universität
wien

DISSERTATION

Titel der Dissertation:

“Methods of Dimensionality Assessment in Psychological Measurement and their Application to Cognitive Assessment”

verfasst von

Mag. Rudolf Debelak

angestrebter akademischer Grad:

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2013

Studienkennzahl lt. Studienblatt:	A 091 298
Dissertationsgebiet lt. Studienblatt:	Psychologie
betreut von:	Ao. Univ.-Prof. Dr. Georg Gittler

Table of Contents

Introduction

The last decades saw the rise of multiple mathematical models which described a respondent's behavior when working on a psychological test. Since these models described the interaction of the respondent population and an item population, models of this kind have been named item response theory (IRT) models (Embretson & Reise, 2000). An overview given by van der Linden and Hambleton (1997a) listed several dozen IRT models, but was still far from exhaustive. In Germany, a similar, but less extensive overview was provided by Rost (2004). Since the publication of these books, even more IRT models have been described and reviewed for their practical applicability.

In their introductory book to IRT, Embretson and Reise (2000) note that IRT models are essential for model-based measurement and describe how a person's level of a psychological trait can be optimally described by the responses given to the items of a psychological test. In general, IRT models make strong assumptions on the underlying relationship between respondents and items. Psychometricians have recommended to test these assumptions, often by the means of statistical and graphical model tests, and to select the model showing the best fit to the test data (cf. Rost, 2004; van der Linden & Hambleton, 1997b).

IRT models have also been used to assess whether a given test meets certain characteristics which have been considered as desirable. A well-known example is the one-parameter logistic (1PL) model of Rasch (1960), which can be used to describe the relationship between respondents and binary items. This model assumes that the probability of a positive response of respondent *v* to item *i* can be described by the formula:

$$P(+|\beta_i,\theta_v) = \frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)} \tag{1}$$

In formula (1), θ_v denotes a respondent-specific parameter, which is interpreted as the respondent's ability in many applications of the Rasch model, since the probability of a positive response is higher for large values of θ_v. β_i is an item-specific parameter, which is often interpreted as the item's difficulty. The Rasch model makes four strong assumptions on the relationship between the item and respondent population (cf. Fischer, 1974): First, the items are regarded to be unidimensional, which means that they measure a single trait. Second, it is assumed that the probability of a correct response is solely determined by the trait level of the respondent and can get arbitrarily close to 0 or 1. Third, the sum of correct responses is a sufficient statistic for the ability of the respondent. The fourth assumption of the Rasch model is that the items and respondents can be considered as locally independent, which implies that the response of a respondent to an item does not depend on other person-item-interactions. Fischer (1974) proved that the Rasch model is the only IRT model which meets these four assumptions.

Numerous methods have been suggested to test the assumptions of the Rasch model, including graphical (cf. Rost, 2004), parametric (e.g. Andersen, 1973; Glas, 1984) and non-parametric (Koller & Hatzinger, 2012; Ponocny, 2001) approaches. Suárez-Falcón and Glas (2003) reported the results of a simulation study which compared the power and sensitivity of some tests for the Rasch model.

Many IRT models including the Rasch model assume the unidimensionality of the test items, and many authors agree that the assessment of the dimensionality of test items is central for test development and test evaluation (e.g. Hattie, 1985). As a consequence, multiple approaches for dimensionality assessment have been described and suggested. An important distinction can be made between exploratory approaches, which aim to determine the number of dimensions underlying an item set, and confirmatory approaches, which assume a specific number of dimensions (cf. Reckase, 2009).

Determining the dimensionality of test items influences practical psychological assessments as well as the development of psychological theories. Hattie (1985) provided a critical review of many early exploratory and confirmatory approaches and generally found most of them unsatisfying. Overviews of specific approaches have been provided by several authors, including Baker and Kim (2004), Embretson and Reise (2000), Reckase (2009) and van der Linden and Hambleton (1997a). The three studies contained in this doctoral thesis extend the findings reported by these authors by investigating new methods of dimensionality assessment which were developed for the evaluation and development of psychological tests.

Outline Study I

Although IRT and factor analytical methods are among the most widely used methods in dimensionality assessment, there are other methods which should be described briefly in this section. Different approaches for dimensionality assessment have suggested the use of cluster analytical methods (e.g. Bartolucci, 2007; Bartolucci, Montanari, & Pandolfi, 2012; Roussos, Stout, & Marden, 1998). Reckase (2009) mentions cluster analysis of items as a measure to confirm the dimensionality of an item set. In this approach, a similarity measure is defined which describes to which extent two items measure the same dimension. A possible similarity measure described in the literature is the covariance between two items conditional on the respondent's ability, i.e. the number correct score (cf. Roussos et al., 1998). The criterions to determine the dimensionality of a test using cluster analysis are somewhat subjective, and different clustering methods may lead to differing results. Moreover, there seems to be no definite conclusion in the literature which method leads to optimal results. Reckase (2009) thus recommended cluster analysis as an approach to confirm the dimensionality of a test, and remarked that cluster analysis may overestimate the number of dimensions underlying a test.

Van Abswounde, van der Ark and Sijtsma (2004) compared several non-parametric methods for dimensionality assessment which were based on cluster analysis. These methods included a cluster analytical approach based on Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002), cluster analytical approaches based on the concept of essential unidimensionality (Stout, 1987, 1990), and cluster analytical approaches based on the conditional covariance between items (Roussos et al., 1998). They found that some of the non-parametric approaches accurately detected the underlying dimensionality of an item set in the presence of multidimensionality. However, these approaches are based on different theoretical assumptions when compared to common IRT models like the Rasch model, so the clusters built from these procedures do not necessarily show a good fit to the Rasch model or comparable models.

Study I describes a possible application of hierarchical cluster analysis to Rasch measurement. In this approach, fit statistics for the Rasch model are used to construct item clusters showing a good fit to the Rasch model. The proposed method uses a model fit statistic for the Rasch model as a similarity measure in a hierarchical cluster analysis.

This paper further compares the results obtained from this approach to the results obtained from a principal component analysis (PCA) of tetrachoric correlations without applying a smoothing algorithm by the means of a simulation study. In this simulation study, datasets were simulated which consisted of the responses of a person sample answering to two scales each fitting the Rasch model. The simulated datasets differed with respect to the distribution of the item and person parameters, the sizes of the person and item sample, and the correlations between the person parameters of the two scales.

In this simulation study, it was found that the proposed algorithm for clustering items more often led to a correct reconstruction of the two scales when the item set was small, when the person sample was large, when the standard deviation of the person parameters was large when compared to the standard deviation of the item parameters, and when the correlation between the person parameters was small.

The application of PCA often led to more correct results when compared to the cluster analytical algorithm. However, it was also very often observed that the application of this method was not possible due to the presence of indefinite correlation matrices. This study was further examined in study II. Both approaches were further demonstrated by applying them to a real dataset of a person sample which worked on the intelligence test battery IBF (Blum, Didi, Fay, Maichle, Trost, Wahlen, & Gittler, 2005).

Outline Study II

An important assumption of the Rasch model described in study I as well as many other IRT models is the unidimensionality of the item set. A well-known approach for testing unidimensionality includes methods based on exploratory factor analysis (EFA) or PCA. This approach aims to determine the number of factors or components underlying a matrix of correlations or covariances measured in the analyzed item set. Exploratory and confirmatory factor analysis is still a field of active research, and papers describing new approaches are constantly published. The approaches used in practical research differ in several central aspects, including the method used for factor or component extraction, the method used to determine the number of critical factors or components, and the nature of the correlations used for analysis.

Many authors (e.g. Lord & Novick, 1968; Lord, 1980) have already emphasized that the use of Pearson correlations in EFA or PCA assumes that the underlying variables stem from an interval scale, and should therefore not be used for ordinal variables. A common example for a violation of this assumption is the analysis of dichotomous items, which allow only binary responses (e.g., correct and incorrect). It is well documented in the literature that the application of EFA or PCA based on Pearson correlations leads to misleading results, and alternatives have been proposed for this case. An early suggestion was the analysis of tetrachoric correlations instead of Pearson correlations (cf. e.g. Lord, 1980). A major advantage of tetrachoric correlations over Pearson correlations is their independence of the relative frequency of the response categories observed in the data, or in other words, their independence of the item difficulties. As a consequence, the application of PCA or EFA based on tetrachoric correlations has been found to lead to more reliable results than the application of a comparable procedure based on Pearson correlation (e.g. Tran & Formann, 2009; Weng & Cheng, 2005)

An open question in the application of tetrachoric correlations in factor-analytical methods is the problem that very often the calculation of tetrachoric correlation leads to indefinite correlation matrices, i.e. correlation matrices with at least one negative eigen-value. Since it is not possible to carry out factor-analytical methods with indefinite correlation matrices, a solution to this problem is necessary to make factor analytical methods applicable in case that a non-definite correlation matrix is observed in a data set. Possible solutions to this problem have been suggested by multiple authors. Tran and Formann (2009) mentioned a paper of Knol and Berger (1991) which described smoothing algorithms in order to convert a non-definite matrix in a positive definite one. Another solution to this problem was suggested by Bentler and Yuan (2011).

Once the factors or components underlying a data set have been calculated, an important question concerns the determination of the number of critical dimensions. This topic was thoroughly investigated by multiple authors (e.g. Ford, MacCallum, & Tait, 1986; Green, 1983; Guttman, 1954; Horn, 1965; Humphreys & Montanelli, 1975; Kaiser, 1960; Zwick & Velicer, 1986), and a number of different approaches was proposed and evaluated. A promising approach which has been recommended by many authors (e.g. Ledesma & Valero-Mora, 2007; Reckase, 2009) is parallel analysis, which seems to have been proposed first by Horn (1965). Parallel analysis consists of two major steps: First, the eigenvalues of the factors or components underlying the analyzed dataset are calculated. In a second step, a number of datasets which are comparable to the original dataset with respect to the sample size, the number of items and the distribution of the responses, but consist of uncorrelated items, are simulated. Based on these simulated datasets, the distribution of eigenvalues under the condition that there is no relationship between the items is simulated. Once this distribution is determined, the number of dimensions underlying the dataset is determined by comparing the eigenvalues observed in the dataset and the distribution of eigenvalues under the condition that there is no relationship between the items.

Although parallel analysis is no exact mathematical procedure, it has been found to determine the correct number of dimensions under many conditions. Many improvements on the original approach described by Horn (1965) have been suggested (e.g. Glorfeld, 1995; Green, Levy, Thompson, Lu, & Lo, 2012), which lead to improved results under specific conditions when compared to Horn’s method.

Besides parallel analysis, several other procedures have been suggested to determine the number of dimensions in factor analysis, which will not be reviewed here in detail. Alternatives to parallel analysis include the MAP criterion by Velicer (1976), the scree test criterion (Cattell, 1966), the eigenvalue-greater-than-one criterion (Kaiser, 1960), and the Bartlett test (Bartlett, 1950, 1951). Comparisons of these methods showed that under many conditions, parallel analysis leads to the most accurate results (Zwick & Velicer, 1986).

Study II examined the problem of indefinite correlation matrices in the application of a PCA of tetrachoric correlations by the means of a simulation study. Parallel analysis was used to determine the dimensionality of a dataset.

In this simulation study, responses of a person sample working on an item set consisting of two scales which fitted the two-parameter logistic test model of Birnbaum (1968) were simulated. The simulated datasets varied with respect to the size of the item set, the person sample, the distribution of the item and person parameters, the correlation between the person parameters in the simulated scales, and the smoothing algorithm which was applied when a non-definite correlation matrix was obtained. Under each condition, 1000 datasets were simulated. In each dataset, dimensionality was assessed by applying a parallel analysis and PCA of tetrachoric correlation matrices.

It was found that indefinite tetrachoric correlation matrix are most often observed in datasets with large item sets, small person samples, and large discrimination parameters of the scales. Large correlations between the person parameters more often led to incorrect identifications of the dimensionality. In this study, also minor differences between the results obtained by applying the different smoothing algorithms were observed.

In summary, study II proposed and evaluated three smoothing algorithm which can be applied if indefinite correlation matrices are observed in the application of parallel analysis and PCA to tetrachoric correlation matrices. It was found that this procedure of dimensionality assessment often leads to correct results under the conditions simulated in this study. However, it should be noted that in this simulation study, no conditions were simulated which violated the assumptions underlying the tetrachoric correlation coefficient (e.g. the presence of guessing).

Outline Study III

While the first two studies investigated methods for testing the unidimensionality of an item set, the third study used a modern modeling approach, which also assumes the unidimensionality of an item set, to test a specific hypothesis on gender differences in mental rotation tasks. Study III thus exemplifies the application of a modern confirmatory IRT approach to test specific hypotheses on the cognitive processes measured by a psychological test.

There is a wide consensus in the literature that there are gender differences in some spatial abilities, although these differences depend on the specific test. In their meta-analysis of gender differences in spatial abilities, Linn and Petersen (1985) examined three types of spatial ability tests, which they named spatial perception, mental rotation and spatial visualization. They found gender differences in all three task types, with mental rotation showing the largest gender differences with effect sizes of .73 over all examined age groups. For spatial perception tests, an effect size of .44 was found, while for spatial visualization tasks, an effect size of .13 was reported.

Halpern (2011) reported multiple studies which provided evidence for gender differences in the three types of spatial ability tests described by Linn and Petersen (1985). She also listed several studies which reported gender differences in spatiotemporal ability tests (e.g. Contreras, Rubio, Pena, Colom, & Santacreu, 2007; Law, Pellegrino, & Hunt, 1993) and visual imagery tests (e.g. Dror & Kosslyn, 1994).

Voyer (2011) assumed that gender differences in spatial ability and more specifically in mental rotation tasks are related to the presence of time limits. His hypothesis was based on previous studies of Goldstein, Haldane and Mitchell (1990), who found in two experiments that women work slowly and cautiously on mental rotation tasks when compared with men. It should be noted that these results could not always be reproduced (Masters, 1998). Goldstein et al. (1990) suggested that the often reported gender differences in favor of males are caused by gender differences in the processing speed of male and female respondents when working on mental rotation tasks. In a meta-analysis Voyer (2011) examined the presence of gender differences in mental rotation tasks under various time limits. He found gender differences in favor of males over under all time limit conditions. Voyer also reported that these gender differences grew larger in the presence of short time limits, however, they were not removed when no time limits were present, which contradicted the prediction of Goldstein et al. (1990).

Study III investigates the presence of gender differences in mental rotation processing speed by applying a new item response model, which examines the unidimensionality of the responses and also the item response times.

This approach was first described by van der Linden (2007) and uses an explicit model for response behavior and working times; the work of van der Linden and colleagues (van der Linden, 2006, 2007, 2009; Klein Entink, Fox & van der Linden, 2009) mainly elaborated the psychometric theory for describing the response behavior of binary items. In this approach, three levels of psychometric models are described. On the first level, two psychometric models are defined, of which the first describes the response behavior with regard to the given answer. Two typical choices are the one- and two-parameter normal ogive models, which are closely related to the Rasch model used in study I and the two-parameter logistic model of Birnbaum (1968) described in study II (Embretson & Reise, 2000).

The two-parameter normal ogive (2PNO) model defines a person parameter θ_i which marks ability of person i to answer items correctly. The model further defines two item parameters for each item k which define the respective item's difficulty b_k and discrimination a_k. This model contains the one-parameter normal ogive model as a special case, in which the discrimination parameter is regarded as equal for all items, and which is closely related to the Rasch model (Rasch, 1960; cf. Embretson & Reise, 2000). In the 2PNO model, the probability that person i answers item k correctly is given by:

$$P(+|\theta_i, a_k, b_k) = \Phi(a_k\theta_i + b_k) \tag{2}$$

In this formula, $\Phi()$ denotes the cumulative function of the standard normal distribution.

The second model on the first level of this modeling approach describes the response times during the test. Van der Linden (2007, 2009) suggested using the one- or two-parameter log-normal (2PLNO) model for describing the response times. As in the 2PNO model, two item parameters are defined for each item that describe the respective item's time intensity and time discrimination. For each person i, a speed parameter is defined. This model contains the one-parameter log-normal model as a special case, in which the time discrimination parameter is set to a fixed value.

In the 2PLNO model, the log response time T_{ik} of a person i working on item k is given by:

$$T_{ik} = -\phi_k \zeta_i + \lambda_k + \varepsilon_{ik} \quad (3)$$

In this formula, ζ_{i} denotes the respondent's speed, ϕ_{k} denotes an item's time intensity and ϕ_{k} an item's time discrimination. $\varepsilon_{\mathrm{ik}}$ is a residual-term which is assumed to be normal distributed with an item-specific variance.

On the second level of response time modeling, the interaction between the measured item and person parameters is investigated. On this level, two variance-covariance matrices are defined, of which the first matrix describes the variances and covariances of the item parameters of the response and response time models. The second matrix describes the corresponding variances and covariances for the person parameters. The approach of van der Linden thus does not assume the person and item person to be independent, but it assumes a linear relationship between each pair of item parameters as well as between the two person parameters used for describing the responses and response times.

On the third and highest level, the relationship between the observed item and person parameters and covariates for the item and person parameters is described. Similar to models like the linear logistic test model (LLTM) of Fischer (1995; cf. De Boeck & Wilson, 2004), the item and person covariates are assumed to explain the observed item and person parameters. Goldhammer and Klein Entink (2011) presented a study which has successfully applied this approach to model item difficulty and time intensity in a reasoning test.

