.

Proof. This can be derived from three elementary facts about MTP_2 and PF2 variables: (1) independent variables are MTP_2 (Karlin & Rinott, 1980, proposition 3.5), (2) MTP_2 is preserved by increasing transformations (Karlin & Rinott, 1980, proposition 3.6; Ellis, 2015, proposition 8), and (3) the sum of MTP_2 and independent PF2 variables is MTP_2 (Karlin & Rinott, 1980, proposition 3.7). Consequently, assuming MF1-MF5, we obtain that $\boldsymbol{\Theta}$ is MTP_2 because it has independent components, $\boldsymbol{\eta}$ is MTP_2 because it is an increasing function of $\boldsymbol{\Theta}$, and $\boldsymbol{\eta}+\boldsymbol{\varepsilon}$ is MTP_2 because it is a sum of MTP_2 and independent PF2 variables. $\mathbf{X}$ is MTP_2 because it is an increasing transformation of $\boldsymbol{\eta} + \boldsymbol{\varepsilon}$. □

Appendix B

In this appendix, we will prove Theorem 1, which implies that MFMs satisfy MM. For ease of notation, we will first extend the definition of conditional expectations such as $\mathbb{E}\left(X_i | \sum_{j=1, j\neq i}^{J} X_j = r\right)$. Suppose R is a bounded nonnegative integer valued variable and X is a binary variable. If $P(R = r) = 0$, then $\mathbb{E}(X|R = r)$ is not uniquely defined, and we extend the definition of $\mathbb{E}(X|R = r)$ to all $r \in \mathbb{Z}$, including cases with $P(R = r) = 0$, such that

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

it remains increasing in $r \in \mathbb{Z}$, provided that it is increasing to begin with. The precise values are not relevant, and one possibility is defining

$$\mathbb{E}(X|R=r) := 0 \text{ if } P(R \leq r) = 0,$$
$$\mathbb{E}(X|R=r) := 1 \text{ if } P(R \geq r) = 0, \text{ and}$$
$$\mathbb{E}(X|R=r) := \frac{\mathbb{E}(X|R=r^-)+\mathbb{E}(X|R=r^+)}{2} \text{ if } P(R=r) = 0, P(R \leq r) > 0, \text{and } P(R \geq r) > 0$$

where $r^- = \sup\{s \in \mathbb{Z} | s < r, P(R=s) > 0\}$ and $r^+ = \inf\{s \in \mathbb{Z} | s > r, P(R=s) > 0\}$. In other words, if $\mathbb{E}(X|R=r)$ is undefined for some value of r, then we set it equal to 0 if $r < \min(R)$, equal to 1 if $r > \max(R)$, and equal to the average of the two surrounding values otherwise. Similarly, if S is another random variable, we can define $\mathbb{E}(X|R=r, S) := \mathbb{E}(X|R=r)$ in cases with $P(R=r) = 0$, and $\mathbb{E}(X|R+S=t, S=s)$ can be defined accordingly.

Lemma 1. *Let R and S be bounded nonnegative integer valued variables and let X be a variable with finite expectation. If $\mathbb{E}(X|R=r)$ is increasing in $r = 0, 1, \ldots$ and (X, R) is independent of S, then $\mathbb{E}(X|R+S=t)$ is increasing in $t = 0, 1, \ldots$.*

Proof. Since S is independent of X and R, we have for every $s = 0, 1, \ldots, \max(S)$,

$$\mathbb{E}(X|R+S=t+1, S=s) = \mathbb{E}(X|R=t+1-s, S=s) = \mathbb{E}(X|R=t+1-s) \geq \mathbb{E}(X|R=t-s) = \mathbb{E}(X|R=t-s, S=s) = \mathbb{E}(X|R+S=t, S=s)$$

Therefore,

$$\mathbb{E}(X|R+S=t+1) = \sum_{s=0}^{\max(S)} E(X|R+S=t+1, S=s) P(S=s) \geq \sum_{s=0}^{\max(S)} E(X|R+S=t, S=s) P(S=s) = \mathbb{E}(X|R+S=t)$$

□

Theorem 1. *Suppose that the set of test items $\{X_1, \ldots, X_J\}$ can be divided into disjoint subtests such that different subtests are independent while items within the same subtest satisfy MH. Then, MM holds for the entire set of test items, that is, $\mathbb{E}\left(X_i | \sum_{j=1, j\neq i}^{J} X_j = t\right)$ is increasing in t for each $i = 1, \ldots, J$.*

