Measurement and Research Department Reports 2009-6

A Distribution-Based Association Measure and its Application in Dimensionality Assessment

Johan Braeken

Measurement and Research Department Reports 2009-6

A Distribution-Based Association Measure and its Application in Dimensionality Assessment

Johan Braeken

Tilburg University
Cito

Cito
Arnhem, 2009

Abstract

A test dimensionality assessment approach is proposed that builds upon existing approaches within non-parametric item response theory. The core of the procedure is a novel pairwise association measure based upon information theory and boundaries on bivariate distributions. Asymptotic results on the standard error of the measure allow to scan for anomalies in the pairwise item association matrix, allowing for the detection of serious local item dependence issues in the test. To assess the more general underlying dimensionality of the test a divisive clustering procedure is used to search for structure among the test items. A criterion that balances the homogeneity within clusters and the heterogeneity between clusters is suggested to select an optimal partitioning within the set of cluster solutions. The method is illustrated using a range of simulated test data under both strict and essential uni- and multidimensional conditions.

1. Introduction

After the construction of a test or assessment instrument it is important to have a way to assess its dimensionality. It is desirable to find a congruent mapping between the planned theoretical test structure and the empirically prominent test dimensions. Even when prior expectations about the test structure are absent, an assessment of dimensionality can provide useful insights in the test structure and in what exactly is measured by the test. This information can be helpful in deciding on reporting on overall or subscore level, and on item selection and final test composition. In this sense, test dimensionality is an important aspect with respect to test development and large scale test use.

Statistically, test dimensionality is formally defined in reference to local stochastic independence in a test response model. Let Y_{pi} $(i = 1, 2, \ldots, I)$ be the response of person p $(p = 1, 2, \ldots, P)$ to item i of the test, and let $\boldsymbol{\theta}_p$ be a vector of latent dimensions explaining shared variance among the item scores. Given the item-scores of person p on the test $(\boldsymbol{Y}_p)$ and the conditional item response function $(\Pr(Y_{pi} = y_{pi}|\boldsymbol{\theta}_p))$, the vector $\boldsymbol{\theta}_p$ satisfying local stochastic independence (LSI),

$$\Pr(\boldsymbol{Y}_p = \boldsymbol{y}_p|\boldsymbol{\theta}_p) = \prod_{i=1}^{I} \Pr(Y_{pi} = y_{pi}|\boldsymbol{\theta}_p), \tag{1}$$

is considered to represent the dimensionality of the test. To ensure that this assumption has observable and testable consequences (see e.g., Suppes & Zanotti, 1981), a restricting condition needs to be defined on the conditional response functions. Traditionally, the monotonicity condition is added, in which the probability of a higher value on Y_{pi} increases with higher values on the latent dimensions.

Conceptually, all this means that $\boldsymbol{\theta}_p$ is a summary of the information provided by a persons item responses, on the relative test performance of a person in the population. Notice the reference to the test population; the dimensionality of a test

arises due to the heterogeneity resulting from the interaction between test takers and test items. Consider a test comprised of multiple items, requiring a varying level of two dimensions, algebra-knowledge and reading-ability, for each item. If the test population is, for instance, homogeneous with respect to reading-ability, only the algebra dimension will be picked up statistically. In a similar way, when each item in fact requires the exact same levels of both algebra-knowledge and reading-ability; again only one dimension (i.e., the composite ability) will be picked up statistically (Ackerman, 1994; Reckase, Ackerman, & Carlson, 1988). Hence, in both psychometric traditions, classical test theory and item response theory, it is important to define a priori the test population of interest.

Mathematically the dimensionality issue seems to be clear-cut, yet in practice there are some complications. The rigorous yet strict statistical conditions defining test dimensionality are unlikely to be obtainable in most applications, and even if such an ideal case situation exists, a numerically optimal dimensionality structure does not guarantee that it is substantive meaningful or relevant in practice. Furthermore, one has to acknowledge the possibility of statistically equivalent models in a multidimensional context; think of the many rotated factor solutions that exist in the case of classic factor analysis. Hence, there are still aspects of choice and judgement involved in the assessment of test dimensionality.

From a pragmatic perspective, interest only goes out to the more important, "dominant", dimensions among those measured by a test, making abstraction of the minor, "noise", dimensions (for a strong stance on this dimensionality topic, see e.g, Humphreys, 1984). This implies that the focus is on a vector $\boldsymbol{\theta}_p$ that only needs to approximate the above-mentioned statistical conditions, and that minor violations of LSI for specific items are tolerated. Connecting these dominant dimensions to substantive meaningful aspects is usually also more obvious than trying to give

meaning to noise.

This common-sense idea fostered some theoretical formal work in non-parametric item response theory, leading to the definition of essential dimensionality (Stout, 1987, 1990; Junker, 1993). Basicly, this relaxed dimensionality definition applies a weaker form of local stochastic independence, roughly described as the requirement that conditional upon the latent dimensions the expected value of the covariances between items tends to zero when the item pool grows to infinity. Together with work on conditions for strict dimensionality (see e.g., Rosenbaum, 1984; Holland & Rosenbaum, 1986) and work on test homogeneity (see e.g., Loevinger, 1948; Mokken, 1971), the formal work on the statistical definition of dimensionality in latent variable models has provided useful instrumental guidelines for test dimensionality assessment in practice.

In general, a common starting point for dimensionality assessment is a matrix of conditional or unconditional pairwise item association measures. The patterns in these item associations can be utilized to derive the underlying dimensionality structure (see e.g., Roznowski, Tucker, & Humphreys, 1991; Kim, 1994). For instance in Mokken scaling, as implemented in MSP (Molenaar & Sijtsma, 2000) or in R (Ark, 2007), the Loevinger coefficient, which is a normed unconditional covariance, and related aggregated coefficients play an important role in partitioning an item set in homogeneous subsets. This approach primarily makes use of the monotonicity condition on the item response functions. In contrast, conditional covariances are used in the procedures developed by Stout and colleagues (DETECT; Zhang & Stout, 1999a, DIMTEST; Stout, 1987; Nandakumar & Stout, 1993,HCA-CCPROX; Roussos, Stout, & Marden, 1998, circular MDS; Bolt, 2001), making use of the local independence condition to assess test dimensionality. All these approaches result in a partitioning of the total item pool into homogeneous item subsets, which are a

reflection of the aforementioned dominant dimensions in the data. Thus, the fact that items within a subset are so similar, is due to their common denominator, being the most dominant dimension they are measuring. Note that these non-parametric methods only require a limited amount of assumptions and are computationally not too demanding. The focus is on the dominant dimensions in the data, and no further assumptions are made about the exact relationship between items and dimensions. This is also the reason why the more intensive and restrictive parametric methods such as confirmatory and exploratory factor analysis for categorical data (see e.g., Bock, Gibbons, & Muraki, 1988; Bartholomew & Knott, 1999) are disregarded. For a more in depth overview of the different approaches, the interested reader is directed to reviews by Tate (2003) and Van Abswoude, Ark, and Sijtsma (2004).