Study III applied this method to the datasets of two person samples working on two different mental rotation tasks, i.e. a cube comparison task and an endless loop task. It is shown that in both datasets, gender differences in the responses but not in the response times were observed.

This study thus provided further evidence that the well-known gender differences in mental rotation tasks can not be attributed to gender differences in mental rotation speed. Furthermore, a negative correlation between speed and ability was observed, which is in line with results reported by Goldhammer and Klein Entink (2011).

General Discussion

This doctoral thesis contains three papers which examined different aspects of dimensionality assessment in test development and test evaluation. The first and second paper described several methods to assess the dimensionality of an item set, while the third paper described recently developed methods to model responses and response times of psychological tests and applied them to investigate gender differences in mental rotation processing speed.

Based on the results presented in these papers, a number of conclusions can be drawn. A first research question concerns the comparison of the two methods of dimensionality assessment presented in the first two papers. The first paper presented an approach which assessed the dimensionality of an item set by clustering items, and compared it to the results of a PCA of tetrachoric correlations in a simulation study. One of the major results of this first paper was that the PCA could often not be applied, since the corresponding correlation matrix was not positive-definite. The second paper examined three possible solutions to this problem and compared them based on a simulation study.

When compared to the PCA of smoothed tetrachoric correlation matrices, the cluster analytical approach presented in the first study seemed less sensitive to violations of unidimensionality. Since this approach is based on a step-wise decision to measure the dimensionality of an item set, it also capitalizes on chance to some degree, which may lead to more incorrect decisions in larger item sets. This conclusion is in line with the observed results for larger item sets. However, the cluster analytical approach has several major advantages when compared to a PCA of tetrachoric correlations. First, it does not assume that a specific model underlies the item set, but searches for clusters of items which fit a specific model (i.e., the Rasch model) well. Under this perspective, the cluster analytical approach seems to be more general. The PCA of smoothed tetrachoric correlations, on the other hand, is based on explicit model assumptions which concern the whole item set, and may not be appropriate if these assumptions are violated. However, if these assumptions are met, a PCA of tetrachoric correlations seems more appropriate to assess the dimensionality of an item set based on the results of study I and II.

The approaches discussed in the first two studies can also be regarded as approaches for testing specific assumptions of the Rasch model (Rasch, 1960) and the two-parameter logistic test model of Birnbaum (1968). The cluster analytical approach presented in the first study aimed at constructing item cluster which showed a good fit to the Rasch model. If this approach is applied to an item set fitting the Rasch model, it is to be expected that the application of cluster analytical method presented in the first paper to this item set leads to an item cluster containing all items of the initial item set. The PCA of smoothed tetrachoric correlation matrices, which is discussed in the second paper, can be used to test the unidimensionality of an item set, which is a common assumption of many IRT model, including the Rasch model and the two-parameter logistic test model. However, both approaches are restricted to the analysis of binary items, and future work could examine possible extensions of both approaches to other item types.

Since the Rasch and Birnbaum models are closely related to the one- and two-parameter normal ogive models (cf. Embretson & Reise, 2000), the approaches presented in the first two papers are related to the modeling approach of van der Linden (2006), which is used in the third paper. In contrast to study I and study II, study III does not report the results of a simulation study, but demonstrates the application of a recent psychometric modeling approach in the analysis of the cognitive processes measured by specific ability tasks.

On its first level, this modeling approach consists of two models for the responses and response times. The response model is very similar to the two-parameter logistic model of Birnbaum, so the methods presented in the first two papers may be used to test assumptions of van der Linden's response model. The response time model is a characteristic feature of this modeling approach, which marks an important difference to traditional IRT models like the Rasch model. The application of this modeling approach can be used for testing psychological theories, as was exemplified by the third paper of this thesis.

As study III shows, the modeling of responses and response times in two mental rotation tasks led to several conclusions on the cognitive processes underlying mental rotation. First, it was observed that gender differences in untimed mental rotation tasks can not be explained by gender differences in mental rotation processing speed, as it was hypothesized by Goldstein et al. (1990). Second, it was observed that processing speed and ability are negatively correlated in both mental rotation tasks. Although similar results have been reported for other untimed ability tests (e.g. Goldhammer & Klein Entink, 2011), this result led to interesting conclusions on the role of speed-accuracy tradeoff in mental rotation tasks, which were discussed in study III.

The modeling approach of van der Linden (2006) is still a field of active research, and future research may lead to additional tests of model fit for this modeling approach. It appears that no statistical test has been described yet in the literature which allows testing the equivalence of this modeling approach's item parameters over several groups (Klein Entink, personal communication). This drawback leads to a possible limitation for the third study in this paper, since a difference of the item parameter estimations between males and females could affect the interpretation of observed differences in the speed and ability parameters between males and females. However, a separate estimation of the item parameters for males and females led to comparable results for almost all items of both tests examined in study III, so the differences of the item parameters between males and females were considered as negligible; this line of reasoning is comparable to that of the well-known graphical model test for the Rasch model (cf. Rost, 2004). This finding further confirms the theoretical implications of study III on the gender differences in mental rotation processing speed and mental rotation processing ability. Nevertheless, future work may improve on this approach.

In the first and third study, methods of dimensionality assessment have been applied to real data sets of intelligence tests. In the first paper, these data consisted of a person sample who worked on an intelligence test battery, which contained a version of a cube comparison task which was presented with a time limit. In the third paper, both a cube comparison task and an endless loop task were presented without time limit to two separate respondent samples (although in the used version of the cube comparison task, testing was quitted as soon as a time limit of 30 minutes had been reached and the last item presented to the respondent had been answered). The results reported in both papers demonstrate that cube comparison tasks of this type show a good fit to the Rasch model, which is in line with results reported in previous studies (e.g. Gittler, 1984, 1986, 1990, 1992; Tanzer, Gittler & Ellis, 1995).

Overall, the methods presented in this doctoral thesis may help to assess the dimensionality of responses and, if the modeling approach of van der Linden is applied, response times in psychological and educational tests. Their application may thus be of use in the evaluation of psychological tests and for the development of psychological theories.

Conclusions

The studies contained in this doctoral thesis evaluated several methods for dimensionality assessment in psychometric test evaluation. Study I described and evaluated an algorithm for item clustering based on model fit statistics for the Rasch model. This algorithm makes no explicit assumptions on the analyzed item set and may serve as a screening procedure in dimensionality assessment when no assumptions regarding the dimensionality of the analyzed item set can be made. In a simulation study, the reported algorithm was found to lead to lead more often to correct results if the sample of respondents was large and the analyzed item set was small. The results were also found to be dependent on properties of the item subsets which fitted the Rasch model well, like the correlations between the person parameters. However, it was also reported in the first study that under certain conditions, the application of PA and PCA of tetrachoric correlations led to more correct results than the cluster analytical algorithm. Study II thus further examined the problem of indefinite tetrachoric correlation matrices, which was observed in study I. It described and evaluated three different smoothing algorithms which could make PA and PCA applicable if indefinite correlation matrices were observed. In study II, the proposed PCA of smoothed tetrachoric correlation matrices was found to be a valid assessment procedure for the dimensionality of an item set under many conditions.

Study III exemplified a new approach for assessing unidimensionality of responses and response times. By applying an approach developed by van der Linden and colleagues (van der Linden, 2006, 2007, 2009; Klein Entink, Fox & van der Linden, 2009) the presence of gender differences in mental rotation processing speed in two mental rotation tasks was investigated. This modeling approach is closely related to the psychometric models used in study I and II. It was found that the observed gender differences in mental rotation ability can not be attributed to gender differences in mental rotation processing speed.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

Baker, F. B., & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques. New York: Dekker.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, 3, 77-85.

Bartlett, M. S. (1951). A further note on tests of significance in factor Analysis. *British Journal of Psychology*, 4, 1-2.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, *72*, 141-157.

Bartolucci, F., Montanari, G.E., & Pandolfi, S. (2012). Dimensionality of the latent structure and item selection via latent class multidimensional IRT models. *Psychometrika*, *77*, 782-802.

Bentler, P. & Yuan, K.-H. (2011). Positive definiteness via off-diagonal scaling of a symmetric indefinite matrix. *Psychometrika*, *76*, 119-123.

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

Blum, F., Didi, H.-J. , Fay, E., Maichle, U., Trost, G., Wahlen, H.-J., & Gittler, G. (2005). *Basic Intelligence Functions (IBF*). Mödling: Schuhfried.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Contreras, M. J., Rubio, V. J., Peña, D., Colom, R., & Santacreu, J. (2007). Sex Differences in Dynamic Spatial Ability: The Unsolved Question of Performance Factors. *Memory and Cognition*, *35*, 297-303.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Dror, I. E. & Kosslyn, S. M. (1994). Mental imagery and aging. *Psychology and Aging*, *9*, 90-102.

Embretson, S. E. & Reise, S. (2000). *Item Response Theory for Psychologists*. Erlbaum: Erlbaum Publishers.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to mental test theory]. Bern: Huber.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer, & I. W. Molenaar (Eds.), Rasch models. Foundations, Recent Developments, and Applications (pp. 157−180). New York: Springer.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*, 291-314.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Gittler, G. (1984). Entwicklung und Erprobung eines neuen Testinstruments zur Messung des räumlichen Vorstellungsvermögens [Development and evaluation of a new assessment for the measurement of spatial abilities]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *5*, 141-165.

Gittler, G. (1986). Inhaltliche Aspekte bei der Itemselektion nach dem Modell von Rasch [Content facets in item selection in Rasch measurement]. *Zeitschrift für experimentelle und angewandte Psychologie*, *33*, 386-412.

Gittler, G. (1990). *Dreidimensionaler Würfeltest—Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. Theoretische Grundlagen und Manual* [Three dimensional cubes test—A Rasch-calibrated test for the measurement of spatial ability. Theoretical background and Manual]. Weinheim: Beltz.

Gittler, G. (1992). *Testpsychologische Aspekte der Raumvorstellungsforschung - Kritik, Lösungsansätze und empirische Befunde* [Aspects of psychological testing in the examination of spatial abilites]. Wien: Habilitationsschrift der Universität Wien.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525-546.

Glorfeld, L.W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*, 377-393.

Goldhammer, F. & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*, 108-119.

Goldstein, D., Haldane, D., & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition*, *18*, 546–550.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement*, *7*, 139-147.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, *72*, 357-374.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, *19*, 149-161.

Halpern, D. F. (2011). *Sex differences in cognitive abilities*. Mahwah: Erlbaum.

Hattie, J. (1985). Methodology review:Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*, 193-206.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141-151.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21 - 48.

Knol, D., & Berger, M. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457-477.

Koller, I. & Hatzinger, R. (2012, April). Exakte Tests für das Rasch Modell unter besonderer Berücksichtigung von lokaler stochastischer Unabhängigkeit [Exact tests for the Rasch model with particular regard to local stochastic independence]. In: I. Koller, M.J. Maier, K. Gruber, R.W. Alexandrowicz, I.W. Nader. *Item Response Modelle: Weiterentwicklung, Überprüfung und Anwendung.* Symposium gehalten bei der 10. Tagung der Österreichischen Gesellschaft für Psychologie, Graz.

Law, D. J., Pellegrino, J. W., & Hunt, E. B. (1993). Comparing the tortoise and the hare: Gender differences and experience in dynamic spatial reasoning tasks. *Psychological Science*, *4*, 35-40.

Ledesma, E. L., & Valero-Mora, P. (2007). Determing the number of factors to retain in EFA: an easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, *12*, 1-11.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterisation of gender differences in spatial abilities: A meta-analysis. *Child Development*, *56*, 1479–1498.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlsbaum Associates.

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

Masters, M. (1998). The gender difference on the Mental Rotations Test is not due to performance factors. *Memory & Cognition*, *26*, 444–448.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *66*, 437-460.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, NY: Springer.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber.

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.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293-326.

Suárez-Falcón, J.C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127-143.

Tanzer, N. K., Gittler, G. & Ellis, B.B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment, 11*, 170-183.

Tran, U. S. & Formann, A. K. (2009). Performance of Parallel Analysis in Retrieving Unidimensionality in the Presence of Binary Data. *Educational and Psychological Measurement*, *60*, 50-61.

van Abswoude, A. A. H., van der Ark, L. A., & Sijtsma, T. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*, 1, 3-24.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181-204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272.

van der Linden, W. J. & Hambleton, R. K. (1997a). Item Response Theory: Brief History, Common Models, and Extensions. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer.

van der Linden, W. J. & Hambleton, R. K. (Eds.) (1997b). *Handbook of modern item response theory*. New York: Springer.

Velicer, W. E (1976). The relation between factor score estimates, image scores and principal component scores. *Educational and Psychological Measurement*, *36*, 149-159.

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin and Review*, *18*, 267-277.

Weng, L.-J., & Cheng, C.-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*, 697-716.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442.

Study I

Debelak, R., & Arendasy, M. (2012). An algorithm for clustering items and testing unidimensionality in Rasch measurement. *Educational and Psychological Measurement*, *72*, 375-387.

Role of the authors:

Rudolf Debelak wrote the manuscript, wrote all code involved in this study, and set up the final design of the simulation study. He was also responsible for the calculations used for the practical application.

Martin Arendasy provided helpful advice on the design of the simulation study and on the topic of Rasch measurement.

Educational and Psychological
Measurement
72(3) 375–387
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164411426005
http://epm.sagepub.com

An Algorithm for Testing Unidimensionality and Clustering Items in Rasch Measurement

Rudolf Debelak[1] and Martin Arendasy[2]

Abstract

A new approach to identify item clusters fitting the Rasch model is described and evaluated using simulated and real data. The proposed method is based on hierarchical cluster analysis and constructs clusters of items that show a good fit to the Rasch model. It thus gives an estimate of the number of independent scales satisfying the postulates of sufficiency of total number of correctly answered items for a person's proficiency, unidimensionality, and local independence that can be constructed from an item set. The method is also compared with the application of a principal components analysis based on tetrachoric correlations. In general, the proposed method was shown to provide practically usable results especially for large person samples.

Keywords

binary variables, parallel analysis, Rasch model

Introduction

In this article, a statistical method is described that allows the identification of item scales showing a good fit to the unidimensional Rasch model (Rasch, 1960) in a multidimensional item set. It thus gives an estimate of the number of independent scales, satisfying the postulates of sufficiency of total number of correctly answered items

[1]Schuhfried GmbH, Mödling, Austria
[2]University of Graz, Graz, Austria

Corresponding author:
Rudolf Debelak, Schuhfried GmbH, Hyrtlstraße 45, Mödling, 2340, Austria
Email: debelak@schuhfried.at

for a person's proficiency, unidimensionality, and local independence that can be constructed from an item set (Fischer, 1974).

The proposed method is intended to be used prior to the application of other model tests that have power against specific model violations. Appropriate test statistics have been described elsewhere (e.g., Andersen, 1973; Glas, 1988; Martin-Löf, 1973; Ponocny, 2002; van den Wollenberg, 1982; Wright & Panchapakesan, 1969). A discussion of additional test procedures in the context of Rasch measurement was provided by Linacre (1992), E. V. Smith (2002), and other writers.

Because the application of the Rasch model requires the unidimensionality of the data, several statistical tests have been proposed or used in the research literature to test this assumption prior to Rasch analysis. These procedures include principal components analysis (PCA; R. M. Smith, 1996) and other approaches based on factor analysis (e.g., Wirth & Edwards, 2007). Lord (1980), among others, already discussed some of the problems associated with this approach when analyzing binary data. Also, in the application of exploratory factor analysis and PCA, determining the correct number of factors or components plays a crucial role. Recent writers (e.g., Tran & Formann, 2009; Weng & Cheng, 2005) evaluated the performance of parallel analysis, an approach possibly initially suggested by Horn (1965), in solving this problem. Although Weng and Cheng (2005) found that parallel analysis performed well in determining the correct number of factors, Tran and Formann (2009) concluded that the usefulness of classical linear factor analysis and PCA is diminished in the presence of binary data.

In contrast to PCA and methods based on factor analysis, the procedure described in the article at hand does not assume that a specific model underlies the analyzed item set. Instead, it is based on the idea of a partial hierarchical cluster analysis, which uses a test statistic for the Rasch model as a measure of similarity.

There have already been numerous approaches for applying cluster analysis in item response theory (for an overview, see Reckase, 2009, or van Abswoude, van der Ark, & Sijtsma, 2004). Many of them try to assign each item to one of several clusters, and often they provide no clear criterion to determine the number of clusters underlying an item set (for an illustration, see again Reckase, 2009). The application of cluster analysis described in this article addresses this issue by providing a statistic of model fit which is used as a strict criterion for selecting items to yield a unidimensional cluster.

The described method could also serve as an alternative to other approaches for the assessment of unidimensionality in the context of Rasch measurement. If an item set fits the Rasch model, it is to be expected that the suggested procedure assigns all items to a single cluster. For comprehensive reviews of methods for assessing the dimensionality of an item set, see Hattie (1985).

The procedure described in this article will be evaluated using simulation studies. As will be shown, it provides results applicable under many circumstances but generally requires large person samples. To compare the new approach with other methods, its results will be compared with those of a PCA and parallel analysis based on tetrachoric correlations.