Proof. Consider an arbitrary item for which MM should be established. Without loss of generality, we may assume that this item is X_1 and that the first subtest is $(X_1, \ldots, X_k)$. With $R := \sum_{i=2}^{k} X_i$ and $S := \sum_{i=k+1}^{J} X_i$ we can write $\mathbb{E}\left(X_1 | \sum_{i=2}^{J} X_i = t\right) = \mathbb{E}(X_1|R+S=t)$, where X_1 and R are independent of S. Since $(X_1, \ldots, X_k)$ satisfies MH, it must also satisfy MM, that is, $\mathbb{E}(X_1|R=r)$ is increasing in r. Using Lemma 1, we obtain that $\mathbb{E}(X_1|R+S=t)$ is increasing in t. □

Theorem 2. *Suppose that the set of test items $\{X_1, \ldots, X_J\}$ can be divided into two disjoint subtests such that different subtests are independent while items within the same subtest satisfy MH. If two items X_i and X_j belong to different subtests, then $\mathbb{E}\left(Cov\left(X_i, X_j | \sum_{k\neq i,j}^{J} X_k\right)\right) \leq 0$.*

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

Proof. Without loss of generality, we can assume, for notational convenience, that the first subtest is $\{X_1, \ldots, X_{J_1}\}$ and the second subtest is $\{X_{J_1+1}, \ldots, X_J\}$, and that $i = 1$ and $j = J$. As a consequence of MM, $\mathbb{E}\left(X_1 | \sum_{k=2}^{J_1} X_k = t\right)$ is increasing in t. Now, apply Lemma 1 with $X = X_1$, $R = \sum_{k=2}^{J_1} X_k$ and $S = \sum_{k=J_1+1}^{J-1} X_k$. It follows that $\mathbb{E}\left(X_1 | \sum_{k=2}^{J-1} X_k = t\right)$ is increasing in t. Similarly, $\mathbb{E}\left(X_J | \sum_{k=2}^{J-1} X_k = t\right)$ is increasing in t too. Therefore, $Cov\left(\mathbb{E}\left(X_1 | \sum_{k=2}^{J-1} X_k\right), \mathbb{E}\left(X_J | \sum_{k=2}^{J-1} X_k\right)\right) \geq 0$. Since $Cov(X_1, X_J) = 0$ and, by the law of total covariance,

$$Cov(X_1, X_J) = \mathbb{E}\left(Cov\left(X_1,\ X_J \mid \sum_{k=2}^{J-1} X_k\right)\right) + Cov\left(\mathbb{E}\left(X_1 \mid \sum_{k=2}^{J-1} X_k\right), \mathbb{E}\left(X_J \mid \sum_{k=2}^{J-1} X_k\right)\right),$$

it follows that $\mathbb{E}\left(Cov\left(X_1,\ X_J \mid \sum_{k=2}^{J-1} X_k\right)\right) \leq 0$. □

This can be generalized to weighted sum scores, provided that MM still holds with respect to the weighted sum scores of the subtests, as stated in the following lemma and theorem.

Lemma 2. *Let R and S be real valued random variables with finite range and let X be a variable with finite expectation. If $\mathbb{E}(X|R = r)$ is increasing in r, $\forall r \in \mathbb{R}$, and (X, R) is independent of S, then $\mathbb{E}(X|R + S = t)$ is increasing in t, $\forall t \in \mathbb{R}$.*

Proof. Extend the definition of $\mathbb{E}(X|R = r)$ and $\mathbb{E}(X|R = r, S)$ to all $r \in \mathbb{R}$, just as we did prior to Lemma 1. Since S is independent of X and R, we have for every $s \in \mathbb{R}$, $\delta > 0$:

$$\mathbb{E}(X|R + S = t + \delta, S = s) = \mathbb{E}(X|R = t + \delta - s, S = s) = \mathbb{E}(X|R = t + \delta - s) \geq \mathbb{E}(X|R = t - s) = \mathbb{E}(X|R = t - s, S = s) = \mathbb{E}(X|R + S = t, S = s)$$

Therefore,

$$\mathbb{E}(X|R + S = t + \delta) = \sum_{s \in range(S)} \mathbb{E}(X|R + S = t + \delta, S = s) P(S = s) \geq \sum_{s \in range(S)} \mathbb{E}(X|R + S = t, S = s)\, P(S = s) = \mathbb{E}(X|R + S = t)$$