The dimensionality assessment approach proposed in this paper, is based upon information theory and boundaries on bivariate distributions, and will indirectly make use of, modify, and integrate some features and aspects of existing non-parametric approaches. The starting point is again a pairwise item conditional association matrix, also on the local stochastic independence condition (cfr. DETECT), but now based upon a measure labeled "signed information rate" (SIR). This new pairwise association measure between two random variables U and V is build up in terms of their probability distributions, and only takes a 0 value ($SIR = 0$) if and only if their observed joint distribution $\Pr(U, V)$ coincides with the expected joint distribution under independence $\pi(U, V)$. This in contrast to the traditional covariance in which $COV(U, V) = 0$ does not necessarily imply that U and V are independent, because covariance only measures linear dependence. Thus, the new measure SIR gives a more accurate and less restrictive picture of the pairwise association. To account for the influence of the discrete nature of the item responses on association measures and to allow for better comparison, the SIR measure is

normed with respect to the marginal distributions of the random variables (cfr., Mokken scaling, and see also Section 2.1.2), and direction of the association is determined by the relative position of the observed joint distribution $\Pr(U, V)$ in its limiting boundary space. The standard error of the signed information rate provides a means to assess which element of the pairwise item association matrix can be regarded as an extreme outlier, that can be expected to have too much influence on the direction a dimensionality assessment procedure will take for the given item pool. Because the SIR measure can readily be transformed into a distance measure, a hierarchical clustering procedure (cfr. HCA-CCPROX) can be used to construct an item tree. This item tree offers a range of partitioning solutions which can be evaluated in terms of a dimensionality criterion (cfr. DIMTEST). The criterion used is formulated in terms of heterogeneity between partitions and homogeneity within a partition, leading to the selection of a clearly expressed structure corresponding to the main target of the procedure, that is the dominant dimensions underlying the test.

2. Method

The proposed method makes use of a necessary condition implied by LSI. If LSI holds, each item pair should also be independent conditional on the latent trait(s). In practice, this pairwise LSI is more feasible to verify than strict LSI. Although in principle pairwise LSI is merely a necessary and not a sufficient condition for LSI, it is claimed to be sufficient in general practice (see e.g., McDonald & Mok, 1995). Inspecting a $I \times I$ matrix of conditional pairwise item association measures can give a good insight in whether pairwise LSI (and almost certain strict LSI) holds for the data at hand and is a good starting point to assess the dimensionality structure. To construct such a conditional item association measure for an item pair i and j, a few

choices have to be made. The first choice is to determine the conditioning factor. To ensure manifest monotonicity of the empirical conditional response functions (see e.g., Junker & Sijtsma, 2000), the rest score $Y_{p+}^{-i,j} = \sum_{k \neq i,j}^{I} Y_{pk}$ is given this status. This conditioning factor functions as an approximative proxy of the latent traits underlying the test $\boldsymbol{\theta}_p \approx Y_{p+}^{-i,j}$, and can be regarded as a composite trait. The data are grouped according to a binning procedure utilizing $Y_{p+}^{-i,j}$. These grouped data allow for the construction of non-parametric conditional item response functions and of a conditional pairwise association measure, which will form the fundaments for the further dimensionality assessment procedure.

2.1. A Signed Mutual Information Rate

The chosen association measure is an adapted form of mutual information, normed according to the conditional item response functions and modified based upon limiting bounds to the joint conditional item response function. To give the mathematical foundations of the association measure, some less familiar material, originating from information theory, a branch of applied mathematics and engineering (Shannon & Weaver, 1949; Cover & Thomas, 2006), and from the study on distributions with fixed margins (Fréchet, 1951; Hoeffding, 1940) needs to be introduced. The basic building blocks are presented first, gradually making the transition to the final measure.

2.1.1. Mutual Information

The mutual information between two categorical variables U and V measures the amount of information obtainable about one variable by observing the other, and is given in the following expression

$$I(U,V) = H(U) + H(V) - H(U,V),$$

in which $H(X)$ is the entropy or uncertainty around the random variable X, where the function H computes minus the expected value of the log probability of a realization of X as

$$H(X) = -E\left[\log(\Pr(X))\right] = -\sum_{x \in \Omega(X)} \Pr(X = x) \log(\Pr(X = x)),$$

where the sum is over all possible unique realizations x in the outcome space $\Omega(X)$ of the random variable. Conceptually, when both variables U and V are independent, observing one or both of the variables at the same time does not make a difference on the amount of information gained on an individual variable, and hence there is no reduction in uncertainty. In this case, their joint entropy $H(U, V)$ is simply the sum of their self entropies $H(U) + H(V)$, such that $I(U, V) = 0$. In contrast, when both variables are dependent, information can be gained from observing U and V simultaneously, and the uncertainty around their paired observations reduces, hence $H(U, V) < H(U) + H(V)$ and $I(U, V) > 0$.

Moreover, mutual information $I(U, V)$ can also be motivated in terms of a Kullback-Leibler Divergence $\bigtriangledown$,

$$\begin{aligned} I(U, V) &= \bigtriangledown\left(\Pr(U, V) || \pi(U, V)\right) \\ &= E\left[\log\left(\frac{\Pr(U, V)}{\pi(U, V)}\right)\right] \\ &= \sum_{u \in \Omega(U), v \in \Omega(V)} \Pr(U = u, V = v) \log\left(\frac{\Pr(U = u, V = v)}{\pi(U = u, V = v)}\right). \end{aligned}$$

Thus, it can be seen as a measure of how close the joint distribution of U and V, $\Pr(U, V)$, is to their expected joint distribution under independence, $\pi(U, V)$. Furthermore, because mutual information only equals 0 if and only if $\Pr(U, V)$ and $\pi(U, V)$ exactly overlap, it captures all dependencies between the two random variables, not just second-order dependencies as captured by the covariance for instance. Reformulating mutual information as a divergence connects it to the more common

likelihood ratios in a maximum likelihood framework.

2.1.2. Normalizing to a Mutual Information Rate

It is well known that for categorical variables the exact expression of the dependence structure is partly determined by the marginal distributions of the individual variables. A typical example can be found when looking at the attainable limits of the product-moment correlation of 2 Bernoulli distributed random variables, $U \sim Bern(p_u)$ and $V \sim Bern(p_v)$, these are not simply the traditional $[-1, 1]$ bounds, but depend on the marginal distributions as parametrized by p_u and p_v (see e.g. Cureton, 1959; Joe, 1997, p. 210). Similar considerations also motivated the development of the Loevinger coefficient (Loevinger, 1948), an association measure used in the Mokken scaling approach (Mokken, 1971) to dimensionality assessment. Note that Warrens (2008) provides a discussion of a whole range of association measures with respect to some margin-independent properties (i.e., properties that are in general true, and not only applicable for a specific dataset).

To take into account the influence of the marginal distributions of U and V on the assocation structure, the mutual information measure will be normed with reference to the sum of information provided by the individual variables, resulting in a mutual information rate, defined as

$$IR(U,V) = \frac{I(U,V)}{H(U) + H(V)}$$

$$= \frac{H(U) + H(V) - H(U,V)}{H(U) + H(V)},$$

where in case of independence $IR(U,V) = 0$, and $0 < IR(U,V) \leq 1$ otherwise.