The remainder of this article is organized as follows: In the next section, the procedure will be described, which includes a discussion of the R_{1c} statistic of Glas (1988). The subsequent section contains a description of the method of a simulation study that was used to evaluate the procedure. This is followed by the ''Results of the Simulation Study'' section. The penultimate section contains results from an empirical study in which the results of the procedure are compared with other tests of fit to the Rasch model. The article concludes with a discussion of the procedure and directions for future research.

Description of the Procedure

The Basic Structure of the Procedure

The method used to assign items to scales that fit the Rasch model will first be formally described. Its principal idea is similar to that of a *partial* hierarchical cluster analysis. Given an item set O, let O_n be the set of all sets of items of O, consisting of n items. By calculating a statistic of fit to the Rasch model, a function f is defined that assigns the p value of a statistic of model fit to a subset of O. The procedure starts with the analysis of all item subsets that are elements of O_3. The initial subset A_3 in the scale construction is the subset for which f reaches its maximum.

After defining A_3 the procedure begins to expand this subset. Let A_n be a subset of n items already constructed by the procedure. The procedure constructs a new item subset A_{n+1}, containing $(n + 1)$ items, by analyzing all elements of $O_{n + 1}$ that contain all elements of A_n. A_{n+1} is defined as the item subset for which f is maximized. This procedure might terminate as soon as the maximum of f, calculated for each element of O_k with a fixed k value, is below a predefined upper threshold, or as soon as O_k cannot be expanded any further. In the study at hand, the R_{1c} statistic of Glas (1988) was used as the statistic of model fit. This test statistic will be reviewed in the next section.

Assessing the Model Fit

The R_{1c} statistic of Glas (1988) is based on the comparison of the expected and observed frequencies of persons giving a positive or negative response to item i and obtaining a score of exactly r. Following Glas (1988), the R_{1c} statistic is calculated by the formula in Equation (1):

$$R_{1c} = \sum_r N_r^{-1} \boldsymbol{d}'_{.r} W_{.r}^{-1} \boldsymbol{d}_{.r}. \tag{1}$$

In Equation (1), N_r denotes the number of persons obtaining a raw score of r, $\boldsymbol{d}_{.r}$ denotes the vector of deviances between the observed and the expected frequency of persons obtaining a score of r. $W_{.r}$ is a variance–covariance matrix for the vector $\boldsymbol{d}_{.r}$.

The expected frequency of a positive response $E(n_{ri})$ is calculated by the formula in Equation (2):

$$E(n_{ri}) = n_r \frac{\varepsilon_i \gamma_{r-1}^{(i)}}{\gamma_r}. \tag{2}$$

In this formula, γ_r denotes the rth elementary symmetric function, and $\gamma_{r-1}^{(i)}$ denotes the $(r-1)$th elementary symmetric function after item i has been removed from the vector of item parameters. To calculate the elementary symmetric functions, we use the summation algorithm described by Gustafsson (1980). The variable n_r denotes the observed frequency of persons obtaining a score of r, and $\varepsilon_i = \exp(-\beta_i)$, where β_i is the item parameter of item i. The item parameters β_i were estimated using the conditional maximum likelihood approach (e.g., Molenaar, 1995).

Following a proof presented by Glas (1988), the R_{1c} statistic can be regarded as being asymptotically χ^2 distributed with $(k-1)(k-2)$ degrees of freedom, with k being the number of items in the test. The R_{1c} statistic was shown to have power against multiple violations of assumptions of the Rasch model, such as the axioms of unidimensionality and parallel item characteristic curves (Glas & Verhelst, 1995;Suárez-Falcón & Glas, 2003). For this reason, the R_{1c} statistic was chosen as test of global model fit for the study at hand.

Method of the Simulation Study

We now conduct a simulation study to analyze the extent to which the algorithm described in the previous section is correctly able to detect and reconstruct subsets of items that fit the Rasch model.

To evaluate the algorithm, item samples containing two subsets, both of which fit the Rasch model, were simulated. In each simulation, it was assessed whether the scale-constructing algorithm was able to distinguish between the items of the two scales. One reason for choosing this study design was that, to our knowledge, no study has been published so far that investigated the power of the R_{1c} statistic to detect between-item multidimensionality (Adams, Wilson, & Wang, 1997).

In all these simulations, response data were constructed by a computer program using the following algorithm: first, the item and person parameters were defined with previously set means and standard deviations. The distribution of the person parameter was set to be normal, while the item parameters were normally or equally distributed, depending on the simulation. After the parameters of every simulated item and every simulated person were set, the probability of a positive reaction and a random number between 0 and 1 were calculated for every person–item pair. If the random number was found to be smaller than the calculated probability of a positive reaction, the reaction was set to be positive for the person–item pair; otherwise, it was set to be negative.

In every simulation, a set of items was analyzed. Of these items, the first and the second half were independent item sets that fit the Rasch model. A computer program was written that implemented the algorithm described in the previous sections. In every simulation, it was assessed whether one of the two item sets fitting the Rasch model was reconstructed correctly by the algorithm.

The simulations differed in six aspects: the distribution of the item parameters (i.e., approximately normal or equal distribution), the standard deviations of the item and the person parameters, the size of the person sample, the size of the item set, and the correlation between the person parameters of the scales fitting the Rasch model.

Four different types of data sets were defined that varied in the standard deviations of their item and person parameters. In the first type of data set (defined as Type A), the person and the item parameters were set to be normally distributed with standard deviations of 1.0 and 0.5 for the person and item parameters, respectively. In the second type of data set (defined as Type B), the person and item parameters were set to have standard deviations of 1.5 and 0.5, respectively. In the third type of data set (defined as Type C), the standard deviations of the person and item parameters were set to be 2.0 and 1.5, respectively. In the fourth type of data set (defined as Type D), the person and item parameters were set to have standard deviations of 2.5 and 1.5, respectively.

Each of these data sets was combined, such that 10 different combinations of parameter distributions were analyzed. The size of the item sample varied between 10, 30, and 50, and the size of the person sample varied between 250, 500, and 1,000. In the first half of the simulated data sets, the correlations between the person parameters were set to 0.0, and in the second half, they were set to .5.

To perform the simulation study, a computer program was written that implemented the scale construction procedure with the R_{1c} test statistic of Glas (1988) as the test statistic of item fit, as described in the previous sections. As a termination criterion for the scale construction, the scale construction was set to halt as soon as the significance probability would become less than .05 for any further expansion of the scale.

To compare the results of the new method with those of a traditional method, all simulations were repeated with a PCA based on tetrachoric correlations. After each PCA, a varimax rotation was carried out. It was assessed whether the application of PCA resulted in two components, with all items of each item subset fitting the Rasch model showing their highest loading on the same component.

In our study, we calculated the tetrachoric correlations using an algorithm proposed by Brown (1977). In the simulations, parallel analysis (Horn, 1965) was used to determine the number of components to extract. Following previous studies (e.g., Tran & Formann, 2009; Weng & Cheng, 2005), the 95th percentile eigenvalues calculated from 10,000 random data matrices were chosen as the criteria for comparison. To obtain stable results, 10,000 simulations were carried out under each condition.

Results of the Simulation Study

In this section, the results of the simulation study are presented. For each simulated condition, the percentage of correct results over 10,000 replications is presented.

Because of space constraints, only the results of simulations with normally distributed item parameters are reported. Under all simulated conditions, the form of the distribution of the item parameters had only negligible effects on the results of the simulations.

To assess the accurateness of the results, the maximum standard error of all simulations was calculated. For all simulations, the standard error of the obtained results reached a maximum of 0.5.

The time needed to analyze each data set differed depending on the size of the analyzed person sample and item set. To illustrate a typical example, it should be reported that the analysis of the responses of 1,000 persons to 50 items took 7 seconds on the average if two data sets of type B were combined with each other using an Intel Core i7 processor.

Results of the Application of the Cluster Analytical Algorithm

The percentages of correct scale reconstructions under each simulated condition are given in Table 1 for simulations where the item parameters were approximately normally distributed.

Results of the Application of Principal Components Analysis

In the application of the PCA, a solution was considered as correct if (a) a two-component solution was obtained and (b) for each subset fitting the Rasch model, all items showed their highest loading on the same component. The results of the application of PCA and parallel analysis for simulations where the item parameters were approximately normally distributed are shown in Table 2.

A Practical Application

In this section, a practical application of the method described in this article is presented. To further evaluate the method, it was applied to dichotomous data obtained by testing 298 persons with a general battery of intelligence tests, the Basic Intelligence Functions (Intelligenz-Basis-Funktionen or IBF; Blum et al., 2005). The purpose of this analysis was to test whether the scales constructed by the algorithm would also pass traditional tests of fit to the Rasch model.

Description of the Sample

The analyzed sample contained the data of 298 persons (151 male, 147 female) all of whom participated in the IBF subtesting. The mean age of the sample was 23.9 years, with a standard deviation of 3.27. In all, 31 persons (10.4%) had an EU educational level of 2, 49 persons (16.4%) had an EU educational level of 3, 188 persons (63.1%)

Table 1. Percentage of Correct Scale Reconstructions With No Errors Under Different Conditions[a]

n	*i*	*r*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
250	10	.0	27.36	53.18	40.99	44.81	89.28	84.44	87.88	88.92	91.53	93.59
		.5	0.42	8.53	9.08	10.0	16.21	19.54	22.28	28.89	42.86	65.34
	30	.0	16.61	46.92	24.79	28.32	86.68	76.88	81.77	85.07	87.4	89.24
		.5	0.0	4.42	6.85	7.67	4.11	10.72	13.37	14.47	32.71	55.92
	50	.0	7.23	36.91	18.8	21.32	77.07	63.9	72.25	79.32	83.51	86.25
		.5	0.0	1.75	7.04	8.76	1.09	8.21	13.81	6.61	20.65	41.54
500	10	.0	78.01	84.91	73.05	76.89	92.07	91.56	93.02	93.2	93.65	93.96
		.5	3.31	20.31	10.05	9.64	66.3	44.12	39.47	71.29	77.95	89.23
	30	.0	76.83	83.71	63.71	71.4	90.16	88.33	89.11	89.19	90.58	90.91
		.5	0.36	22.22	9.31	7.83	55.36	25.07	22.3	63.54	73.0	84.83
	50	.0	63.11	73.73	49.81	57.85	83.05	81.27	83.38	85.11	86.84	87.01
		.5	0.02	21.63	10.1	9.56	34.91	19.14	17.35	47.73	59.7	76.99
1,000	10	.0	91.48	92.09	89.08	90.91	92.6	92.76	92.28	93.11	93.0	93.57
		.5	29.17	32.02	12.26	10.15	90.93	75.46	71.73	88.94	90.48	92.26
	30	.0	87.39	88.99	85.71	87.63	90.45	90.23	90.27	89.85	90.4	89.88
		.5	22.25	33.91	11.09	9.4	87.21	59.45	54.45	84.27	86.35	87.69
	50	.0	78.63	82.62	76.06	79.87	84.55	85.07	85.09	84.42	85.97	86.84
		.5	10.84	33.35	13.22	10.15	77.12	43.24	37.4	75.65	79.25	82.99

[a] Values are percentage of correct scale reconstructions with no errors under different conditions of person sample size (*n*), item set size (*i*), and correlation between the person parameters (*r*) for each combination of data sets A, B, C, and D when the item parameters were normally distributed.

had an EU educational level of 4, and 30 persons (10.1%) had an EU educational level of 5.

From the original sample of 298 persons, persons were excluded if they (a) cancelled the test, (b) showed very short response times combined with poor test performance, or (c) did not answer at least 75% of the items in a subtest. After the exclusion of persons showing deviant response behavior, the data of between 281 and 284 persons were used for the analysis of the four subtests of the IBF.

Description of the Test

The IBF test battery is a computerized intelligence test that consists of six subtests assessing verbal and numerical intelligence functions, long-term memory, and visualization. Verbal and numerical intelligence functions are each assessed by two subtests, and long-term memory and visualization are each assessed by one subtest. For each subtest there is a time limit. Items that are not answered before the end of the time limit are counted as not solved by the test participant.

In the IBF test, the test result contains the number of correct answers for each subtest and the factor scores for the verbal, numerical, visualization, and long-term memory tasks. The two subtests assessing numerical intelligence were not analyzed

Table 2. Percentage of Correct Results After Application of Principal Components Analysis Under Different Conditions[a]

n	*i*	*r*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
250	10	.0	98.31	99.79	99.75	99.76	100.0	99.84	99.68	97.97	96.47	94.35
		.5	49.5	67.43	69.91	71.65	92.14	95.25	96.65	95.94	95.93	94.31
	30	.0	82.47	49.12	0.14	0.0	18.2	0.0	0.0	0.0	0.0	0.0
		.5	70.1	42.9	0.18	0.0	22.03	0.02	0.0	0.0	0.0	0.0
	50	.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
500	10	.0	99.97	100.0	100.0	100.0	100.0	100.0	100.0	99.99	99.97	100.0
		.5	85.29	94.94	95.83	96.05	99.83	99.92	99.97	99.98	99.96	99.98
	30	.0	100.0	100.0	83.39	71.01	100.0	74.2	58.06	15.68	6.01	1.75
		.5	99.63	98.67	81.3	70.11	100.0	76.72	62.21	18.34	8.07	2.76
	50	.0	99.46	88.31	0.04	0.0	48.31	0.0	0.0	0.0	0.0	0.0
		.5	99.25	88.2	0.06	0.0	56.01	0.0	0.0	0.0	0.0	0.0
1,000	10	.0	100.0	100.0	99.97	99.98	100.0	99.98	99.97	99.94	99.91	99.91
		.5	99.46	99.97	99.9	99.93	100.0	99.97	99.96	99.96	99.95	99.91
	30	.0	100.0	100.0	99.49	99.64	100.0	99.46	99.44	98.67	98.04	96.64
		.5	100.0	99.98	99.42	99.43	100.0	99.48	99.55	98.68	98.11	97.39
	50	.0	100.0	100.0	93.58	86.95	100.0	90.46	77.97	26.84	9.53	1.9
		.5	100.0	100.0	93.68	88.18	100.0	90.78	80.62	30.36	12.68	3.23

[a] Values of percentage of correct results after application of principal components analysis under different conditions of person sample size (*n*), item set size (*i*), and correlation between the person parameters (*r*) for each combination of data sets A, B, C, and D when the item parameters were normally distributed.

because fewer than 240 test participants were able to answer 75% of the items within the test's time limit; we regard the resulting samples as too small for an analysis with the method described in this article. Items with missing responses were excluded from the test analysis. After the exclusion, 13 of the 17 items of the visualization subtest, 15 of the 20 items of the long-term memory subtest, 12 of the 16 items of the first verbal intelligence functions, and 15 of the 19 items of the second verbal intelligence functions subtest were analyzed.

Procedure

A computer program called Raschcon,[1] which implemented the scale-constructing algorithm analyzed in the simulation study, was used to analyze the IBF data set. The item set of each subtest was analyzed separately. The scale construction stopped as soon as the *p* values corresponding to the R_{1c} test statistics reported by Raschcon became less than .05. After four item sets had been constructed by Raschcon, their fit to the Rasch model was assessed by calculating Andersen likelihood ratios (Andersen, 1973), using the partitioning criteria of age, gender, and mean split on the basis of the test performance. The Andersen tests were computed using the eRm software package (Mair & Hatzinger, 2006). A level of significance of .01 was chosen, and an alpha adjustment was performed using the method of Holm–Bonferroni. Additionally, several item fit statistics for the Rasch model were calculated using Winsteps (Linacre, 2007), and a PCA of the residuals was calculated for each subtest. As in the simulation study, a PCA and parallel analysis of tetrachoric correlations were also performed for the data of each subtest.

Results

In the case of all subtests, the scale constructed by Raschcon was identical to the analyzed item set of the respective subtest. Each of the item sets constructed by Raschcon were subsequently analyzed for their fit to the Rasch model.

The Andersen tests indicate that the scales constructed by Raschcon fit the Rasch model with a level of significance of .01 in each subtest. It should be noted, however, that if a level of significance of .05 had been used, the test statistics used would have detected violations in three of the four subtests. The mean square infit and outfit statistics ranged between 1.33 and 0.65 for all four scales. After performing a PCA of the residuals, the eigenvalues of the first components reached values of 1.4 or less for the two verbal intelligence subtests and the visualization subtest, indicating that no additional dimensions were present in the data. In the case of the long-term memory subtest, the eigenvalue of the first component was 2.0, which could indicate that a second dimension is present in this subtest. In line with these results, a PCA of tetrachoric correlations led to a two-component solution for this subtest. For the visualization subtest, the PCA of tetrachoric correlations led to a one-component solution, whereas indefinite correlation matrices were obtained in the remaining two subtests. In general, the results of the analysis with Winsteps and eRm indicate that the scales constructed by Raschcon show a good fit to the Rasch model.

Discussion

In this article, a new algorithm was presented for the construction of scales that show a good fit to the Rasch model. The algorithm is based on a partial hierarchical cluster analysis and makes no specific assumptions regarding the model underlying the analyzed item set.

The R_{1c} test statistic of Glas (1988) was chosen as a test statistic to evaluate the practical use of the algorithm in a simulation study. The same algorithm was later implemented in the Raschcon computer program to apply it to real data. To assess its usefulness for practical research, a simulation study was carried out that compared the cluster analytic approach with another well-known method of exploratory data analysis, a PCA of tetrachoric correlations with varimax rotation.