□

Theorem 3. *Suppose that the set of test items $\{X_1, \ldots, X_J\}$ can be divided into two disjoint subtests $\{X_1, \ldots, X_{J_1}\}$ and $\{X_{J_1+1}, \ldots, X_J\}$, such that different subtests are independent. Let $a_1, \ldots, a_J \in \mathbb{R}$ be such that both $\mathbb{E}\left(X_i \mid \sum_{k=2}^{J_1} a_k X_k = t\right)$ and $\mathbb{E}\left(X_j \mid \sum_{k=J_1+1}^{J-1} a_k X_k = t\right)$ are increasing in t. Then, $\mathbb{E}\left(Cov\left(X_1,\ X_J \mid \sum_{k=2}^{J-1} a_k X_k\right)\right) \leq 0$.*

Proof. By Lemma 2, $\left(X_1 | \sum_{k=2}^{J-1} a_k X_k = t\right)$ and $\mathbb{E}\left(X_J | \sum_{k=2}^{J-1} a_k X_k = t\right)$ are both increasing in t. Therefore, their covariance is nonnegative, i.e., $Cov\left(\mathbb{E}\left(X_1 | \sum_{k=2}^{J-1} a_k X_k\right), \mathbb{E}\left(X_J | \sum_{k=2}^{J-1} a_k X_k\right)\right) \geq 0$. The hypothesis of the theorem implies that $Cov(X_1, X_J) = 0$, and the conclusion follows by the law of total covariance. □

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

References

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*, 141–157. https://doi.org/10.1007/s11336-005-1376-9

Bartolucci, F., Farcomeni, A., & Scaccia, L. (2017). A nonparametric multidimensional latent class IRT model in a Bayesian framework. *Psychometrika, 82*, 952–978. https://doi.org/10.1007/s11336-017-9576-7

Bartolucci, F., & Forcina, A. (2000). A likelihood ratio test for MTP_2 within binary variables. *The Annals of Statistics, 28*, 1206–1218. https://doi.org/10.1214/aos/1015956713

Bartolucci, F., & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika, 70*, 31–43. https://doi.org/10.1007/s11336-001-0934-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*, 1165–1188. https://doi.org/10.1214/aos/1013699998

Brusco, M. J., Köhn, H. F., & Steinley, D. (2015). An exact method for partitioning dichotomous items within the framework of the monotone homogeneity model. *Psychometrika, 80*, 949–967. https://doi.org/10.1007/s11336-015-9459-8

Clarke, B., & Yuan, A. (2001). Manifest characterization and testing for certain latent properties. *Annals of Statistics, 29*, 876–898. https://doi.org/10.1214/aos/1009210693

Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2002). Testing unidimensionality in polytomous Rasch models. *Psychometrika, 67*, 563–574.

De Gooijer, J. G., & Yuan, A. (2011). Some exact tests for manifest properties of latent trait models. *Computational Statistics and Data Analysis, 55*, 34–44. https://doi.org/10.1016/j.csda.2010.04.022

Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2005). *Actuarial theory for dependent risks*. Wiley.

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*, 234–243. https://doi.org/10.1177/01466210122032046

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association, 105*(491), 1042–1055. https://doi.org/10.1198/jasa.2010.tm09129

Efron, B. (1965). Increasing properties of Pólya frequency function. *The Annals of Mathematical Statistics, 36*(1), 272–279. https://doi.org/10.1214/aoms/1177700288

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika, 79*, 303–316. https://doi.org/10.1007/s11336-013-9341-5

Ellis, J. L. (2015). MTP2 and partial correlations in monotone higher-order factor models. In Roger E. Millsap, Daniel M. Bolt, L. Andries van der Ark, en Wen-Chung Wang (Eds.), *Quantitative Psychology Research. The 78th Annual Meeting of the Psychometric Society* (pp. 261-272). Springer. https://doi.org/10.1007/978-3-319-07503-7_16

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62*, 495–523. https://doi.org/10.1007/bf02294640

Ellis, J. L., Pecanka, J., & Goeman, J. J. (2020). Gaining power in multiple testing of interval hypotheses via conditionalization. *Biostatistics, 21*(2), e65–e79. https://doi.org/10.1093/biostatistics/kxy042

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika, 81*, 434–460. https://doi.org/10.1007/s11336-014-9428-7

Fortuin, C. M., Kasteleyn, P. W., & Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics, 22*(2), 89–103. https://doi.org/10.1007/bf01651330

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525–546. https://doi.org/10.1007/BF02294405

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence, 15*(1), 79–103. https://doi.org/10.1016/0160-2896(91)90023-7

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Kluwer Nijhoff Publishing.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46*(1), 79–92. https://doi.org/10.1007/bf02293920

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543. https://doi.org/10.1214/aos/1176350174

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378. https://doi.org/10.1214/aos/1176349262

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*. https://doi.org/10.1214/aos/1069362751

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65–81. https://doi.org/10.1177/01466216000241004

Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *Journal of Multivariate Analysis, 10*(4), 467–498. https://doi.org/10.1016/0047-259x(80)90065-2

Kieftenbeld, V., & Nandakumar, R. (2015). Alternative hypothesis testing procedures for DIMTEST. *Applied Psychological Measurement, 39*(6), 480–493. https://doi.org/10.1177/0146621615577618

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. Academic Press.
Kuritz, S. J., Landis, J. R., & Koch, G. G. (1988). A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health, 9*(1), 123–160. https://doi.org/10.1146/annurev.pu.09.050188.001011
Ligtvoet, R. (2022). Incomplete tests of conditional association for the assessment of model assumptions. *Psychometrika*. https://doi.org/10.1007/s11336-022-09841-1
Ligtvoet, R., & Vermunt, J. K. (2012). Latent class models for testing monotonicity and invariant item ordering for polytomous items. *British Journal of Mathematical and Statistical Psychology, 65*, 237–250.
Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimen-sional contingency tables. *Psychometrika, 71*, 713–732. https://doi.org/10.1007/s11336-005-1295-9
McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.
Milgrom, P. R., & Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica, 50*(5), 1089. https://doi.org/10.2307/1911865
Mokken, R. J. (1971). *A theory and procedure of scale-analysis*. Mouton.
Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417–430. https://doi.org/10.1177/014662168200600404
Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion". *Applied Psychological Measurement, 10*, 279–285. https://doi.org/10.1177/014662168601000306
Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows*. A program for Mokken scale analysis for polytomous items, Groningen, The Netherlands: iecProGAMMA.
Rinott, Y., & Scarsini, M. (2006). Total positivity order and the normal distribution. *Journal of Multivariate Analysis, 97*, 1251–1261. https://doi.org/10.1016/j.jmva.2005.07.008
Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*(3), 425–435. https://doi.org/10.1007/bf02306030
Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1–30.
Sarkar, T. K. (1969). Some lower bounds of reliability. Tech. Report, No. 124, Dept. of Operations Research and Statistics, Stanford University.
Saumard, A., & Wellner, J. A. (2014). Log-concavity and strong log-concavity: A review. *Statistics Surveys, 8*, 45–113. https://doi.org/10.1214/14-SS107
Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.
Sijtsma, K., & Van der Ark, L. A. (2021). *Measurement models for psychological attributes*. Chapman and Hall/CRC.
Spearman, C. (1904). 'General intelligence', objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–292. https://doi.org/10.2307/1412107
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617. https://doi.org/10.1007/bf02294821
Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 19*, 331–354. https://doi.org/10.1177/014662169602000403
Stout, W., Froelich, A.G., Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In: Boomsma, A., van Duijn, M.A.J., Snijders, T.A.B. (eds) Essays on Item Response Theory. Lecture Notes in Statistics, vol 157. Springer, New York, NY. https://doi.org/10.1007/978-1-4613-0169-1_19
Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 75–99. https://doi.org/10.1007/s00357-013-9122-y
Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*, 117–123. https://doi.org/10.1027/1614-2241/a000115
Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408. https://doi.org/10.1007/bf02294363
Tijmstra, J., & Bolsinova, M. (2019). Bayes factors for evaluating latent monotonicity in polytomous item response theory models. *Psychometrika, 84*, 846–869. https://doi.org/10.1007/s11336-019-09661-w
Tijmstra, J., Hessen, D. J., van der Heijden, P. G. M., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statistical inference. *Psychometrika, 78*, 83–97. https://doi.org/10.1007/s11336-012-9297-x
Tversky, A., & Russo, J. E. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology, 6*(1), 1–12. https://doi.org/10.1016/0022-2496(69)90027-3
Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.
Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123–140. https://doi.org/10.1007/BF02296270
Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1–19.
Van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*, 1–27.
Van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519–538.
Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*, 283–294.

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press

Walther, G. (2009). Inference and modeling with log-concave distributions. *Statistical Science*. https://doi.org/10.1214/09-sts303

Wasserman, J. D. (2019). Deconstructing CHC. *Applied Measurement in Education, 32*, 249–268. https://doi.org/10.1080/08957347.2019.1619563

Whitt, W. (1980). Uniform conditional stochastic order. *Journal of Applied Probability, 17*, 112–123. https://doi.org/10.2307/3212929

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*(2), 129–152. https://doi.org/10.1007/bf02294532

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249. https://doi.org/10.1007/BF02294536

Manuscript Received: 19 JAN 2022
Published Online Date: 18 MAR 2023

https://doi.org/10.1007/s11336-023-09905-w Published online by Cambridge University Press