2.1.3. Determining Directionality

A disadvantage of this association measure is the lack of direction; $IR(U,V)$ makes no difference between positive or negative dependencies. A novel signed version of this measure $SIR(U,V)$ is proposed based upon the Frechet-Hoeffding bounds of bivariate distributions with given margins (Fréchet, 1951; Hoeffding, 1940). The probability distribution space of bivariate cumulative distribution functions (cdf) $F(U = u, V = v) = \Pr(U \leq u, V \leq v)$ with given margins $F(U = u) = \Pr(U \leq u)$ and $F(V) = \Pr(V \leq v)$, can be defined using 3 essential distributions,

$$\Pi(U,V) = F(U)F(V);$$

$$W(U,V) = max(F(U) + F(V) - 1, 0);$$

$$M(U,V) = min(F(U), F(V)).$$

The cdf Π is recognized as the bivariate cdf for the given margins $F(U = u) = \Pr(U \leq u)$ and $F(V) = \Pr(V \leq v)$ when U and V would be independent; the cdf W is then the bivariate distribution corresponding to absolute negative dependence between U and V, and the cdf M is the bivariate distribution corresponding to absolute positive dependence between U and V. For all possible bivariate distributions given margins $F(U = u) = \Pr(U \leq u)$ and $F(V) = \Pr(V \leq v)$ it holds that

$$W(U,V) < \Pi(U,V) < M(U,V),$$

$$W(U,V) \leq F(U,V) \leq M(U,V),$$

such that together these 3 cdfs define the limiting boundaries of $F(U = u, V = v) = \Pr(U \leq u, V \leq v)$.

Using these boundary conditions, the observed joint probability $\Pr(U,V)$ can be located in $\Omega|[\Pr(U), \Pr(V)]$, the entire probability distribution space given fixed univariate margins $\Pr(U)$ and $\Pr(V)$. The translation of the limiting bounds from

cdf's to pdf's can easily be done using the following recursive formula based upon quadrant probabilities (see e.g., Mood, Graybill, & Boes, 1974)

$$Pr(U = u, V = v) = \sum_{k_1=0}^{1} \sum_{k_2=0}^{1} (-1)^{k_1+k_2} CDF(U = u - k_1, V = v - k_2),$$

where CDF corresponds to the joint cumulative probability. If $k_1 = 0$ then $F(U = u - k_1) = F(U = u)$, and if $k_1 = 1$ then $F(U = u - k_1)$ is the cdf of a realization of U that falls into the next-lower ordinal category of U. Note that when u already falls in the lowest category, by definition of a cdf, $F(U = u - k_1) = F(-\infty) = 0$.

To take into account that the probability distribution space $\Omega|[\Pr(U), \Pr(V)]$ is not necessarily symmetrically centered around the independence case $\pi(U, V)$, normalized divergences are used to assess the relative position of the observed joint distribution $\Pr(U, V)$ in $\Omega|[\Pr(U), \Pr(V)]$. These normalized divergences Z_∇ are defined as the ratio of the divergence between the joint distribution and a limiting bound and the divergence between the independence distribution and that same bound:

$$Z_\nabla(w) = \frac{\nabla\left(\Pr(U, V)||w(U, V)\right)}{\nabla\left(\pi(U, V)||w(U, V)\right)}$$
$$Z_\nabla(m) = \frac{\nabla\left(\Pr(U, V)||m(U, V)\right)}{\nabla\left(\pi(U, V)||m(U, V)\right)}.$$

Comparing these normalized divergences then allows for the assignment of direction to the mutual information rate $IR(U, V)$ in the following way:

$$Z_\nabla(w) < Z_\nabla(m) \Rightarrow sign(IR(U, V)) = -1;$$
$$Z_\nabla(w) = Z_\nabla(m) \Rightarrow sign(IR(U, V)) = 0;$$
$$Z_\nabla(w) > Z_\nabla(m) \Rightarrow sign(IR(U, V)) = 1.$$

Using these results, the novel signed and normed association measure, labeled

Signed Information Rate ($SIR(U,V)$) is defined as

$$SIR(U,V) = sign\left(Z_{\triangledown}(w) - Z_{\triangledown}(m)\right)\frac{I(U,V)}{H(U)+H(V)}.$$

2.1.4. conditional SIR

A conditional version of this association measure can easily be obtained by making use of conditional entropies and Kullback Leibler Divergences between conditional probabilities, such that for an item pair Y_{pi} and Y_{pj} the conditional association measure is formulated as

$$SIR(Y_{pi},Y_{pj}|Y_{p+}^{-i,j}) = sign\left(Z_{\triangledown}(w) - Z_{\triangledown}(m)\right)\frac{I(Y_{pi},Y_{pj}|Y_{p+}^{-i,j})}{H(Y_{pi}|Y_{p+}^{-i,j})+H(Y_{pj}|Y_{p+}^{-i,j})} \qquad (2)$$

where a conditional entropy can be computed as the expected value of the entropy given a realization of $Y_{p+}^{-i,j}$

$$\begin{aligned}
&H(Y_{pi}|Y_{p+}^{-i,j})\\
&\quad= -\sum_{y_{p+}^{-i,j}}\Pr(Y_{p+}^{-i,j}=y_{p+}^{-i,j})\sum_{y_{pi}}\Pr(Y_{pi}=y_{pi}|Y_{p+}^{-i,j})\log\left(\Pr(Y_{pi}=y_{pi}|Y_{p+}^{-i,j})\right),\\
&\quad= -\sum_{(y_{pi}\cap y_{p+}^{-i,j})}\Pr(Y_{pi}=y_{pi}\cap Y_{p+}^{-i,j}=y_{p+}^{-i,j})\log\left(\Pr(Y_{pi}=y_{pi}\cap Y_{p+}^{-i,j}=y_{p+}^{-i,j})\right)
\end{aligned}$$

and a similar construction holds for the Kullback Leibler divergence $\triangledown$ (The outcome spaces Ω were left out for notational parsimony).

This signed conditional information rate $SIR(Y_{pi},Y_{pj}|Y_{p+}^{-i,j})$ is symmetric $SIR(U,V) = SIR(V,U)$, applicable for both binary and polytomuously scored items, only takes zero-value in case of conditional independence, and is normed with reference to the information provided by the 2 items when conditional independence would hold. These properties result in ease of use and interpretation. Furthermore, the Signed Information Rate has an intuitive meaning in the context of local stochastic independence. Once one knows someone's position on the latent dimensions $\boldsymbol{\theta}_p$, here approximated by $Y_{p+}^{-i,j}$, one can not learn anything more about the response to one

item from the response to the other item. Hence, these items are not assumed to share more information than already contained in $Y_{p+}^{-i,j}$. If there still is excess shared information present, this can be used to determine the general test dimensionality structure; negative information indicates that both items are probing a different aspect, while positive information indicates them to measure a similar aspect.