We can try to identify the conditions under which the algorithm leads to correct results, and the reasons for the observed incorrect results. In the simulation study, the algorithm performed best if the simulated person sample was large and if the correlation between the person parameters was low. This trend is demonstrated by the high rate of correct scale reconstruction in simulations with a person sample of 1,000 and in simulations with a correlation of 0.0 between the person parameters. By comparison, the distribution of the item parameter seems to have only small effects on the correctness of the results of the algorithm. These results are in line with the results of previous studies on the R_{1c} test statistic (e.g., Suárez-Falcón & Glas, 2003).

It can also be observed that the algorithm performed better in data sets with smaller scales than in data sets that required the reconstruction of large scales. To some extent, this result can be explained by the higher probability in large item samples that at least one item has a response vector that is improbable under the assumption of the Rasch model, so the algorithm does not add it to the scale it constructs. The scale-constructing algorithm generally combines items that show highly probable response patterns under the assumption of a Rasch model. Therefore, items that happen to show an improbable response pattern are not included in the scale constructed by the algorithm, even if they are an a priori part of a scale that fits the Rasch model. The probability of such errors increases in larger item pools.

The simulation study revealed two different conditions under which the algorithms failed to correctly reconstruct one of the two scales that fitted the Rasch model. First, the algorithm leads to incorrect results more often if the person sample is small. Suárez-Falcón and Glas (2003) have obtained similar results by showing that the power of the R_{1c} test statistic to detect multiple model violations decreases in data sets with small samples. The second condition that leads to a failure of the algorithm is the combination of small variances of the item and person parameters and a significant correlation between the person parameters of the two scales.

The comparison of the cluster analytical algorithm with PCA and parallel analysis of tetrachoric correlations showed that the cluster analytical algorithm was in some cases to be preferred over this classical approach. This is most notably the case with the analysis of large item sets and small person samples, since the application of PCA was often not possible in these cases because of the occurrence of indefinite correlation matrices. Similar results have been previously reported by Weng and Cheng (2005) and Tran and Formann (2009). As a comparison of the results of both approaches suggests, the application of PCA may be preferred if there is a significant correlation between the person parameters. The generalizability of these results is limited by the generalizability of the conditions simulated in our simulation study. Since the application of PCA on tetrachoric correlations and the cluster analytical algorithm is based on different theoretical assumptions, there might be further conditions, such as the presence of guessing, which influence the relative efficiency of both approaches as well.

Notably, the result of the algorithm should not conclude the analysis process. The scales constructed by the algorithm meet several of the conditions that are necessary for fitting the Rasch model, but the application of additional statistical test procedures is recommended to determine the fit of the scales constructed by the Raschcon algorithm to the Rasch model. Several authors (e.g., Linacre, 1992; van der Linden & Hambleton, 1997) have already emphasized the importance of checking model assumptions with multiple model tests.

Future research could investigate the performance of the cluster analytical algorithm if alternative test statistics for the fit of the Rasch model are used. Based on the research by Suárez-Falcón and Glas (2003), it can be assumed that most of the test statistics analyzed in their study (e.g., the likelihood ratio test of Andersen, 1973) would lead to comparable results. Future research could also investigate the application of test statistics of other IRT models, such as the three-parameter logistic model (Birnbaum, 1968).

The application of Raschcon to the IBF data showed that the scales constructed by Raschcon may fit the Rasch model, even if the sample used for analysis is relatively small. The results of the simulation study still suggest that the use of larger samples would lead to more reliable results.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. A copy of Raschcon can be obtained from the first author.

References

Adams, R., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

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

Blum, F., Didi, H.-J., Fay, E., Maichle, U., Trost, G., Wahlen, H.-J., & Gittler, G. (2005). *Basic intelligence functions (IBF*). Mödling, Austria: Schuhfried.

Brown, M. B. (1977). Algorithm AS 116: The tetrachoric correlation and its asymptotic standard error. *Applied Statistics*, *26*, 343-351.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to mental test theory]. Bern, Switzerland: Huber.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H., Fischer & I. W., Molenaar (eds.), *Rasch models. Their foundations, recent developments and applications* (pp. 69-96). New York: Springer.

Gustafsson, J.-E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement*, *40*, 377-385.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525-546.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Linacre, J. M. (1992). Prioritizing misfit indicators. *Rasch Measurement Transactions*, *9*, 422-423.

Linacre, J. M. (2007). *Winsteps* (Version 3.61.2) [Computer software]. Chicago, IL: Winsteps. Retrieved from http://www.winsteps.com

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlsbaum Associates.

Mair, P., & Hatzinger, R. (2006). *eRm: Extended Rasch models. R package version 0.3.2*. Retrieved from http://CRAN.R-project.org/

Martin-Löf, P. (1973). *Statistiska modeller* [Statistical models]. Notes from seminars 1969-70 by Rolf Sundberg, 2nd ed.). Stockholm, Sweden: University of Stockholm.

Molenaar, I. W. (1995). Evaluation of item parameters. In Fischer, G. H., & Molenaar, I. W. (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 39-51). Berlin, Germany: Springer.

Ponocny, I. (2002). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *67*, 315.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*, 205-231.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, *3*, 25-40.

Suárez-Falcón, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127-143.

Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement*, *60*, 50-61.

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 statistics for the Rasch model. *Psychometrika*, *47*, 123-140.

Van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer.

Weng, L.-J., & Cheng, C.-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*, 697-716.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58-79.

Wright, B. D., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Study II

Debelak, R., & Tran, U.S. (2013). Principal component analysis of smoothed correlation matrices as a measure of dimensionality. *Educational and Psychological Measurement*, *73*, 63-77.

Role of the authors:

Rudolf Debelak wrote the first draft of the manuscript and also worked on all further drafts. He was responsible for the R code used in the simulation study and also designed the simulation study.

Ulrich S. Tran helped improving all drafts of the manuscript and also provided helpful feedback on the design of the simulation study.

Educational and Psychological
Measurement
73(1) 63–77
© The Author(s) 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164412457366
http://epm.sagepub.com
SAGE

Principal Component Analysis of Smoothed Tetrachoric Correlation Matrices as a Measure of Dimensionality

Rudolf Debelak[1] and Ulrich S. Tran[2]

Abstract

The application of principal component analysis and parallel analysis to smoothed tetrachoric correlation matrices was investigated in a simulation study. To evaluate the effect of several smoothing algorithms, 360 different types of data sets were simulated. Under each simulated condition, two item sets, each fitting a unidimensional two-parameter logistic model, were combined with each other. The simulations differed in the size of the simulated item sets, the size of the person samples, the distribution of the difficulty and discrimination parameters, and the correlation between the person parameters. In general, the application of a smoothing algorithm led to an improved performance in the assessment of dimensionality, but minor differences between the three investigated smoothing algorithms were found. Procedures to apply two of the three investigated smoothing algorithms via R software packages are presented.

Keywords

factor analysis, binary variables, parallel analysis, two-parameter logistic model

Introduction

The assessment of the dimensionality of an item set is a central issue in test theory, and many approaches to dimensionality assessment have been proposed and

[1] SCHUHFRIED GmbH, Mödling, Austria
[2] Department of Basic Psychological Research and Research Methods, University of Vienna, Vienna, Austria

Corresponding Author:
Rudolf Debelak, SCHUHFRIED GmbH, Hyrtlstraße 45, Mödling 2340, Austria
Email: debelak@schuhfried.at

discussed (e.g., Hattie, 1985; Reckase, 2009; Stout, 1987, 1990). Many writers have already emphasized the impact of a correct dimensionality assessment on practical psychological measurements (Green, 1983; Hattie, 1985) and the development of psychological theories (Weng & Cheng, 2005). Some of the standard methods to assess the dimensionality of an item set are based on exploratory factor analysis (EFA), which also includes principal axis factoring, and the related principal components analysis (PCA). The differences between these methods have been discussed, among others, by Crawford et al. (2010); Fabrigar, Wegener, MacCallum, and Strahan (1999); and Glorfeld (1995). These methods are typically based on the analysis of the matrix of Pearson product–moment correlations. In the case of binary items, this approach has been considered as critical. The factoring of Pearson product–moment correlations (or phi coefficients) may lead to spurious factors. The magnitude of the product–moment correlation of two binary items is limited by their difficulties (Carroll, 1945; Lord & Novick, 1968). Alternatively stated, their bivariate relation is not linear but nonlinear (McDonald & Ahlawat, 1974). Nonlinearity gives rise to extra factors.

As an alternative, the use of tetrachoric correlations instead of phi coefficients has been proposed. In contrast to phi coefficients, tetrachoric correlations are invariant to the item difficulties as long as the assumption of bivariate normality holds (Carroll, 1945; Lord & Novick, 1968).

To determine the correct number of components underlying a data set, a common suggestion is the application of parallel analysis (PA), a method originally proposed by Horn (1965). Several studies found evidence that PA is an accurate method to determine the number of underlying components (e.g., Humphreys & Montanelli, 1975; Zwick & Velicer, 1986). Since the original presentation of PA, several modifications to the original method have been proposed, like using the 95th percentile of the eigenvalue distribution from the simulated data as criterion for assessing the number of underlying components (e.g., Glorfeld, 1995). Recently, Crawford et al. (2010) compared multiple approaches for PA. Their study suggested that no single approach is clearly superior to the others, but that the results depended on the underlying factor structure. A recent study by Green, Levy, Thompson, Lu, and Lo (2012) also suggested a revised version of PA, which takes into account of the existence of prior factors in the determination of the critical eigenvalues of subsequent factors. Their results suggest that the suggested method leads to improved results in the presence of certain factor structures when compared with traditional methods.

Recent studies have already investigated the performance of the PA in retrieving unidimensionality in binary data (Tran & Formann, 2009; Weng & Cheng, 2005). One of the main problems in the application of PA and PCA of tetrachoric correlations to binary data identified by these studies is the presence of indefinite correlation matrices, which makes the application of PCA impossible.

The purpose of the present study was to investigate the performance of PCA and PA of smoothed tetrachoric correlations as an assessment of the dimensionality of sets of binary items in uni- and multidimensional item sets. We show that the investigated procedures perform well as assessments of dimensionality.

Two of the three investigated smoothing methods can be carried out with freely available software packages written in R (R Development Core Team, 2011). We present code that allows performing necessary computations.

This article is organized as follows. In the next section, we discuss several smoothing algorithms for indefinite correlation matrices described in the literature. Then previous studies on the application of PA to binary data are summarized. Following this the methods and results of a simulation study on the application of the improved PCA to matrices of tetrachoric correlations are presented. Finally, results are summarized and discussed.

Several Smoothing Procedures for Indefinite Matrices

One objection to the application of principal component analysis and factor analytic methods to tetrachoric correlation matrices is that these matrices might be indefinite (Lord, 1980). To apply factor analytical methods in these cases, several smoothing procedures have been suggested, of which some selected algorithms will be described below. For a general overview of earlier methods, see Devlin, Gnanadesikan, and Kettenring (1975); for an overview of current methods, see Yuan, Wu, and Bentler (2010).

The algorithm of Higham (2002) searches for the symmetric positive semidefinite matrix X with unit diagonal that is nearest to a given matrix A. The distance between two matrices is defined by the Frobenius norm, which is defined by $\|A - X\|_F = \sum_{i,j}(a-x)^2_{i,j}$. In his original work, Higham (2002) also investigated a more general approach that allowed the specific weighting of rows, columns, or even specific entries of $(A - X)$ in order to control changes to the correlation matrix during the smoothing process. The smoothing procedure of Higham (2002) has been implemented in the R package Matrix (Bates & Maechler, 2011).

Knol and Berger (1991) used the following smoothing procedure in their simulation studies: Let R be a possibly indefinite matrix of correlations, and let $R = VDV^{-1}$ be its eigen decomposition. A positive-definite correlation matrix $R^+(\delta)$ can be obtained by

$$R^+(\delta) = [Diag(VD^+V^{-1})]^{-0.5}VD^+V^{-1}[Diag(VD^+V^{-1})]^{-0.5}. \tag{1}$$

In this formula, D^+ denotes a modified diagonal matrix of eigenvalues obtained from D by replacing each eigenvalue below a predefined nonnegative threshold δ with δ. $Diag(VD^+V^{-1})$ denotes a diagonal matrix that contains the diagonal elements of the matrix VD^+V^{-1}. Although this algorithm was shown to provide acceptable results in the simulation studies reported by Knol and Berger (1991), to our knowledge it is not available in a publicly available software package.

Recently, Bentler and Yuan (2011) described another approach for obtaining a positive definite correlation matrix from an indefinite one. Given a symmetric indefinite matrix R, Bentler and Yuan (2011) show that a positive definite matrix R^* can be obtained by

$$R^* = \Delta R_0 \Delta + D_R. \tag{2}$$

In Equation (2), R_0 results from R by setting its diagonal entries to 0, D_R is the diagonal matrix containing the diagonal elements of R. Δ is a positive definite matrix that meets the condition that $D_R - \Delta^2(R - D)$ is positive definite, where D is a diagonal matrix with nonzero elements such that $(R - D)$ is positive semidefinite. Like the smoothing approach of Knol and Berger (1991), this method has not been explicitly implemented in a widely available software package yet. However, a variation of this method uses minimum trace factor analysis to determine R^*, as is demonstrated by Bentler and Yuan (2011). Functions contained in the R packages psych (Revelle, 2012) and Rcsdp (Bravo, 2010) may be used to carry out this approach in R.

The algorithm of Higham (2002) has been fully implemented in a software package written in R and can be easily applied to tetrachoric correlation matrices, which may also be calculated using R software (e.g., with the package psych; Revelle, 2012). In the next section, we investigate the results of several studies on the application of PCA on tetrachoric correlation matrices without applying a smoothing algorithm.

The Use of Parallel Analysis With Binary Variables

Determining the number of factors underlying a data set is important in the application of factor analytical methods, and a number of studies have focused on this issue. One of the most commonly used methods lies in the application of the Kaiser–Guttman rule, according to which all factors with eigenvalue greater than 1 (Guttman, 1954; Kaiser, 1960) are to be retained. This approach was evaluated by Bernstein and Teng (1989) with PCA of phi coefficients. They concluded that the Kaiser–Guttman rule leads to the overextraction of factors when applied to dichotomized variables.

Horn (1965) criticized this rule because it does not take sampling errors into account. He proposed to calculate a correlation matrix of random data of the same sample size and variable set size to determine the critical eigenvalues. Horn's approach became known as parallel analysis and was recently advocated by several writers for its use with PCA and EFA (Reckase, 2009; Weng & Cheng, 2005).

Recent studies also discussed the application of PA in detecting the number of factors in data sets of binary variables. Turner (1998) analyzed the application of PA to PCA and EFA and concluded that the common approach to determine the critical eigenvalues based on the size of the item set and the sample size may lead to the underextraction of factors or components because the presence of real factors in the data may influence the size of critical eigenvalues of random factors.

Green (1983) investigated the application of PA to EFA of phi coefficients of uni- and multidimensional data in which guessing was present. He found that PA of phi coefficients led to the extraction of spurious factors but argued that these factors could be identified in well-designed tests.

Weng and Cheng (2005) found that the application of PA performed well if the method was applied using the 95th or 99th percentile eigenvalues as the criterion for comparison, an approach that was also suggested by Glorfeld (1995). In their simulations, PA performed better with increasing sample size. Weng and Cheng (2005) found no difference in the performance of PA, using phi coefficients and tetrachoric correlations.

Tran and Formann (2009) further studied the performance of PA in PCA based on tetrachoric correlations. Using simulation studies, they showed that the latent structure of an item set that conforms to a unidimensional normal ogive model is not reliably uncovered by PA and PCA based on Pearson correlations, but the performance of this approach depends on the item discrimination and difficulty parameters. They also found that PA based on tetrachoric correlations performs better than PA based on Pearson correlations when applied to PCA of tetrachoric correlations. Nevertheless, they concluded that the usefulness of PCA is diminished in the presence of binary data. One reason for their verdict was the problem of indefinite correlation matrices, which makes the application of PCA impossible. The frequent presence of indefinite tetrachoric correlations matrices was also reported in other studies (e.g., Knol & Berger, 1991; Weng & Cheng, 2005). However, to our knowledge, a systematic investigation on prevalence rates under various conditions was never carried out.

More recently, Timmerman and Lorenzo-Seva (2011) investigated the application of PA to polytomous items. In their study, they found that a PA based on polychoric correlations performed best in determining the dimensionality in PCA. However, as in previous studies, nonconvergence of PCA due to indefinite matrices again posed a serious problem, with the total convergence rate reaching only 37.01% across their 10,400 simulated data matrices.

Method

We carried out a simulation study to compare results of PCA and PA with and without applying a smoothing algorithm and to investigate the prevalence of indefinite tetrachoric correlation matrices under different conditions in a more systematic way than previous research. In each simulation, the responses of a person sample to an item set consisting of two separate scales each fitting the two-parameter logistic model (Birnbaum, 1968) were simulated. The simulations varied in the following aspects: (a) the distributions of the item difficulty and item discrimination parameters in the simulated item sets (10 distribution combinations), (b) the size of the person sample (3 sizes), (c) the size of the item set (3 sizes), (d) the correlation between the person parameters in the simulated multidimensional data sets (4 correlations), and (e) the applied smoothing algorithm. Under each condition of the 10 × 3 × 3 × 4 ×

3 design, 1,000 data sets were simulated. Similar study designs were used by Tran and Formann (2009) and van Abswoude, van der Ark, and Sijtsma (2004).