2.1.5. Standard error of SIR

Derivation of an approximate standard error for SIR is based upon the delta method and the observation that the essential sample data underlying the statistic are actually the n_{suv}'s, the counts (see Roulston, 1999) of persons for which hold that $Y_{pi} = u$, $Y_{pj} = v$, and $Y_{p+}^{-i,j} = s$. Based upon these counts the conditional response probabilities are estimated as

$$\begin{aligned} \Pr(Y_{pi} = u | Y_{p+}^{-i,j}) &= \sum_{v \in \Omega(Y_{pj})} n_{suv}/N_s \\ \Pr(Y_{pj} = v | Y_{p+}^{-i,j}) &= \sum_{u \in \Omega(Y_{pi})} n_{suv}/N_s \\ \Pr(Y_{pi} = u, Y_{pj} = v | Y_{p+}^{-i,j}) &= n_{suv}/N_s, \end{aligned} \tag{3}$$

with N_s the number of persons with $Y_{p+}^{-i,j} = s$.

Given that $n_{suv} \sim Bin(N_s, \Pr(u, v | Y_{p+}^{-i,j}))$, its sampling variance equals

$$\begin{aligned} VAR(n_{suv}) &= N_s \Pr(u, v | Y_{p+}^{-i,j})(1 - \Pr(u, v | Y_{p+}^{-i,j})) \\ &= n_{suv} - n_{suv}^2/N_s. \end{aligned}$$

If we disregard the dependence direction and consider $\mu(x) = abs(SIR(Y_{pi}, Y_{pj} | Y_{p+}^{-i,j})) =$

$IR(Y_{pi}, Y_{pj}|Y_{p+}^{-i,j}))$ as the mean of x_{suv} with

$$h_{su} = -\log\left(\sum_{v\in\Omega(Y_{pj})} n_{suv}/N_s\right)$$

$$h_{sv} = -\log\left(\sum_{u\in\Omega(Y_{pi})} n_{suv}/N_s\right)$$

$$h_{suv} = -\log\left(n_{suv}/N_s\right)$$

$$x_{suv} = \frac{h_{su} + h_{sv} - hsuv}{h_{su} + h_{sv}}$$

then an approximate standard error can be computed as

$$SE\left[SIR(Y_{pi}, Y_{pj}|Y_{p+}^{-i,j})\right] = \sqrt{\sum_{u\in\Omega(Y_{pi})}\sum_{v\in\Omega(Y_{pj})}\sum_{s\in\Omega(Y_{p+}^{-i,j})}(x_{suv} - mu(x))^2}.$$

The standard error of the signed information rate provides a means to assess which element of the the $I \times I$ pairwise item association matrix $\boldsymbol{SIR}$ can be regarded as an outlier and might point at the presence of local item dependency similar to the problem of multi-colinearity in multiple regression, rather than to be considered a general problem at the underlying dimensionality level. Using a common standard normally distributed Z-test it can be checked whether the observed $SIR(i, j|\theta)$ is a statistically extreme value. The test statistics is defined as

$$Z = \frac{SIR(Y_{pi}, Y_{pj}|Y_{p+}^{-i,j})}{SE\left[SIR(Y_{pi}, Y_{pj}|Y_{p+}^{-i,j})\right]} \sim N(0, 1). \tag{4}$$

Hence, a usefull first step in assessing the dimensionality of the test might be to detect outliers, that show such an extreme excess item interdependence. In practice one item of such a detected item pair could be argued to be redundant, and it could be opted to drop one such item in favour of test efficiency. If not dropped, it should be explicitly modeled to prevent distortion of test reliability and other model aspects and parameters.

2.2. Divisive Hierarchical Clustering

When searching for a structure of dominant dimensions, a hierarchical approach appears a natural choice. Using the $I \times I$ $\boldsymbol{SIR}$ matrix of pairwise conditional association measures $SIR(Y_{pi}, Y_{pj}|Y_{p+}^{-i,j})$ as starting point, a divisive (i.e., top-down) clustering algorithm will be adopted in search for hierarchical cluster structure among the items. For the purpose of clustering, the signed mutual information matrix $\boldsymbol{SIR}$ is transformed to a dissimilarity matrix $\boldsymbol{D} = (1 - \boldsymbol{SIR})/2$. Note that each element d_{ij} of the matrix $\boldsymbol{D}$ satisfies the properties of a distance metric,

$$
\begin{aligned}
d_{ij} &\geq 0 && \text{(non-negativity);} \\
d_{ij} &= 0 \Leftrightarrow i = j && \text{(identity of indiscernibles);} \\
d_{ij} &= d_{ji} && \text{(symmetry);} \\
d_{ij} &<= d_{ik} + d_{kj} && \text{(triangle inequality).}
\end{aligned}
$$

Hence, $\boldsymbol{D}$ is indeed appropriate for use in cluster analysis. Divisive cluster methods start by considering the whole set of items as one cluster, and then split up the set into successive subclusters until each object is a singleton cluster. There are $I - 1$ successive splitting steps to be made. In each of these steps the cluster with the largest dissimilarity (within the cluster) is selected and then split into two new clusters. A variant of the method of MacNaughton-Smith, Williams, Dale, and Mockett (1964) is used, which recursively repositions items from the start group, initially equivalent with the original cluster, into a new splinter group based upon their average distances with respect to all other items in these two groups, until no improvement can be made (see also, Kaufman & Rousseeuw, 1990). The resulting cluster hierarchy can be represented graphically by means of a divisive tree-diagram or dendrogram, in which the stem represents the entire item set, and where a branch

splits at a vertical coordinate corresponding to the diameter of that cluster before splitting.

Divisive (top-down) algorithms are often more accurate than their more commonly known agglomerative (bottom-up) counterparts. Bottom-up clustering methods, as for instance used in HCA-CCPROX, make clustering decisions based on local patterns without initially taking into account the full dissimilarity matrix, increasing the risk of running into sub-optimal solutions. In contrast, top-down clustering methods involve a bit more computation at each step, but benefit from complete information about the full dissimilarity matrix when making top-level partitioning decisions. If one is only interested in a limited amount of clusters, a top-down algorithm can terminate early, while a bottom-up algorithm needs to go through the whole tree (and do the computations) to reach that same top-level. Both algorithms remain prone to errors due to their greedy nature (mistakes in an earlier step can not be undone in a subsequent step of the hierarchy).

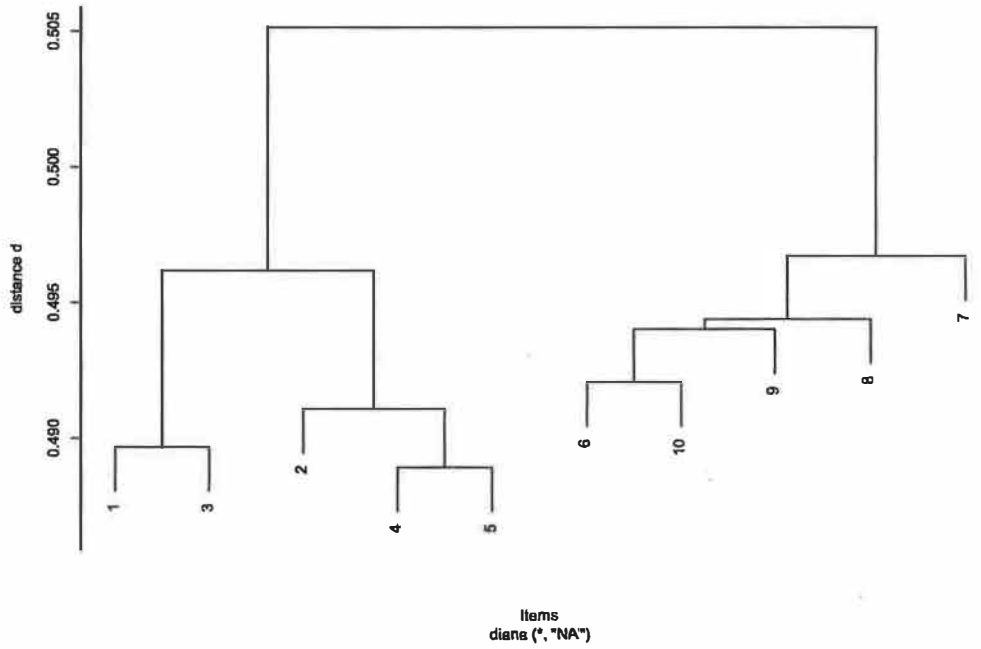
2.2.1. Dimensionality criterion

Although a cluster hierarchy is informative, it is not the end goal in practice. The end goal is an optimal non-hierarchical dimensional representation of the item set; optimal in the sense of demarcating the dominant dimensions in the item responses, yet making abstraction of the minor noise dimensions (cfr., Introduction). To formalize this, a criterion is proposed that balances between-cluster heterogeneity and within-cluster homogeneity, to allow the selection of that solution in the hierarchy with the most outspoken item partitioning. A clear partitioning clarifies what is exactly measured by the test and allows for straightforward test scoring and communication.

Thus, the hierarchical cluster tree need to be cut at a certain level, resulting in

FIGURE 1.

Hierarchical cluster tree



a partitioning $\wp(K) = \{C_1, C_2, \ldots, C_K\}$, with C_k an item cluster. It is proposed to evaluate the I possible partitioning solutions $\mathbb{S}(K)$, corresponding to the initial set of items $\mathbb{S}(K=1)$ and the $I-1$ successive solutions of the hierarchical algorithm, and select the solution $\mathbb{S}(K)$ maximizing the following criterion:

$$\Psi(\mathbb{S}(K)) = \frac{\bar{d}_{between}}{\bar{d}_{within}}|\mathbb{S}(K) - \frac{\bar{d}_{between}}{\bar{d}_{within}}|\mathbb{S}(1)$$

where

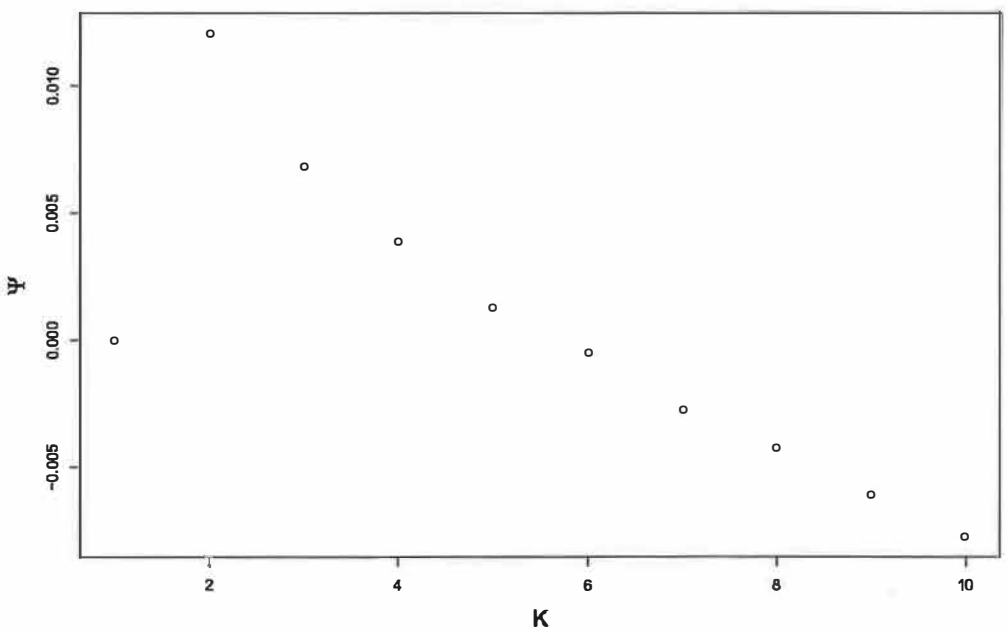
$$\bar{d}_{within} = \frac{\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} 1\left[\exists\{i,j\} \subset C_k\right] d_{ij}}{\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} 1\left[\exists\{i,j\} \subset C_k\right]}$$

$$\bar{d}_{between} = \frac{\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} 1\left[\nexists\{i,j\} \subset C_k\right] d_{ij}}{\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} 1\left[\nexists\{i,j\} \subset C_k\right]}.$$

For $K = 1$ $\bar{d}_{between} = 0.5$ corresponding to the independence situation (i.e., $SIR = 0$

FIGURE 2.

Ψ-plot over the I possible partitioning solutions $\mathbb{S}(K)$



and $D = (1 - SIR)/2 = 0.5$) and such that $\Psi(\mathbb{S}(1)) = 0$. The criterion favors homogeneous clusters (i.e., small $\bar{d}_{within}$) that are well separated (i.e., large $\bar{d}_{between}$). Note that only a statistically 'nice' multidimensional solution $\mathbb{S}(K)$) gives rise to a positive value of $\Psi[\mathbb{S}(K)]$ (i.e., between-cluster distance is larger than twice the average distance within the total item set).

3. Example

Consider the following example test consisting of 10 items generated under a compensatory multidimensional model (see e.g., McKinley & Reckase, 1982)

$$\Pr(Y_{pi} = 1|\boldsymbol{\theta}_p) = \frac{\exp(\sum_{k=1}^{K} \alpha_{ik}\theta_{pk} - \beta_i)}{1 + \exp(\sum_{k=1}^{K} \alpha_{ik}\theta_{pd} - \beta_i)}. \quad (5)$$

The item difficulties β_i follow a standard normal distribution. The loadings $\boldsymbol{\alpha}$ are chosen such that the first 5 items only measure a relatively stronger first dimension ($\alpha_{i1} \sim N(1.1, 0.05)$ and $\alpha_{i2} = 0\ \forall i \in \{1, \ldots, 5\}$), while the next 5 items only measure a relatively weaker second dimension ($\alpha_{i2} \sim N(0.9, 0.05)$ and $\alpha_{i1} = 0$ $\forall i \in \{6, \ldots, 10\}$). The latent traits $\boldsymbol{\theta_p} = \{\theta_{p1}, \theta_{p2}\}$ are uncorrelated and standard normally distributed. Notice that this resembles the dominant dimensions versus noise dimensions idea. Data was generated for $P = 1000$ persons.