Distributions of the Item Difficulty and Item Discrimination Parameters

Four types of scales were defined that differed in the standard deviation of the item difficulty parameters and in the mean of the distribution of the item discrimination parameters. The item difficulty parameters of two scale types (named A and B) followed a uniform distribution in the interval $\left[\frac{-\sqrt{3}}{2};\frac{\sqrt{3}}{2}\right]$ (resulting in a standard deviation of 0.5), whereas the item difficulty parameters in the other two scale types (named C and D) followed an uniform distribution in the interval $\left[\frac{-3\sqrt{3}}{2};\frac{3\sqrt{3}}{2}\right]$ (resulting in a standard deviation of 1.5). The distribution of the item discrimination parameters was log normal in all scale types. In scale types A and C, the item discrimination parameters were distributed LN(ln(0.5), 1), whereas in scale types B and D, the item discrimination parameters were distributed LN(ln(1.5), 1). In all scales types, the person parameter followed a standard normal distribution.

Size of the Item Sets

The simulated item sets consisted of 10, 30, and 50 items, respectively.

Size of the Person Samples

Three person sample sizes were used, containing 250, 500, and 750 simulated respondents, respectively.

Correlations Between the Person Parameters

The correlations between the person parameters was set 0, 0.4, 0.8, and 1.0, with the first three values resulting effectively in two-dimensional item sets, while the last correlation value resulted in unidimensional item sets. Given the person parameter θ_1 in the first scale, the person parameters θ_2 in the second scale were calculated using the formula

$$\theta_2 = Rand * \sqrt[2]{1 - r^2} + \theta_1 * r. \tag{3}$$

In this formula, r is the correlation between the person parameters, and *Rand* is a random variable that is normally distributed with standard deviation 1.

After the parameters of every simulated item and every simulated person were set, the probability of a positive reaction and a random number between 0 and 1 were calculated for each person–item pair. If the random number was smaller than the calculated probability of a positive reaction (‘‘1’’), the reaction was set as positive for the person–item pair; otherwise, it was set as negative (‘‘0’’).

In our study, tetrachoric correlation coefficients and PCA were calculated and carried out using functions of the R package psych (Revelle, 2012). The critical eigenvalues for the PCA were determined using a PA based on tetrachoric correlations, which were also carried out by using functions of the psych package. Horn (1965) advocated in his original approach comparing whether observed eigenvalues were greater than expected eigenvalues, that is, the mean of the respective eigenvalue distribution. We used the median (= 50th percentile) of the respective eigenvalue distribution in our study, calculated from 1,000 random data matrices. Means and medians differed by less than 0.01 across all eigenvalue distributions in our study, which is in line with Glorfeld (1995).

If an indefinite correlation matrix was obtained in the simulations, a smoothing algorithm was applied. In these cases, PCA was applied to the smoothed tetrachoric correlation matrices. All simulations were replicated using each of the smoothing algorithms described in the section ‘‘Several Smoothing Procedures for Indefinite Matrices.’’ The smoothing algorithm of Higham (2002) was applied with the R package Matrix (Bates & Maechler, 2011). The smoothing algorithm of Bentler and Yuan (2011) was applied in R using functions of the packages psych (Revelle, 2012) and Rcsdp (Bravo, 2010). See the appendix for R code to use these two algorithms. The smoothing algorithm of Knol and Berger (1991) was applied using software which was developed specifically for this study in Java. In this software, the constant δ was set to 0.

Results of the Simulation Study

Our simulations confirmed the results of previous studies, which indicated the frequent presence of indefinite tetrachoric correlation matrices in the application of PCA to binary variables (Timmerman & Lorenzo-Seva, 2011; Tran & Formann, 2009; Weng & Cheng, 2005). Table 1 displays the relative frequencies of indefinite correlation matrices in all simulated data sets in which the correlation between the person parameters was 0.

To assess the practical usefulness of PCA of smoothed tetrachoric correlations, we report the principal results of our simulation study in several tables. Table 2 shows the results without application of a smoothing algorithm to the data. In unidimensional data sets, similar results were obtained. Table 3 shows the results of the analysis of multidimensional data sets when the correlation matrices were smoothed according to the algorithm of Higham (2002) before applying PCA. The smoothing algorithm of Knol and Berger (1991) led to results mostly comparable to those of the Higham algorithm; therefore, they will not be reported in detail here. The results of the smoothing algorithm by Bentler and Yuan (2011) are reported in Table 4.

Table 1. Rate (in Percent) of Indefinite Correlation Matrices in Simulated Data Sets With a Correlation of 0.0 Between the Person Parameters for Given Sizes of the Samples (*n*) and the Item Sets (*i*) for Combination of the Scale Types A, B, C, and D

n	*i*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
250	10	0	1.9	0.4	6.5	8	3.3	17.5	0.6	8.8	25.2
	30	43.4	98.1	65.2	99.6	100	98.4	100	76.2	99.8	100
	50	100	100	100	100	100	100	100	100	100	100
500	10	0	0.1	0	1.2	0.9	0.1	3.3	0	1.8	8.1
	30	3.5	61.5	10.7	87.5	92.9	69.7	98.2	20.2	91.1	99.8
	50	47.1	100	72.7	100	100	99.9	100	87.8	100	100
750	10	0	0	0	0.7	0.4	0.1	1.7	0	0.7	4.3
	30	0.4	31.7	3.3	68.3	71.4	39	89.6	5.7	73.9	96.7
	50	13.7	93.4	34.5	99	99.8	97	99.9	51.8	99.8	100

Table 2. Rate (in Percent) of Correctly Detected Dimensionality for Each Combination of the Scale Types A, B, C, and D, With Given Size of the Person Sample (*n*), Item Set (*i*), and Correlation Between the Person Parameters (*R*) When Smoothing Was Not Applied

R	*n*	*i*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
0.0	250	10	40.9	64.1	37.6	55.9	84.4	58.3	72.2	35.3	51.9	61.2
		30	23.6	1.7	11.6	0.4	0	1.2	0	5.9	0.1	0
		50	0	0	0	0	0	0	0	0	0	0
	500	10	47.8	75.9	45.1	70.4	95	73.9	89.5	43.7	66.8	84.6
		30	61.8	38.7	49.4	11.7	7.7	30.4	2.2	36.8	8.5	0.3
		50	38.8	0.2	15.8	0	0	0.1	0	6	0	0
	750	10	52.8	80.6	50.4	76.8	96.9	77.9	94.2	50.8	74.7	90.1
		30	73.8	68	63.3	30.6	29.5	60.7	11.6	53.4	24.9	4.1
		50	73.6	6.9	47.9	0.7	0.2	3.8	0	29.8	0.4	0
0.4	250	10	38.2	58.2	37.2	51.9	81.7	56.7	71	36.5	47.5	62
		30	25.5	2	13.8	0.3	0	0.9	0	7.3	0.1	0
		50	0	0	0	0	0	0	0	0	0	0
	500	10	46.4	70.2	44.5	67.6	95.3	69.3	89.6	43.1	65.9	85.6
		30	63.4	37.8	49	11.9	8	27.5	1.9	37.5	8.9	0.7
		50	37.7	0.4	17.5	0	0	0.2	0	6.3	0	0
	750	10	52.8	77	52.4	74.2	97.6	74.6	94.3	50.3	72.9	91.3
		30	72.9	70.7	63.6	31	30.7	60.6	11.7	53.2	25.4	3.1
		50	73.3	7	48.6	0.5	0.1	4.4	0	29.7	0.4	0
0.8	250	10	25.9	16.2	26.8	18.7	6.2	18.6	8.7	26.9	21	10.9
		30	14.2	0.4	10.3	0	0	0.2	0	5.8	0	0
		50	0	0	0	0	0	0	0	0	0	0
	500	10	22.4	11.4	24.1	14	4.1	12.5	5.2	24.5	15.3	7.8
		30	25.8	11.9	26	4.9	7.6	8.8	1.8	24.9	3.7	0.3
		50	22.6	0.4	9.5	0	0	0.1	0	5.3	0	0
	750	10	20.5	8.8	21.4	10.3	2.3	10.3	3.9	24.5	13	5.4
		30	30.7	30.9	30.5	14	33.4	27.7	10.6	30.9	12.4	3.5
		50	54.9	8	39.4	0.4	0.3	3.6	0	24.8	0.6	0

Table 3. Rate (in Percent) of Correctly Detected Dimensionality for Each Combination of the Scale Types A, B, C, and D, With Given Size of the Person Sample (*n*), Item Set (*i*), and Correlation Between the Person Parameters (*R*) When the Smoothing Algorithm of Higham Was Applied

R	*n*	*i*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
0.0	250	10	41.4	66.4	38.3	59.6	92.2	62	87.4	34.9	57.4	83.1
		30	49.8	89.8	38.5	66.7	99.4	86.4	89.7	30.9	58.1	75.5
		50	63.2	96.6	49.9	60.5	99.9	94	84.1	34.9	48.5	54.7
	500	10	49.4	75.4	48	72.2	95.5	75.4	93.8	43	70.4	92.9
		30	65.5	95.7	57	81.2	100	92.9	96.3	48.5	77.3	88.6
		50	80.1	99.1	67.3	70.1	100	98.2	87.8	53.8	68.7	69.5
	750	10	53.9	81.5	49.1	78.9	98	78.6	97.8	51.4	76.5	96.5
		30	76.5	98.4	68	90.5	100	96.5	98.3	57.5	85.9	95.4
		50	85.1	99.6	76.8	79.7	100	98.8	92.8	66.1	74.2	76.2
0.4	250	10	38.7	58.9	37.3	54.5	90.1	57.5	85.8	36.8	54.7	83.2
		30	49.6	88.9	43.8	64.6	99.7	85.8	91.5	32.9	58.9	77.3
		50	64.7	97.7	51.6	60.9	100	93.8	85	39.7	52.2	58.7
	500	10	47.3	69.6	44.8	69.1	96.1	70.6	94	44.9	67.1	93.3
		30	65.9	95.1	57.8	83.1	100	93.9	97.4	51.2	76.6	89.1
		50	76	99.5	67.2	73.3	100	97.7	90.4	53.9	68.8	72.6
	750	10	54.5	79.1	52.8	74.2	98	75.3	97.3	50.6	74.1	96.8
		30	73.2	98	67.5	89.7	100	96.8	98.7	58.2	86.5	95.8
		50	86	99.6	77.9	79.6	100	99.7	91	67.7	75.5	75.6
0.8	250	10	24.9	16.4	27.4	20.8	7.4	18.2	10.6	24.6	23.2	16.2
		30	24.2	17.8	29.5	28.2	57.3	21.9	57.4	24.4	31.6	59.6
		50	33.2	40.9	34	50.3	95.5	44.8	85	33.4	48.6	54
	500	10	22.5	11.2	24.4	14.8	3.5	14.7	5.1	23.2	14.6	10
		30	25.5	34.4	31	40.7	91.8	36.6	88	32.8	39.8	82.4
		50	50.5	78.4	49.8	70.5	100	79.8	93	47.2	68.6	61.6
	750	10	19	7.7	20.4	10.6	2.5	9.4	3.3	24.4	11	4.6
		30	29.6	42.7	31.2	48.5	98.4	50.7	96.7	31.8	52.4	91.6
		50	65.1	94.2	65.6	80.8	100	91.5	92.5	62.4	77.4	65

As can be seen from Tables 3 and 4, the algorithm of Bentler and Yuan (2011) led to more correct solutions compared with the algorithm of Higham (2002) when the correlation between the person parameters was low, but to fewer correct solutions when it was high. As was to be expected, the number of correct results in general increased with person sample size and decreased with the number items. In general, the number of correct results did not decrease when the correlation between the person parameters of the two item sets increased.

The results of the simulations without application of smoothing in unidimensional data sets are not reported here in detail, since the results were comparable to those observed in bidimensional data sets. For the unidimensional data sets, the results of the application of the smoothing algorithm by Higham (2002) are presented in Table 5. Table 6 shows the results of the application of the smoothing algorithm of Bentler and Yuan (2011).

Table 4. Rate (in Percent) of Correctly Detected Dimensionality for Each Combination of the Scale Types A, B, C, and D, With Given Size of the Person Sample (*n*), Item Set (*i*), and Correlation Between the Person Parameters (*R*) When the Smoothing Algorithm of Bentler and Yuan Was Applied

R	*n*	*i*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
0.0	250	10	40.8	63.7	35.5	59.8	92.1	60.2	86.5	36.2	58.4	83.3
		30	52.3	91.9	46	80.5	99.7	88.1	97.6	40	77.9	91.6
		50	85.5	99	83.7	95.8	100	99.1	99.4	81.1	94.1	97.2
	500	10	46.3	75.8	43.2	71	96.4	73	93.2	45.5	68.4	93.2
		30	61.9	95.3	56.1	87.3	99.9	94.5	98.2	48.3	83.8	96.4
		50	77.3	99.4	73.2	92	100	98.7	98.9	66.5	88.2	93.8
	750	10	51.9	78.2	52.9	78	97.7	78.6	97.2	51	76.3	96.3
		30	72.5	97.1	66.2	91.4	100	96.6	99.3	60.1	91.5	96.9
		50	85.6	99.7	76.6	91.4	100	99.2	98	69.4	89.9	92.9
0.4	250	10	36.5	59.5	37.1	56.8	90.6	59.2	88	38.8	51.9	82.2
		30	53.2	92	47.3	80.5	99.4	88.6	97.4	42.6	77.5	92.6
		50	88.6	99.5	85.2	96.1	100	99.2	99.4	84.6	94.7	98
	500	10	45.5	68.7	46.3	69.5	97.1	68.1	94.3	43.2	68	93.6
		30	66	95.8	57.3	87.4	99.9	94.1	99.1	49.1	84.6	96.8
		50	80.6	99.7	74.8	89.5	100	98.6	98.7	67	89.1	96.7
	750	10	51.9	76.1	51.7	77.4	97.7	75.2	96.1	49	75.6	97.2
		30	72.9	97.9	66.7	91.9	100	96.6	99.2	59.4	89.9	97.6
		50	86	99.7	80.6	92	100	99.3	98.3	71.5	90.4	95.2
0.8	250	10	26.6	16.9	27.8	18.5	6.9	17.9	10	27	21.8	12.4
		30	23.1	13.5	25.7	16.6	44.7	13.7	37.2	26.9	19.9	36.5
		50	10.3	16.4	11	12.6	80.6	15.2	63.3	12.2	11.3	47.9
	500	10	21	11.1	23.5	13.1	4.1	11.2	4.7	24.6	15.7	7.9
		30	27.6	29.8	28.9	33.5	88.6	31	82.5	33.9	37.4	79.5
		50	46.1	72.4	42	69.7	100	68.1	98.9	42	64.9	93.3
	750	10	20.1	8.9	22.2	10.1	2.6	10.2	3.8	25.2	13.8	6.1
		30	31.3	44	30.8	42.5	98.4	45.8	96.7	34.8	51	94.1
		50	66	91.4	66.1	87.2	100	91	99	59.8	83	91.2

In unidimensional item sets, the number of correct results increased with person sample size but decreased with the number of items. Both in bidimensional and in unidimensional item sets, the size of the change depended on the distributions of the item parameters. Similar results have been previously reported by Tran and Formann (2009) and Weng and Cheng (2005). The smoothing algorithm by Bentler and Yuan (2011) led to improved results when compared with the smoothing algorithm of Higham (2002), especially in the analysis of large item sets.

Across all 90 simulation conditions of unidimensional data sets, the algorithm of Bentler and Yuan led 54 times to better results than the other two algorithms. In two-dimensional data sets, this number was 52 out of 90. However, when the correlation was 0.8, the Bentler and Yuan algorithm achieved better results only in 18 out of 90 conditions. Under these conditions, the algorithm of Knol and Berger performed best.