The first step in the proposed procedure is to construct a pairwise item association matrix making use of the non-parametric conditional response functions given in Equation 3. Note that one could also work with smoothed versions of these functions (see e.g., Ramsay, 1991; Habing, 2001). The upper-triangle of the resulting pairwise item association matrix $\boldsymbol{SIR}$ (Equation 2) together with the corresponding lower-triangle of p-values according to the Z-test in Equation 4 is presented in the matrix below

$$[up.tri(\boldsymbol{SIR}) + low.tri(\boldsymbol{p})] =$$

$$\begin{bmatrix}
. & 0.022 & 0.021 & 0.020 & 0.019 & -0.005 & -0.003 & -0.007 & -0.003 & -0.004 \\
.036 & . & 0.008 & 0.021 & 0.018 & -0.003 & 0.001 & -0.004 & -0.003 & -0.004 \\
.041 & .251 & . & 0.021 & 0.010 & -0.005 & -0.002 & 0.006 & -0.004 & -0.002 \\
.045 & .037 & .037 & . & 0.022 & -0.007 & -0.005 & -0.008 & -0.010 & -0.005 \\
.054 & .064 & .181 & .031 & . & -0.006 & -0.003 & -0.007 & -0.005 & -0.005 \\
.334 & .407 & .324 & .267 & .301 & . & 0.010 & 0.012 & 0.012 & 0.016 \\
.380 & .459 & .426 & .331 & .395 & .201 & . & 0.007 & 0.007 & 0.007 \\
.271 & .354 & .307 & .236 & .278 & .150 & .282 & . & 0.011 & 0.011 \\
.388 & .395 & .369 & .182 & .322 & .146 & .286 & .158 & . & 0.012 \\
.374 & .346 & .419 & .330 & .339 & .091 & .263 & .162 & .137 & .
\end{bmatrix}$$

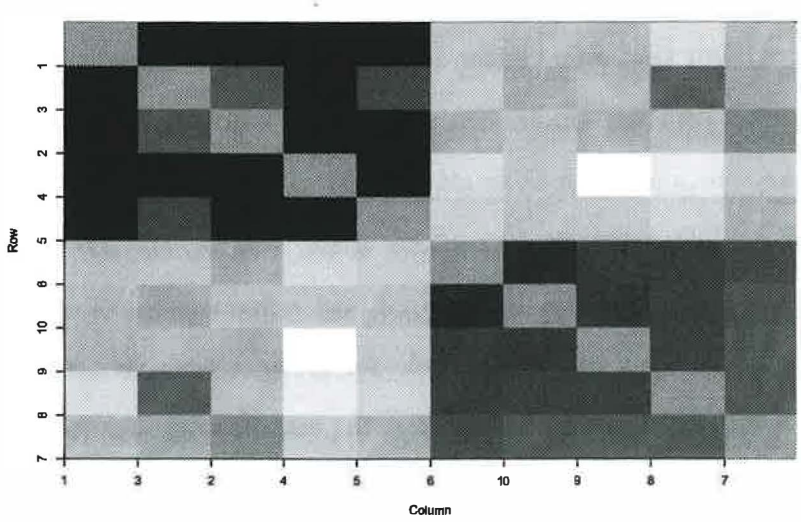
Visual inspection of this matrix shows that the items might be measuring different aspects (see e.g., the sign pattern). Furthermore some item pairs are evaluated as showing rather extreme local dependence ($p < 0.05$) given the test composite $Y_{p+}^{-i,j}$; although adopting a more strict significance level to take into account the multiple testing would weaken this observation. In any case, it are signs that a unidimensional structure might not be adequate for this test.

A second step is to search for structure by means of a divisive clustering proce-

dure. The resulting hierarchical tree is shown in Figure 1, and a plot of the proposed partitioning criterion $\Psi(\mathbb{S}(K))$ over the I possible partitioning solutions in this tree is given in Figure 2. It can be seen that the 2-cluster solution ($K = 2$) is to be preferred, but that even cluster solutions up to $K = 5$ might be considered a better representation than unidimensionality $K = 1$ in statistical terms (Beyond 5 clusters the solutions loose their attractivity in terms of the between-within criterion $\Psi(\mathbb{S}(K))$). When a strong structure arises, this should also be visible in a sorted contrast version of $\boldsymbol{SIR}$. In Figure 3, the 2-dimensional structure clearly surfaces. Darker colors of an $\{i, j\}$ entry indicate stronger positive interdependence between items i and j of the test. Notice that the matrix rows and columns are sorted such that items belonging to the same cluster are next to each other.

Aggregated pairwise SIR measures similar to $\bar{d}_{\boldsymbol{within}}$ and $\bar{d}_{\boldsymbol{between}}$ can be utilized as a relative description of the homogeneity within a cluster. For instance, the mean SIR value for the item pairs belonging to cluster 1 equals 0.015, while for cluster 2 this is 0.008, and hence this confirms the simulation design in which dimension 1 was relatively stronger than dimension 2. In this way, the $\boldsymbol{SIR}$ matrix can be summarized in a set of structural informative indices, that can be used for further situating the dimensionality assessment results. For instance, to further refine a given dimension, one might consider using a similar approach to item selection as in Mokken scaling, but now based upon an item-aggregated SIR measure, instead of the usual scalability coefficient. However, note that, especially in small item clusters, removing an item might have large consequences, not only for the cluster at hand, but also for the quality of the overall partitioning.

FIGURE 3.
Sorted contrast ***SIR*** matrix



4. Simulation Study

4.1. Design

To evaluate the above-presented way of looking at the dimensionality structure of a test, a small simulation study was set up. The design considers unidimensional and multidimensional test data in terms of both strict and essential dimensionality.

(1) A first condition consists of unidimensional test data generated under the two-parameter logistic model (Birnbaum, 1968) for $P = 1000$ persons and $I = 18$ items,

$$\Pr(Y_{pi} = 1|\theta_p) = \frac{\exp(\alpha_i(\theta_p - \beta_i))}{1 + \exp(\alpha_i(\theta_p - \beta_i))}.$$

Difficulty parameters β_i and person parameters θ_p were each generated from a standard normal distribution. Item discrimination parameters were chosen to show a relatively large variation ($\alpha_i \sim N(1, 0.1)$, approximate range $[0.7, 1.3]$) to resemble

tests consisting of items with variable strength as indicator of the latent dimension. To manipulate the homogeneity of the scale, 4 types of unidimensiomal test data were generated. The first type resembled a strong homogeneous test where all items are relatively good indicators of the underlying latent variable ($\alpha_i \sim N(1, 0.1)$); the second type resembled a less homogeneous test containing a few less efficient indicators of the latent dimension ($\forall i \in \{5, 6, \ldots, 18\}, \alpha_i \sim N(1, 0.1)$, and $\alpha_i \sim N(0.2, 0.1)$ truncated at $[0.1, 0.3]$ otherwise); the third and forth type resembled a strong homogeneous test in which the item pair $\{i = 1, j = 5\}$ showed a degree of redundancy (i.e., local item dependence). The local item dependence is simulated making use of the copula IRT models proposed by Braeken, Tuerlinckx, and De Boeck (2007). Frank copula which induces local item dependency in a symmetric fashion around the item locations, was chosen. The degree of redundancy was either minor, comparable to Kendall's τ of 0.2, or mediocre, comparable to Kendall's τ of 0.4.

(2) A second condition consists of threedimensional test data generated under a compensatory multidimensional model (cfr. Equation 5) for $P = 1000$ persons and $I = 18$ items. The relation between items and dimensions underlying the test is manipulated, giving rise to two types of tests. In a test of Type A, each item only loaded on a single dimension k ($\alpha_{ik} \sim N(1, 0.1)$, $\alpha_{ik'} = 0$ otherwise) resembling a simple structure factor loading pattern, whereas in a test of type B, each item primarily loaded on a single dimension k, but also loaded to a smaller extent on the other dimensions k' ($\alpha_{ik} \sim N(1, 0.1)$, and $\alpha_{ik'} \sim N(0.2, 0.1)$ truncated at $[0.1, 0.3]$ otherwise). The latter corresponds the most to the essential dimensionality idea (Stout, 1987, 1990), with dominant and noise dimensions for each item, yet is suprisingly not commonly studied in other simulation excercises in this area. The correlations $\rho_{kk'}$ between the 3 dimensions were either all equal to 0, 0.4, or 0.8.

4.2. Results

A summary of the simulation study results for the first condition is given in Table 1. The results are indicative for the performance of the procedure given essentially unidimensional test data. Overall the performance of the clustering criterion $\Psi(\mathbb{S}(K))$ for unidimensionality assesment is quite satisfactory. All unidimensional strong-homogeneous datasets were correctly identified as unidimensional, while 95% of the unidimensional weak-homogeneous datasets were correctly identified as unidimensional. Note that the 5 cases that were misidentified are characterized by relative lower discrimination α_i on the less-efficient indicators compared to the other cases. Even in the presence of minor LID for an item pair, the clustering criterion correctly identified 96% of the datasets as essentialy unidimensional. In the presence of mediocre LID for an item pair, the clustering criterion identified 42% as unidimensional and 58% as multidimensional. Cases that were assessed to be multidimensional are characterized by a larger expression of local item dependence compared to the other cases, reflected in larger maximum SIR values.

Detection of outliers within the $\boldsymbol{SIR}$ matrix was rather succesful. For the datasets generated under strict unidimensionality no false positive results occured. In 46% of the datasets with one minor LID item pair, 1 pair was detected to show excess local item dependence, each time the correct pair $\{1, 5\}$. Given the prior results in which the unidimensionality assessment was rather robust against a minor violation of strict LSI, the relative power of the test in detecting a minor LID pair is higher than expected. In each dataset with one mediocre LID item pair the correct LID pair was detected, and only in one case an additional, false positive, item pair occured ($p = 0.047$).

A summary of the simulation study results for the second condition is given in Table 2. The results are indicative for the performance of the procedure given

		1 dimension: homogeneity ($n = 100$ datasets)			
		strong	weak	minor LID pair	mediocre LID pair
$\Psi(\mathbb{S}(K))$					
chosen	$K = 1$	100	95	96	42
	$K > 1$	0	5	4	58
mean		0	$6.436e^{-6}$	$4.17e^{-5}$	$1.27e^{-3}$
SIR					
mean		$3.49e^{-3}$	$3.05e^{-3}$	$3.53e^{-3}$	$3.73e^{-3}$
max		$1.22e^{-2}$	$1.30e^{-2}$	$2.12e^{-2}$	$5.62e^{-2}$
min		$-5.93e^{-3}$	$-6.24e^{-3}$	$-6.10e^{-3}$	$-5.96e^{-3}$
$SE(SIR)$					
$p \leq 0.05$		0	0	0.46*	1.01**
mean($p \leq 0.05$)				0.023	0.003
p {1,5}				0.074	0.002

(*) In 46 cases the correct LID pair was detected (false negative otherwise).

(**) In all cases the correct LID pair was detected. In one case a second pair was detected: false positive pair $\{3, 6\}$ with $p = 0.047$.

essentially multidimensional test data. In case of type A (resembling simple structure factor loadings), the correct dimensionality is determined in at least 97% of the cases, unless the correlations among the dimensions are high ($\rho = 0.8$), because then the test is assessed to be unidimensional. In case of type B (resembling the essential dimensionality idea), the correct dimensionality is determined in at least 94% of the cases when the dimensions are uncorrelated, while this decreases to 32% with medium-sized dimensional intercorrelations ($\rho = 0.4$). In the latter case, the main choice of the criterion is unidimensionality. When dimensions correlate highly, the criterion always points at unidimensionality. Note that when the correct dimensionality is assessed, the majority of items is also correctly partitioned. The probability of having at least one misclassification is 4% given type A and 25% given type B. When the interdimensional correlation increases from $\rho = 0$ to $\rho = 0.4$, these probabilities increase to 29% and 56%, respectively.

The average number of detected extreme LID pairs and the maximum observed SIR value for an item pair decreases with increasing intercorrelations and when items also load on other dimensions (type B). The minimal observed SIR value for

an item pair follows the opposite pattern. Notice that in the case of high dimensional intercorrelations, no extreme LID pairs are detected.

		3 dimensions: structure (n = 100 datasets)					
		type A			type B		
	$\rho_{kk'}$	0	0.4	0.8	0	0.4	0.8
$\Psi(\mathbb{S}(K))$							
chosen	$K=1$	0	0	97	0	57	100
	$K=2$	0	2	0	1	4	0
	$K=3$	99	97	2	94	32	0
	$K=4$	1	1	1	5	7	0
	$K>4$	0	0	0	0	0	0
mean		$8.95e^{-3}$	$3.43e^{-3}$	3.15^{-6}	$3.79e^{-3}$	$2.30e^{-4}$	0
SIR							
mean		$1.96e^{-3}$	$2.79e^{-3}$	$3.30e^{-3}$	$2.85e^{-3}$	$3.53e^{-3}$	$3.90e^{-3}$
max		$2.26e^{-2}$	$1.84e^{-2}$	$1.31e^{-2}$	$1.88e^{-2}$	$1.61e^{-2}$	$1.40e^{-2}$
min		$-8.92.e^{-3}$	$-7.68e^{-3}$	$-6.4e^{-3}$	$-8.14e^{-3}$	$-6.45e^{-3}$	$-5.45e^{-3}$
$SE(SIR)$							
$p \leq 0.05$		1.36	0.21	0	0.28	0.10	0
mean($p \leq 0.05$)		0.035	0.040		0.038	0.041	

5. Discussion

The dimensionality assessment approach introduced here, is based upon information theory and boundaries on bivariate distributions, and builds on existing nonparametric approaches (cfr., DETECT and Mokken scaling). The starting point from the procedure is a pairwise conditional item association matrix, now based upon a newly-proposed association measure, labeled "signed information rate" (SIR). This SIR measures is not restricted to capture only linear association, and only takes a zero-value in case of independence. To account for the influence of the discrete nature of the item responses on association measures and to allow for better comparison, the SIR measure is normed with respect to the marginal distributions of the item scores, and direction of the association is determined by the relative position of the observed joint distribution $\Pr(U, V)$ in its limiting boundary space.