Table 5. Rate (in Percent) of Correctly Detected Dimensionality for Each Combination of the Scale Types A, B, C, and D, With Given Size of the Person Sample (*n*) and Item Set (*i*) for Each Data Set Combination When the Smoothing Algorithm of Higham Was Applied

n	*i*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
250	10	65.7	88.2	62.1	80.5	97.4	84.7	95.9	58.4	78	91.7
	30	79.7	97.2	71	84	100	96.7	94.2	61.9	74.8	66.6
	50	87.6	99.8	82.3	73.9	100	98.3	84	68	58.8	29.5
500	10	75.9	93.1	70.4	90.5	99.5	92	99.1	67.8	89.5	96.8
	30	87.3	99.3	81.9	93.5	100	98.7	98.1	78.3	86.9	74.5
	50	95.1	99.9	90.2	83.5	100	99.6	88.5	79.4	69.6	33.2
750	10	77.6	95.3	76.5	94.5	99.6	94.3	99.8	73	92.1	98.4
	30	92.9	99.5	88.5	95.7	100	99.1	98.9	85	93	81.7
	50	97.6	99.9	93.6	78.8	100	99.8	91.4	91	72.8	38.8

Table 6. Rate (in Percent) of Correctly Detected Dimensionality for Each Combination of the Scale Types A, B, C, and D, With Given Size of the Person Sample (*n*) and Item Set (*i*) for Each Data Set Combination When the Smoothing Algorithm of Bentler and Yuan Was Applied

n	*i*	A-A	A-B	A-C	A-D	B-B	B-C	B-D	C-C	C-D	D-D
250	10	63.1	87.6	59.3	80	97.8	85.2	96.3	57.6	77.1	91.6
	30	80.1	98.9	77.7	96.1	100	97.7	98.7	76	92.6	88.3
	50	98.4	100	97.4	99.8	100	100	99.9	97	98.6	90
500	10	74.4	93.6	71.2	90.3	99.4	93.1	99.3	70	86.4	97.8
	30	87.4	99.1	83.7	95.2	100	99.1	99.3	78	93.5	90.4
	50	94.8	100	93.2	97.9	100	99.8	99	89.5	94.3	81
750	10	76.5	94.7	75	93.6	99.6	94.1	99.7	74.5	92.5	98.7
	30	92.3	99.9	90.7	98.5	100	99.8	99.3	87.2	96.5	91.5
	50	97.1	99.9	94.4	97.8	100	99.8	99	91	93.6	75.1

We also evaluated how often each smoothing algorithm achieved an accuracy percentage of 95% or higher for all simulated conditions. In this evaluation, the Bentler–Yuan algorithm generally achieved again the best results. In unidimensional data sets, an accuracy percentage of more than 95% was achieved for 44 out of 90 conditions. The Higham and the Knol–Berger algorithms achieved this accuracy only in 32 and 21 conditions, respectively. In bidimensional data sets, high accuracy was achieved less often when the correlation between the person parameters in both scales increased. When this correlation was 0, the Bentler–Yuan algorithm achieved an accuracy percentage of more than 95% in 29 simulated conditions, compared with 13 for the Knol–Berger algorithm and 21 for the Higham algorithm. When the correlation between the person parameters was high, all smoothing algorithms achieved accuracy percentages of more than 95% only in four to six conditions.

Discussion

The results obtained in the simulation study are in line with the results of past studies on the PA of tetrachoric correlations (Tran & Formann, 2009; Weng & Cheng, 2005). The presence of indefinite correlation matrices often precludes the application of PCA and PA and underlines the need for a smoothing algorithm. In our simulation study, we observed that indefinite matrices of tetrachoric correlation matrices tended to occur if the analyzed item set is large, the analyzed person sample is small, and the discrimination parameters of the item sets are large. We emphasize that the calculation of the tetrachoric correlation coefficients is not advisable if the assumption of a bivariate normal distribution of the latent variables is likely to be violated in the data (e.g., Lord, 1980). This may be the case if guessing is present. It should be noted that this assumption was not violated in our simulations.

Application of smoothing algorithms generally improved correct identification of dimensionality when the correlation between the latent dimensions was 0.0 or 0.4 in our simulations. When the correlation between the person parameters of the two scales was 0.8, the results were less reliable and more dependent on the size of the person and item sample. In general, the results improved when the size of the analyzed multidimensional item set and person sample increased and when the discrimination parameter increased. We also observed minor differences in the performance of the three smoothing algorithms used in our study. In data sets with a clear dimensional structure, that is, in unidimensional data sets and multidimensional data sets with a low correlation between the underlying dimensions, the algorithm of Bentler and Yuan (2011) performed best, especially when large item sets were analyzed.

In summary, our results seem to indicate that the application of PCA and PA to binary data seems to assess the dimensionality of a multidimensional item set correctly if the correlations between the dimensions are low to medium and when a smoothing procedure is applied.

When applied to unidimensional item sets, PCA and PA of smoothed correlation matrices led to correct results in most cases that were investigated. However, this was not the case with unidimensional data sets in which a large item pool with high discrimination parameters and a wide range of difficulty parameters was combined with a small person sample. Our interpretation of this observation is that in these cases each simulated data set contained at least some items with a very high or very low difficulty parameter that were positively answered by almost all (or very few) persons in the person sample. The tetrachoric correlation coefficients between items with extreme difficulty parameters were generally very low, which led to the extraction of multiple components in PCA.

Although our study was based on the original approach of Horn (1965), future studies should investigate the performance of smoothing algorithms when used in more recent variations of PA (i.e., using the 95th percentile of the eigenvalue distribution as criterion). The evidence collected so far in this and similar studies (Crawford et al., 2010; Green et al., 2012) seems to suggest that no single algorithm leads to optimal results, but the results of each algorithm depend on the underlying factor structure.

Appendix

Sample Code for Matrix Smoothing in R

In our simulation study, we applied two smoothing algorithms using functions of open source software written in R. This appendix contains sample code that demonstrates how these algorithms can be applied. It is assumed that R is a symmetric indefinite matrix that should be approximated by a positive-definite matrix R2.

The first smoothing algorithm, which was proposed by Higham (2002), uses code from the Matrix package. The code for using this algorithm is as follows:

```
library(Matrix)
R2 <- nearPD(R, corr=TRUE)[[1]]
```

The second smoothing algorithm can be used by functions of the packages psych and Rcsdp. The code for using this algorithm will be presented in two steps:

```
library(psych)
library(Rcsdp)
av <- as.vector(rep(2.5,i))
ew <- glb.algebraic(R,UpBounds=av)$solution
```

This code calculates the communalities of each variable after applying a minimum trace factor analysis as values of the vector ew. The used functions require the definition of upper limits for these communalities. In our code, these limits were defined as 2.5. The number of items is denoted by i. In a second step, the correlations are rescaled using a scaling constant. In line with the recommendations of Bentler and Yuan (2011), we used the scaling constant .96. A simple code for rescaling R is as follows:

```
const <- .96
R2 <- R
for (a in 1:i){
    if (ew[ a] >= 1.0) {
        for (b in 1:i) {
            if(a!=b) {
                 R2[a,b] <- R[a,b] * sqrt(const/ew[a])
                 R2[b,a] <- R[a,b]
            }
        }
    }
}
```

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Bates, D., & Maechler, M. (2011). *Matrix: Sparse and dense matrix classes and methods* (R Package Version 1.0-6). Retrieved from http://CRAN.R-project.org/package=Matrix

Bentler, P., & Yuan, K.-H. (2011). Positive definiteness via off-diagonal scaling of a symmetric indefinite matrix. *Psychometrika, 76*, 119-123.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison & Wesley.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics, 30*, 213-225.

Bravo, H. C. (2010). *Rcsdp: R interface to the CSDP semidefinite programming library* (R Package Version 0.1-41). Retrieved from http://CRAN.R-project.org/package=Rcsdp

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika, 10*, 1-19.

Crawford, A., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D. S., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement, 70*, 885-901.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika, 62*, 531-545.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*, 377-393.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary variables. *Applied Psychological Measurement, 7*, 139-147.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement, 72*, 357-374.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-161.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Higham, N. (2002). Computing the nearest correlation matrix: A problem from finance. *IMA Journal of Numerical Analysis, 22*, 329-343.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*, 193-206.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Knol, D., & Berger, M. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457-477.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

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

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

R Development Core Team. (2011). *R: A language and environment for statistical computing* (ISBN 3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reckase, M. D. (2009). *Multidimensional item response theory*. Dordrecht, Netherlands: Springer.

Revelle, W. (2012). *psych: Procedures for personality and psychological research* (R Package Version 1.2.4). Retrieved from http://CRAN.R-project.org/package=psych

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55*, 293-326.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*, 209-220.

Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement, 60*, 50-61.

Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement, 58*, 541-568.

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

Weng, L.-J., & Cheng, C.-P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*, 697-716.

Yuan, K.-H., Wu, R., & Bentler, P. M. (2010). Ridge structural equation modeling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology, 64*, 107-133.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Study III

Debelak, R., Gittler, G., & Arendasy, M. (submitted). On gender differences in mental rotation processing speed. *Learning and Individual Differences.*

Submission confirmation received on 23th February 2013.

Role of the authors:

Rudolf Debelak wrote the first draft of the manuscript and also worked on all further drafts. He carried out all calculations reported in this manuscript.

Georg Gittler provided the data of the endless loop task. He also gave helpful feedback on psychological theories behind mental rotation and provided feedback on the manuscript.

Martin Arendasy wrote part of the theoretical introduction and the description of the cube comparison task. He also provided feedback on the manuscript.

On Gender Differences in Mental Rotation Processing Speed

Rudolf Debelak

SCHUHFRIED GmbH

Georg Gittler

University of Vienna

Martin Arendasy

University of Graz

Author Note

Rudolf Debelak, Department of Psychology, SCHUHFRIED GmbH, Mödling, Austria; Georg Gittler, Department of Applied Psychology: Health Development, Enhancement and Intervention, University of Vienna, Austria; Martin Arendasy, Psychological Methods and Computer-based Modeling Group, University of Graz, Austria

Correspondence concerning this article should be addressed to Rudolf Debelak, Department of Psychology, SCHUHFRIED GmbH, Hyrtlstraße 45, 2340 Mödling, Austria.

Abstract

There is a wide consensus in the literature that gender differences can be observed in tasks measuring mental rotation ability. A possible explanation of this finding is the presence of gender differences in the processing speed of mental rotation tasks. In two studies, we investigated the dimensionality and the presence of gender differences in mental rotation processing speed in two mental rotation tasks. By applying a joint modeling approach for responses and response times, we found that, in both tasks, mental rotation ability and mental rotation processing speed can be regarded as unidimensional constructs. We replicated previous findings that gender differences in mental rotation ability can be observed in both tasks, although we could not find gender differences in mental rotation processing speed. Our results thus indicate that the observed gender differences in mental rotation ability cannot be explained by gender differences in mental rotation processing speed.

Keywords: Mental Rotation Ability, Mental Rotation Speed, IRT model, Gender Differences

1. Introduction

Spatial abilities constitute an important component in current models of human intelligence (cf. Carroll, 1993; Johnson & Bouchard, 2005; McGrew, 2005). Studies on gender differences in human intelligence indicate that spatial ability measures exhibit considerable gender differences in favor of male subjects. Meta-analytic studies show that gender difference is particularly pronounced in case of three-dimensional mental rotation tasks (e.g., Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995). Although there is evidence that the observed gender difference in favor of male subjects is stable across cohorts (cf. Masters & Sanders, 1993; Voyer et al., 1995), age (cf. Linn & Petersen, 1985), and culture (Silverman, Choi & Peters, 2007), there is evidence that the magnitude of the gender difference varies with item design characteristics (cf. Arendasy & Sommer, 2010) and general design characteristics such as time limit. Numerous explanations have been proposed to account for the observed male superiority in three-dimensional mental rotation performance (for an overview: Halpern, 2000). Some models attributed gender differences in mental rotation tasks to gender differences in working speed. This explanation is based on findings, which indicate that gender differences in favor of male subjects decrease in effect size once time limits had been removed from the test (cf. Goldstein, Haldane & Mitchell, 1990). Although this finding has not usually been consistently replicated (e.g., Masters, 1998), a recent meta-analysis conducted by Voyer (2011) indicate that gender differences in paper-pencil mental rotation tasks indeed decrease in size when the psychometric measures were administered without time limits. This finding could be due to at least two different reasons: (1) the removal of time limits may allow female respondents that are not well trained in this population to utilize effective mental rotation strategies (cf. Arendasy, Sommer, Hergovich, & Feldhammer, 2011; Arendasy, Sommer, & Gittler, 2010), or (2) the observed reduction of the gender difference in the untimed administration condition could be due to a ceiling effect in the male population

(Voyer, 2011). Because none of these previous studies assessed mental rotation processing speed, it is hard to differentiate between these two explanations. With the model by Goldstein et al. (1990) as basis, one would expect that gender differences in mental rotation processing speed are either more pronounced or of the same magnitude than gender differences in mental rotation accuracy. Furthermore, accuracy and processing speed in solving three-dimensional mental rotation tasks should be at least moderately correlated. By contrast, if the reduced effect size of the gender difference in three-dimensional mental rotation performance is mainly attributable to a ceiling effect in the male population in case of untimed mental rotation tasks, one would expect either no or small effect sizes of the gender differences in mental rotation processing speed, whereas observed gender differences in mental rotation accuracy should be large in magnitude compared with the processing speed measure.

1.1. Formulation of the Problem

In this article, we want to evaluate these two conflicting hypotheses using an item response theory model that enables the simultaneous estimation of accuracy and processing speed parameters (Klein Entink, Fox & van der Linden, 2009). Another advantage of this psychometric approach is the possibility of simultaneously evaluating the dimensionality of accuracy and processing speed measures of mental rotation performance, which have been debated in the literature for some time because of the possibility of solving mental rotation tasks using different solution strategies. We investigated this problem in two separate studies, which used different mental rotation tasks.

2. Method

2.1. A Multivariate Multilevel Approach for Modeling Speed and Ability

In the literature, multiple approaches for modeling speed have been described (for an overview of early approaches, see van der Linden & Hambleton, 1997). This study chose an approach that has been originally proposed by Klein Entink, Fox et al. (2009). To simultaneously model speed and ability, this approach defines a multivariate multilevel approach for modeling responses and response times under a Bayesian framework. Under this framework, prior distributions for each model parameter are assumed, which reflect the researcher's beliefs on each parameter's distribution before data are collected. After data have been collected, the prior distributions are updated based on the data and Bayes' theorem, resulting in a posterior distribution for each model parameter, which can be used for making inferences. For an introduction to Bayesian item response theory, see Fox (2010).

On the first level of this approach, two separate models for responses and response times are defined. The model for responses is the two-parameter normal ogive model, which defines a person parameter θ_i, which marks the ability of person i to answer items correctly. The model further defines two item parameters for each item k, which define the respective item's difficulty b_k and discrimination a_k. This model contains the one-parameter normal ogive model as a special case, in which the discrimination parameter is regarded as fixed for all items and which is closely related to the Rasch model (Rasch, 1960; cf. Embretson & Reise, 2000). In the two-parameter normal ogive model, the probability that person i answers item k correctly is given by the following:

$$P(+|\theta_i, a_k, b_k) = \Phi(a_k\theta_i + b_k) \qquad (1)$$

In this formula, $\Phi()$ denotes the cumulative function of the standard normal distribution. The response times are described by the two-parameter log-normal model. As in the two-parameter normal ogive model, two item parameters are defined for each item that describes the respective item's time intensity and time discrimination. For each person i, a speed parameter is defined. This model contains the one-parameter log-normal model as a special case, in which the time discrimination parameter is set to a fixed value.

In the two-parameter log-normal model, the log response time T_{ik} of a person i working on item k is given by the following:

$$T_{ik} = -\phi_k \zeta_i + \lambda_k + \varepsilon_{ik} \tag{2}$$

In formula (2), ζ_i denotes the respondent's speed, λ_k denotes an item's time intensity, and ϕ_k is an item's time discrimination. ε_{ik} is a residual term, which is assumed to be normally distributed with an item-specific variance.

Following Goldhammer and Klein Entink (2011), the approach of Klein Entink et al. (2009) assumes that the speed and ability of a person can be regarded as fixed as long as a person is working on the test and that the responses and response times are independent and conditional on the respective person parameter.

On the second level, the approach of Klein Entink et al. (2009) defines additional models for the person and item parameters of the first level models. These second-level models define joint distributions for the person and item parameters of the first-level models. For the person parameters defining ability and speed, a bivariate normal distribution is defined as a common prior distribution:

$$(\theta, \zeta) = (\mu_\theta, \mu_\varsigma) + e_p \tag{3}$$

In formula (3), e_p follows a bivariate normal distribution with mean 0.

This second level model provides information on the variance of speed and ability in the investigated population and on the correlation between them. As Klein Entink, et al. (2009) noted, it can be extended to include person level covariates that may explain some of the variance of the person parameters.

For the item parameters describing difficulty, discrimination, time intensity, and time discrimination, an analogous model, which uses a multivariate normal distribution as prior distribution, is defined. This second-level model provides information on the variance of all item parameters and the dependencies between them.

2.2. Model Selection and Estimation

Under the presented Bayesian framework, several criteria have been proposed for model selection, one of them being the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, van der Linde, 2002; see also Fox, 2010; Gelman, Carlin, Stern & Ruben, 2004). This criterion combines a term measuring the deviance of a model with a term measuring its complexity. It has been already used in a number of studies for model selection (e.g., Goldhammer & Klein Entink, 2010).

In our study, we estimated all model parameters using a Gibbs sampling approach, which has been implemented in the software package cirt (Klein Entink, 2011) for the statistical software R (R core development team, 2011). This approach is based on the principal idea of simulating the multivariate posterior distribution of all model parameters. The distribution of values drawn from the Gibbs sampler converges to the posterior distribution; therefore, convergence has to be tested. Values, which were drawn before convergence was reached, are denoted as burn-in phase and usually not used for further analysis. Based on the drawn values, the mean of the posterior distribution (EAP) and the highest posterior density (HPD) intervals can be calculated. HPD intervals are the smallest intervals that contain a given percentage (e.g., 95%) of the values of the posterior distribution and can be used to test the statistical significance of model parameters.

2.3. Model testing

The fit of the response models was assessed in two steps and based on methods described by Sinharay and colleagues (Sinharay, 2005; Sinharay, Johnson & Stern, 2006). First, the assumption of local independence of the items was tested using the odds ratio statistic. The principal idea underlying this model test is to compare the frequencies of identical responses (0-0 or 1-1) with those of differing responses (0-1 or 1-0) by calculating these frequencies' ratio for each item pair. If an item response model assuming local independence fits the data well, it should accurately describe the observed ratio of these frequencies. If the model's prediction differs significantly from the observed data, this result indicates the presence of dependencies between certain items.

In a second step, the frequency of observed score distribution is compared with the score distribution predicted by the model as an overall measure of model fit. The principal idea underlying this model test goes back to Hambleton and Han (2004) and Ferrando and Lorenzo-seva (2001). A severe discrepancy between the observed and predicted score distribution indicates a misfit of the model to the data.

The fit of the item response time model was measured using a graphical model test, which compared the distribution of the response times predicted by the model with the observed response time distributions for each item. The values predicted by the model are plotted against their observed values. For each item, perfect model fit would be indicated by a linear plot.

2.4. Investigation of gender differences in mental rotation ability and speed

The original model for the responses and response times was further expanded to contain gender as a distinct person covariate G_i (which took the value 0 for the male population and 1 for the female population) and used to measure the influence of gender on speed and ability by a linear regression model:

$$\theta_i = \gamma_{00} + G_i \times \gamma_{01} + e_{oi} \tag{4}$$

$$\zeta_i = \gamma_{10} + G_i \times \gamma_{11} + e_{1i} \tag{5}$$

In this model, *e* is a residual term, which is assumed to be normally distributed. The gender effect on speed or ability is considered to be insignificant if the respective HPD intervals for γ_{01} or γ_{11} do not contain 0.

3. Study I

Participants. A sample of 208 respondents completed a computer-based test battery including a computerized cube comparison task. After excluding 9 respondents who did not show instruction-conforming test behavior, the final sample consisted of 108 female subjects and 91 male subjects aged 17 to 63 years (mean = 37.29, SD = 11.71). Nineteen of the respondents have completed 9 years of schooling but without vocational training, and 72 of the respondents have completed vocational training. Eighty-nine of the respondents graduated from high school and are qualified for an entrance in a university, and 19 of the respondents had an academic degree.

Measure. Mental rotation was measured by means of a cube comparison task. The mental rotation test consisted of k = 17 items. The task of the respondent was to compare a reference cube with a set of six comparison cubes. The respondents were asked to indicate the comparison cube, which merely differs from the reference cube in terms of its orientation. To rule out response elimination strategies, Gittler (1990) also included the response alternative "none of the comparison cubes are identical to the reference cube." Previous studies using this item set indicate the item set measures as unidimensional latent trait and that measurement invariance across age, gender, and educational level can be assumed because of the fit of the

1PL Rasch model and the invariance of the 1PL item difficulty parameters across these subpopulations (cf. Gittler, 1990; Tanzer et al., 1995). Furthermore, these studies also showed that item design features hypothesized to affect the processing demands of mental rotation tasks account for the estimated 1PL item difficulty parameters as indicated by the good fit of the linear logistic test model (LLTM; Fischer, 1995). Further evidence on the construct validity of this measure has been obtained in several exploratory and confirmatory factor analytic studies, which jointly indicate that items of this type load on the same factor as mental rotation tasks similar to the Vandenberg and Kuse (1978) test (cf. Arendasy, 2000; Arendasy, Hergovich & Sommer, 2008; Arendasy et al., 2011).

General Results. In a first step, two models were fitted, which used the one-parameter normal ogive model for describing responses and the one- or two-parameter log-normal model for describing response times. For fitting each model, 10.000 iterations using a Gibbs sampler were used, of which, 1.000 iterations were used as burn-in phase. The convergence of the iterations was tested using the convergence diagnostics of Geweke (1992) and Heidelberger and Welch (1983).

We compared the DIC values of the two models using the one- and two-parameter log-normal model, respectively, for describing the response times and the one-parameter normal-ogive model for describing the responses. We found a lower DIC value for the model containing the two-parameter log-normal model and, therefore, used this model for our further analysis.

The fit of the one-parameter normal-ogive model to the observed responses was tested with the two approaches described in section 2.3. In two item pairs, we observed significant posterior p-values of the odds ratio statistic, which lie below 0.025 or above 0.975. Overall, the assumption of local independence did not seem to be violated. By comparing the observed and predicted score distributions, only for two sum scores, significant deviations could be observed. In summary, we found that the one-parameter normal ogive model fits the data well.

The fit of the response time model was measured using the procedure described in section 2.3. The results of this analysis are presented in Figure 1 for 12 selected items. For each item, perfect model fit would be indicated by a straight line going from the lower left corner of the plot to the upper right corner, which would indicate a perfect correlation between observed and predicted response times. As can be seen, all items fitted the item response model well.

[Insert Figure 1]

The variances and covariances for the item and person parameters are displayed in Table 1. There is a remarkable negative correlation between speed and ability, indicating that more proficient respondents tend to work slower on the test items. The values of the ability and speed parameters of all respondents are displayed in Figure 2. Only small correlations are observed between the difficulty, the time discrimination, and the time intensity parameters.

[Insert Table 1]

[Insert Figure 2]

Gender Differences in the Cube Comparison Task. We measured the effects of gender on speed and ability using the procedure described in section 2.4. The estimations of the standardized effects are displayed as EAP values in Table 2. We further tested the significance of the observed gender differences by calculating a 95% HPD interval for the measured coefficient. The results of this analysis suggest that there is a significant gender difference in the ability level of the mental rotation task but not in the working speed.

[Insert Table 2]

Discussion. Our analysis of the data of the cube comparison task provides evidence that both response and response time can be described well in our modeling framework, which is based on the work of van der Linden (2006) and Klein Entink, Fox et al. (2009). It follows from our findings that in the used mental rotation task, both speed and ability can be considered as unidimensional constructs. This finding is an important prerequisite to compare the speed and ability of male and female subjects in the cube comparison task. We found a strong negative correlation between speed and ability in our sample. This finding is in line with similar findings reported in the literature for other ability tasks (e.g., Goldhammer & Klein Entink, 2011).

As expected, we found significant gender differences in the performance in the cube comparison task but no gender differences in the mental rotation processing speed, suggesting that female subjects do not use more time-consuming strategies if no time limit is given for working on the task.

4. Study II

Participants. A sample of 245 respondents completed k = 13 items, which were based on the endless loop paradigm, which will be described below. Seven participants were excluded from our analysis because they showed a bad test performance combined with short response times. In the final sample, there were 125 (51.43%) female subjects and 119 (48.57%) male subjects aged 18 to 64 years (mean 29.5 years, standard deviation 10.8). Seven of the respondents have completed 9 years of schooling but no vocational training, and 30 of the respondents have completed vocational training. One hundred fifty-two of the respondents graduated from high school and are thus qualified for entrance in a university, and 56 of the respondents had an academic degree.

Measure. This study investigates the response time and responses in an item set based on the endless loop paradigm, which has been already described in a line of studies (Arendasy, 2000, 2005; Arendasy & Sommer, 2010; Arendasy, Sommer & Gittler, 2010; Gittler & Arendasy, 2003). Its basic design is comparable to those of classical mental rotation tasks such as those described by Shepard and Metzler (1971). In this task, a closed and convoluted tube is presented in two different positions to the respondent. For each tube, the respondent has to decide under which viewing angle the second position equals the first. Arendasy (2005), Arendasy and Sommer (2010), and Gittler and Arendasy (2003) described a row of studies, which investigated the psychometric properties of this task and led to the definition of a set of rules for the definition of new items. The results of these authors suggest that item sets, which were constructed according to this rationale showed a good fit to the 1PL Rasch model and invariance of the 1PL item difficulty parameters across age, gender, and educational level. Earlier work by Arendasy (2000) further suggests that the ability measured by this task is highly related to mental rotation ability as assessed by similar tasks. Arendasy and Sommer (2010) further provided evidence that the 1PL difficulty parameters in items of this type are determined by item design features, which have been hypothesized to affect the cognitive processes of mental rotation tasks.

General Results. As in the first study, we first fitted two different models to the data. Both models used the one-parameter normal ogive model for describing the responses. The first model described the response times with the one-parameter log-normal model, whereas the second model used the two-parameter log-normal model. Again, we did not investigate the fit of any model, which used the two-parameter normal ogive model for describing the responses because we expected the one-parameter normal ogive model to fit the data well. For fitting each model, 10.000 iterations of a Gibbs sampler were used, of which, 5.000 iterations were used as burn-in phase. The convergence of the iterations was tested again using the convergence diagnostics of Geweke (1992) and Heidelberger and Welch (1983). We found that the model, which used the two-parameter log-normal model for describing the response times, showed a lower DIC value and selected this model for our further analysis.

We found the item response model to fit the data well. In all item pairs, we observed insignificant posterior p-values of the odds ratio statistic between 0.025 and 0.975. Our results indicate that the assumption of local independence was not violated. In a second step, we compared the frequency of observed sum scores with the frequency expected under the model. For no sum scores, significant deviations could be observed.

The fit of the two-parameter log-normal model to the data was assessed following the posterior predictive assessment described in section 1.6. The results of this analysis are presented in Figure 3 for all 15 items. In summary, the model for the response times fits the data well.

[Insert Figure 3]

The variances and covariances for the item and person parameters are displayed in Table 3. There is a remarkable negative correlation between speed and ability, indicating that more proficient respondents tend to work slower on the test items. The values of the ability and speed parameters of all respondents are displayed in Figure 4. Only small correlations are observed between the difficulty, the time discrimination, and the time intensity parameters.

[Insert Table 3]

[Insert Figure 4]

Gender Differences in the Endless Loop Task. The estimations of the standardized effects are displayed as EAP values in Table 4. We further tested the significance of the observed gender differences by calculating a 95% HPD interval for the measured coefficient. The results of this analysis suggest that there is a significant gender difference in the ability level of the mental rotation task but not in the working speed.

[Insert Table 4]

Discussion. Our analysis of the data of the endless loop task suggests that our chosen modeling framework (Klein Entink, Fox et al., 2009), which models the responses with the one-parameter normal ogive model and the response times with the two-parameter log-normal model, describes the observed data well. Again, it follows from this finding that speed and ability can be regarded as unidimensional constructs, which is a prerequisite for interpreting any observed gender differences in them.

As was the case in the cube comparison task, we found significant gender differences in the performance in the endless loop task but no gender differences in the mental rotation processing speed. These results suggest that female subjects do not use more time-consuming solution strategies in solving the endless loop tasks and, thus, contradict the predictions we made based on the model of Goldstein et al. (1990).

5. General Discussion

Our study investigated the relation between mental rotation processing speed and mental rotation ability as well as the presence of gender differences in both traits. Our principal research question concerned the presence of gender differences in mental rotation processing speed. We investigated this question in two studies, which used two different item types for assessing mental rotation ability, the first being a cube comparison task based on the rationale of Gittler (1990; cf. Arendasy et al., 2011) and the second being an endless loop task (Arendasy, 2005; Arendasy & Sommer, 2010; Gittler & Arendasy, 2003). Both tasks measure similar but not identical facets of mental rotation (cf. Arendasy, Hergovich & Sommer, 2008). Voyer (2011) found in a recent meta-analysis that the well-known gender differences in mental rotation ability are diminished in paper-pencil tests when no time limit is imposed. He explained this result by possible ceiling effects in male subjects. The earlier model of Goldstein et al. (1990) explained the gender differences in mental rotation ability by differences in mental rotation processing speed but without directly modeling mental rotation processing speed.

By applying a modeling approach described by Klein Entink, Fox et al. (2009), we showed that mental rotation processing speed can be regarded as a unidimensional measure in both investigated mental rotation tests. As expected, we found significant gender differences in mental rotation ability in both mental rotation tasks but no significant gender differences in mental rotation processing speed. Our results thus indicate that the observed gender differences in mental rotation ability cannot be explained by gender differences in mental rotation processing speed, as was predicted in the model of Goldstein et al. (1990).

A second major finding of our study was the observed negative correlation between mental rotation processing speed and mental rotation ability. Similar results have been already reported for figural reasoning tasks (Klein Entink, Kuhn, Hornke, & Fox, 2009; Goldhammer & Klein Entink, 2011) as well as quantitative and scientific reasoning tasks (Klein Entink, Fox et al., 2009). The present study extends these findings to mental rotation tasks, which were presented without time limits. Our findings cannot necessarily be generalized to visual ability tests, which are presented with time constraints. Future studies will have to investigate if the modeling approach used in our study can be applied to tests of this type. A discussion of speed-accuracy tradeoff in spatial ability tests with limited stimulus presentation times was provided by Lohman (1986).

For the observed negative correlations, multiple explanations have been offered in previous studies. Klein Entink, Kuhn et al. (2009) pointed out that test takers who care more about their results take more time to complete a test. Goldhammer and Klein Entink (2011) explained the negative correlations of reasoning tasks by the necessity to monitor and validate decisions while working on the task. A similar explanation may apply to our findings because several writers suggested that a conformational stadium is part of the cognitive processes involved in solving mental rotation tasks (e.g., Arendasy & Sommer, 2010; Just & Carpenter, 1985). A respondent who chooses to work fast but not very accurately may achieve a higher speed parameter but a lower ability parameter than a respondent with comparable cognitive abilities who decides to work accurately. A recent description of this speed-accuracy trade-off (Luce, 1986) in the context of educational measurement has been provided by van der Linden (2009).

References

Arendasy, M. (2000). *Psychometrischer Vergleich computergestützer Vorgabeformen bei Raumvorstellungsaufgaben: Stereoskopisch dreidimensionale und herkömmlich-zweidimensionale Darbietung* [Psychometric comparison of computer based presentation modes of spatial ability tasks: Stereoscopic-3D and traditional-2D presentation modes]. PhD Dissertation: University of Vienna.

Arendasy, M., Hergovich, A., & Sommer, M. (2008). Investigating the 'g' saturation of various stratum-two factors using automatic item generation. *Intelligence*, *36*, 574–583.

Arendasy, M., Hornke, L. F., Sommer, M., Häusler, J.,Wagner-Menghin, M.,Gittler, G., Bognar, B., & Wenzl, M. (2007). *Manual Intelligence-Structure-Battery (INSBAT)*. Mödling: SCHUHFRIED GmbH.

Arendasy, M., & Sommer, M. (2010). Evaluating the contribution of different item design features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence*, *38*, 574-581.

Arendasy, M., Sommer, M., & Gittler, G. (2010). Combining automatic item generation and experimental designs to investigate the contribution of cognitive components to the gender difference in mental rotation. *Intelligence*, *38*, 506-512.

Arendasy, M., Sommer, M., Hergovich, A., & Feldhammer, M. (2011). Evaluating the impact of depth cue salience in working three-dimensional mental rotation tasks by means of psychometric experiments. *Learning and Individual Differences*, *21*, 403-408.

Carroll, J. B. (1993). *Human Cognitive Abilities*. New York: Cambridge University Press.

Embretson, S. E. & Reise, S. (2000). *Item Response Theory for Psychologists*. Erlbaum: Erlbaum Publishers.

Ferrando, P. J., & Lorenzo-seva, U. (2001). Checking the appropriateness of item response theory models by predicting the distribution of observed scores: The program EP-fit. *Educational and Psychological Measurement*, *61*, 895-902.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models. Foundations, Recent Developments, and Applications* (pp. 157−180). New York: Springer.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). Washington, DC: Chapman & Hall/CRC.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J.M. Bernardo, J. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 4: proceedings of the fourth Valencia international meeting* (pp. 169–193). Oxford: Oxford University Press.

Gittler, G. (1990). *Dreidimensionaler Würfeltest—Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. Theoretische Grundlagen und Manual* [Three dimensional cubes test—A Rasch-calibrated test for the measurement of spatial ability. Theoretical background and Manual] Weinheim: Beltz.

Gittler, G., & Arendasy, M. (2003). Endlosschleifen: Psychometrische Grundlagen des Aufgabentyps EC [Endless loops: Psychometric foundations of EC]. *Diagnostica*, *49*, 164–175.

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, 39, 108-119.

Goldstein, D., Haldane, D., & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition*, *18*, 546–550.

Halpern, D. F. (2000). *Sex differences in cognitive abilities*. Mahwah: Erlbaum.

Hambleton, R. K., & Han, N. (2004, April). *Assessing the fit of IRT models: Some approaches and graphical displays.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Heidelberger, P., & Welch, P.D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109–1144.

Johnson, W., & Bouchard, T. J., Jr. (2005). The structure of human intelligence: It's verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393-416.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92*, 137-172.

Klein Entink, R. H. (2011). *cirt: Conjoint IRT modeling of Responses and Response Times.* cirt version 2.5.32. Retrieved Dezember 9, 2013, from http://www.kleinentink.eu/download/cirt_2.5.32.zip

Klein Entink, R. H., Kuhn, J. -T., Hornke, L. F., & Fox, J. -P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*, 54−75.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to simultaneous modeling of accuracy and speed on test items. *Psychometrika*, *74*, 21-48.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterisation of gender differences in spatial abilities: A meta-analysis. *Child Development*, *56*, 1479–1498.

Lohman, D. F. (1986). The effect of speed-accuracy tradeoff on sex differences in mental rotation. *Perception and Psychophysics*, *39*, 427–436.

Masters, M. (1998). The gender difference on the Mental Rotations Test is not due to performance factors. *Memory & Cognition*, *26*, 444–448.