A basic Z-test making use of the approximate standard error of the association measure provides a way to scan the item association matrix for anomalies and de-

tect extreme locally dependent item pairs that can be expected to have too much influence on the direction that a dimensionality assessment procedure will take for the given item pool. The results showed that the test has good power in detecting item pairs that show local item dependence. Even when the local item dependence is minor and ignorable in the sense of essential independence, the test detected the target item pair in half of the cases. Furthermore, when one of the items within the detected LID pair was left out of the test, all datasets were assessed to be unidimensional. Hence, evaluating item pairs like this shows practical value for test construction.

Because the SIR measure can readily be transformed into a distance measure, a hierarchical clustering procedure (cfr. HCA-CCPROX) can be used to construct an item tree. This item tree offers a range of partitioning solutions which can be evaluated in terms of a dimensionality criterion (cfr. DIMTEST). For this general dimensionality assessment purpose, the criterion $\Psi(\mathbb{S}(K))$ is suggested to select an optimal partitioning, finding a balance between the homogeneity within item clusters and the heterogeneity between item clusters. For the assessment of unidimensionality, the results show promise for the robustness as well as the sensitivity of the procedure. The presence of a minor LID item pair or a small set of atypical items did not prevent the procedure in correctly identifying the dominant unidimensional structure of the test.

For the assessment of specific multidimensionality, the performance of the criterion is not fully optimal from a strict statistical perspective. When the dimensionality structure is clearly expressed, performance is excellent and the correct number of dimensions is identified. However, in case of high inter-dimension correlations the criterion judges the test to be unidimensional (which might well be the case from a substantive viewpoint). In case of essential dimensionality and intermediate cor-

relations, the criterion either correctly identifies the number of dimensions or again opts for unidimensionality. In both these cases, the homogeneity within the $\boldsymbol{SIR}$ matrix has increased to the extent that the between-within criterion $\Psi(\mathbb{S}(K))$ does not differentiate between different cluster solutions. Thus, from a practical point of view, the criterion inherently chooses for more parsimonious solutions when the multidimensionality of the test is not clearly expressed in the data. In fact, one can consider this a welcome characteristic for a dimensionality assessment procedure and in line with the main goal of only locating the dimensions that are dominantly present in the data.

To increase the power of tests and procedures within dimensionality assessment, fundamental developments are needed towards the relationship between the theoretical latent trait and the proxy being used, as well as more specific results in how the pairwise conditional associations relate to the strict LSI condition, and these proxies. Even under the strict conditions of a Rasch model, in which the total sum score is a sufficient statistic for the latent trait, one can merely state that the corresponding pairwise conditional item associations are necessarily non-positive (see e.g., Junker, 1993). In the more general case, in which such a sufficient statistic is unavailable, one can merely state that conditional upon the restscore pairwise item associations are necessarily non-negative (see e.g., Rosenbaum, 1984). Another argument is that asymptotically, when the size of the item set moves towards infinity, the pairwise conditional association measures move towards 0, the value corresponding to independence (see e.g., Stout, 1990). The disparity of the results makes it difficult to derive expected values for pairwise conditional association measures and formulate exact tests in the general context of latent variable models.

Furthermore, these results and remaining issues might raise the statistical question towards the feasibility of distinguishing between specific dimensionality struc-

tures. Given the possibility of many equivalent models in parametric measurement (see e.g., the structural equation literature), our capacity in determing a specific and unique dimensionality structure among a set of item responses might currently be overrated. However, this cautious note should not prevent the further development of procedures that generate and evaluate potential candidate structures, but merely stresses the importance of the interplay between theory and data in the developement of an assessment instrument, allowing us to make at least a well-informed and supported assessment of test dimensionality.

6. Acknowledgement

The author wish to thank Gunter Maris, Hendrik Straat, en Klaas Sijtsma for helpful comments and discussion on an early version of this manuscript.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255–278.

Ark, L. A. Van der. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*, 1–19.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Hodder Arnold.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–497). Reading: Addison-Wesley.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Bolt, D. M. (2001). Conditional covariance based representation of multidimensional test structure. *Applied Psychological Measurement*, *25*, 244–258.

Braeken, J., Tuerlinckx, F., & De Boeck, P. (2007). Copulas for residual dependency. *Psychometrika*, *72*, 393–411.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory.* New York: Wiley.

Cureton, E. E. (1959). Note on ϕ/ϕ_{max}. *Psychometrika*, *24*, 89–91.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université Lyon: Série 3*, *14*, 53–77.

Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement*, *25*, 221–233.

Hoeffding, W. (1940). Masstabinvariante Korrelations Theorie. *Schriften des Matematischen Instituts und des Instituts für angewandte Mathematik der*

Universität Berlin, *5*, 179–223. [Reprinted as Scale-invariant correlation theory in the Collected Works of Wassily Hoeffding, N.I. Fischer, and P.K. Sen (Eds.), New York: Springer.].

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

Humphreys, L. G. (1984). *A theoretical and empirical study of the psychometric assessment of psychological test dimensionality and bias (onr research proposal).* Washington, DC: Office of Naval Research.

Joe, H. (1997). *Multivariate models and dependence concepts.* London: Chapman & Hall.

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

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65–81.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data.* New York: John Wiley & Sons.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized tests.* Urbana-Champaign: Unpublished doctoral dissertation, Univerity of Illinois at Urbana-Champaign, Departement of Statistics.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, *45*, 507–549.

MacNaughton-Smith, P., Williams, W. T., Dale, M. B., & Mockett, L. G. (1964). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature*, *202*, 1034–1035.

McDonald, R. P., & Mok, M. M. (1995). Goodness of fit in item response models.

Multivariate Behavioral Research, *30*, 23–40.

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general rasch model with multidimensional item response data.* Iowa City, IA: American College Testing.

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

Molenaar, I. W., & Sijtsma, K. (2000). *User manual msp5 for windows.* Groningen, The Netherlands: ProGamma.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics.* New York: McGraw-Hill.

Nandakumar, R., & Stout, W. F. (1993). Latent and manifest monotonicity in item response models. *Journal of Educational Statistics*, *18*, 41–68.

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

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193–203.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425–435.

Roulston, M. S. (1999). Estimating the errors on measured entropy and mutual information. *Physica D*, *125*, 285–294.

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.

Roznowski, M., Tucker, L. R., & Humphreys, I. G. (1991). Three approaches to determining the dimensionality of binary items. *Applied Psychological Mea-*

surement, *15*, 109–127.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *52*, 589–617.

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

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, *48*, 191–199.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*, 159–203.

Van Abswoude, A. A. H., Ark, L. A. Van der, & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric irt models. *Applied Psychological Measurement*, *28*, 3–24.

Warrens, M. (2008). On association for 2x2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *7*, 777–789.

Zhang, J., & Stout, W. F. (1999a). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.