Masters, M.S. & Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behavior Genetics*, *23*, 337-341.

McGrew, K. S. (2005). The Cattell–Horn–Carroll (CHC) theory of cognitive abilities: Past, present and future. In D. Flanagan, & Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 136−202). New York: Guilford Press.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Silverman, I., Choi, J., & Peters, M. (2007). The hunter-gatherer theory of sex differences in spatial abilities: Data from 40 countries. *Archives of Sexual Behavior*, *36*, 261–268.

Sinharay, S. (2005). Assessing fit of unidimensional iterm response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375-394.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, *64*, 583−639.

Tanzer, N. K., Gittler, G. & Ellis, B.B. (1995). Cross-cultural validation of item complexity in a LLTM-calibrated spatial ability test. *European Journal of Psychological Assessment*, *11*, 170-183.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247−272.

van der Linden, W. J., & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotation, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*, 599–604.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270.

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin and Review*, *18*, 267-277.

Table 1. EAP values for the components of the person and item parameter covariance matrices in the cube comparison task

Component		EAP	Correlation
Person Parameters	σ_{θ}^{2}	1.00	-
	$\sigma_{\theta\zeta}$	-0.32	-0.65
	σ_{ζ}^{2}	0.24	-
Item Parameters	σ_{b}^{2}	0.84	-
	$\sigma_{b\phi}$	0.13	0.17
	$\sigma_{b\lambda}$	0.08	0.11
	σ_{ϕ}^{2}	0.67	-
	$\sigma_{\lambda\phi}$	0.07	0.11
	σ_{ϕ}^{2}	0.64	-

Table 2. Estimated standardized effects of gender on mental rotation ability and speed in the cube comparison task

	EAP	95% HPD
Effect on Ability	-0.30	[-0.59; -0.01]
Effect on Speed	0.07	[-0.08; 0.23]

Table 3. EAP values for the components of the person and item parameter covariance matrices in the endless loop task

Component		EAP	Correlation
Person Parameters	σ_{θ}^{2}	1.00	-
	$\sigma_{\theta\zeta}$	-0.05	-0.35
	σ_{ζ}^{2}	0.02	-
Item Parameters	σ_{b}^{2}	1.32	-
	$\sigma_{b\phi}$	0.28	0.13
	$\sigma_{b\lambda}$	0.15	0.13
	σ_{ϕ}^{2}	3.33	-
	$\sigma_{\lambda\phi}$	0.78	0.42
	σ_{ϕ}^{2}	1.05	-

Table 4. Estimated standardized effects of gender on mental rotation ability and speed in the endless loop task.

	EAP	95% HPD
Effect on Ability	-0.40	[-0.68; -0.13]
Effect on Speed	0.03	[-0.01; 0.08]

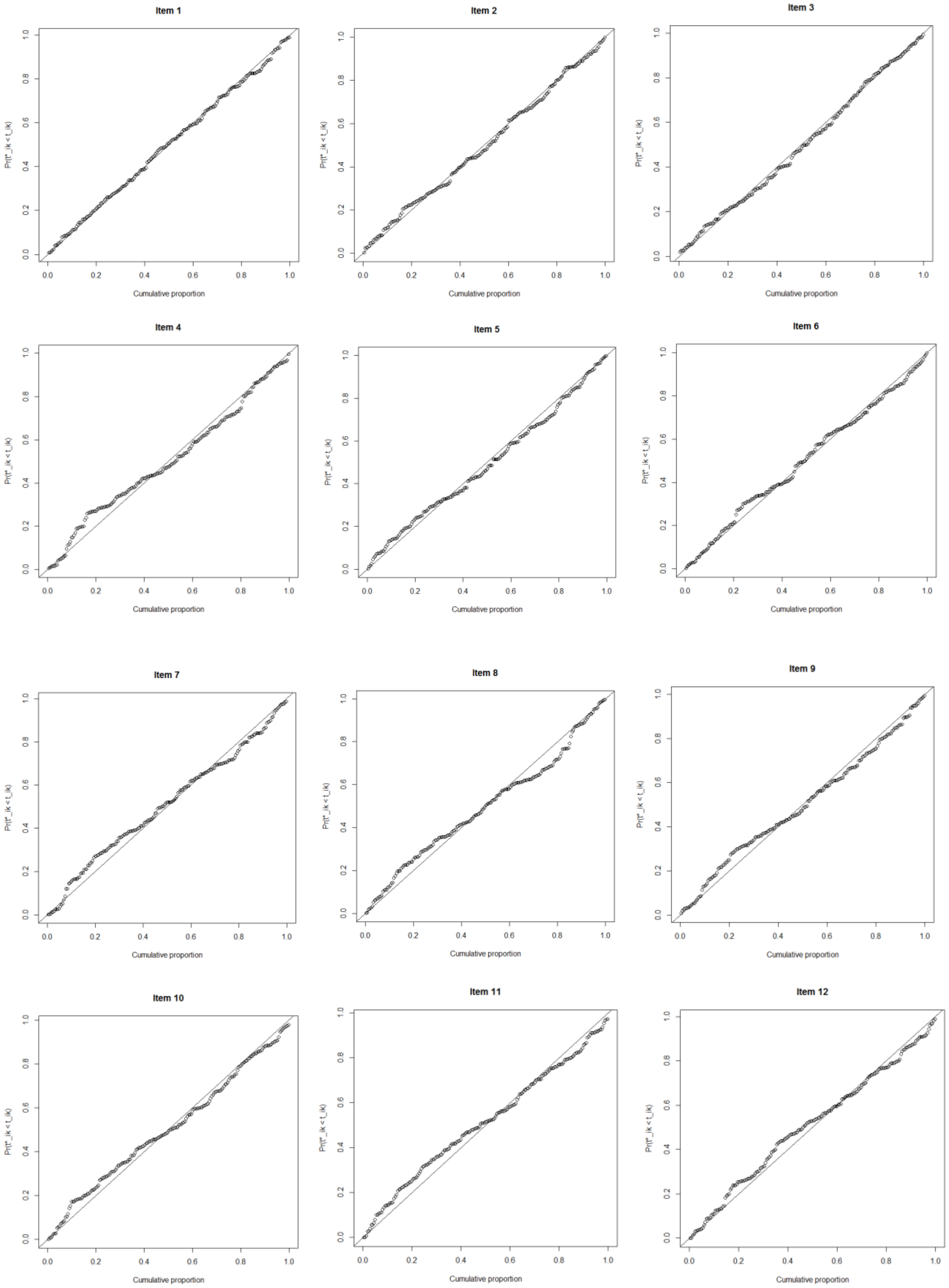

Figure 1. Cumulative probability plots for 12 selected items of the cube comparison task

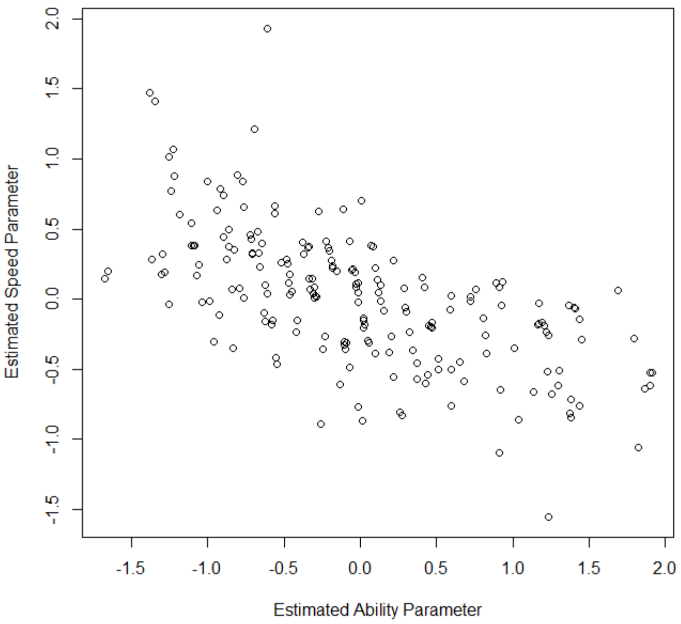

Figure 2. Estimated speed and ability parameters for the analyzed sample working on the cube comparison task.

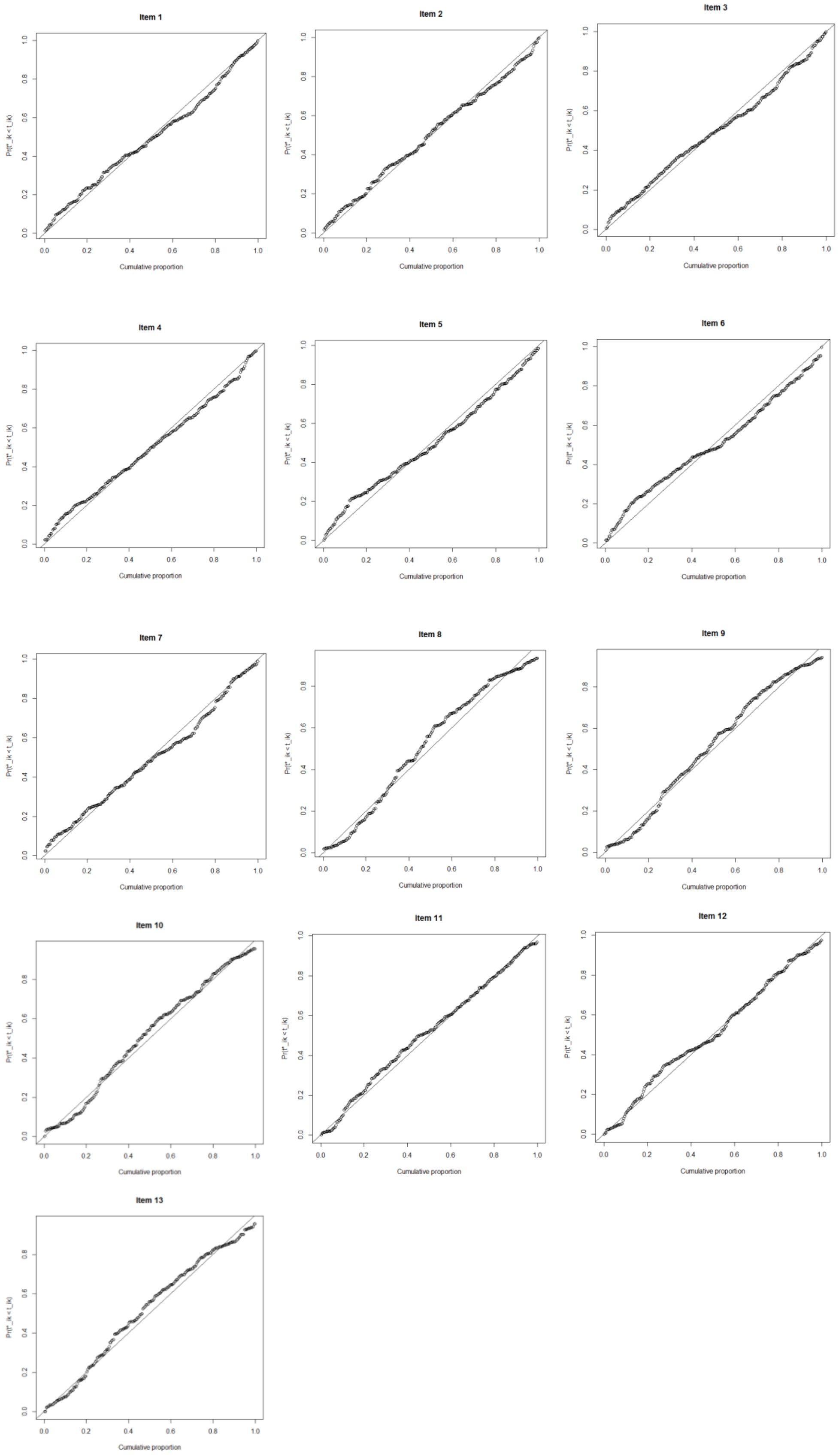

Figure 3. Cumulative probability plots for 13 items of the endless loop task

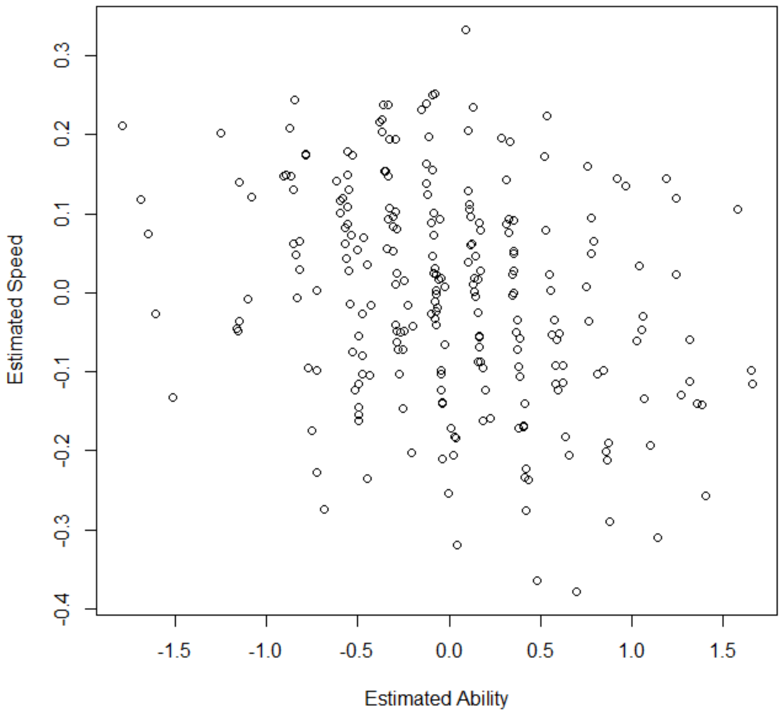

Figure 4. Estimated speed and ability parameters for the analyzed sample working on the endless loop task.

Abstract

This doctoral thesis describes and evaluates different methods of dimensionality assessment in psychological test evaluation and development. It consists of three studies, of which the first two investigate exploratory procedures to determine the dimensionality of an item set. The first study examined an approach for clustering items in Rasch measurement. The purpose of the examined algorithm is to find item clusters which show a good fit to the Rasch model while excluding items which show model violations. This approach was evaluated by the means of a simulation study, which compared the results of this algorithm with the results obtained from the application of a principal component analysis of tetrachoric correlations. It was found that the examined algorithm leads to practically usable results, especially for the analysis of data from large person samples.

The second study investigated the principal component analysis of smoothed tetrachoric correlation matrices as a measure of dimensionality, again by the means of a simulation study. By comparing the results of several smoothing algorithms, it was found that the application of smoothing algorithms to the principal component analysis of tetrachoric correlations led to improved results in the assessment of dimensionality under multiple conditions.

The third study exemplified a confirmatory approach for modeling responses and response times by applying it to two types of mental rotation tasks. This study investigated the hypothesis that the well-known gender differences in mental rotation ability are caused by gender differences in mental rotation speed. After modeling both speed and ability, it was found for both task types that gender differences could only be observed for mental rotation ability, but not for mental rotation processing speed. Furthermore, a negative correlation between speed and ability could be observed in both mental rotation tasks, which was in line with results obtained for other ability tests.

Zusammenfassung

Die vorliegende Arbeit beschreibt und beurteilt verschiedene Methoden zur Erfassung der Anzahl zugrundeliegender Dimensionen in der Bewertung und Entwicklung psychologischer Testverfahren. Sie umfasst drei Studien, von denen die ersten beiden exploratorische Verfahren zur Beurteilung der Dimensionalität einer Aufgabengruppe untersuchen. Die erste Studie untersuchte einen Algorithmus zur Gruppierung von Aufgaben im Kontext des Rasch-Modells. Das Ziel des untersuchten Algorithmus besteht im Finden von Aufgabengruppen, die eine gute Passung auf das Rasch-Modell zeigen, und dem Ausschließen von Aufgaben, welche Modellverletzungen zeigen. Dieser Ansatz wurde mit Hilfe einer Simulationsstudie untersucht, welche die Resultate dieses Algorithmus mit denen einer Hauptkomponentenanalyse tetrachorischer Korrelationen verglich. Es zeigte sich, dass der untersuchte Algorithmus zu praktisch verwertbaren Ergebnissen führt, insbesondere wenn Daten von großen Personenstichproben untersucht werden.

Die zweite Studie untersuchte die Hauptkomponentenanalyse geglätteter tetrachorischer Korrelationsmatrizen als Maß für die Dimensionalität, wieder durch Anwendung einer Simulationsstudie. Indem die Ergebnisse verschiedener Glättungsalgorithmen verglichen wurden, zeigte sich, dass die Anwendung von Glättungsalgorithmen unter zahlreichen Bedingungen zu verbesserten Ergebnissen in der Messung der Dimensionalität führte.

Die dritte Studie veranschaulichte einen konfirmatorischen Ansatz zur Modellierung von Antworten und Antwortzeiten durch dessen Anwendung auf zwei Aufgabentypen zur Erfassung mentaler Rotation. Diese Studie prüfte die Hypothese, dass die allgemein bekannten Geschlechtsunterschiede in der Fähigkeit zur mentalen Rotation durch Geschlechtsunterschiede in der Verarbeitungsgeschwindigkeit mentaler Rotationsaufgaben verursacht werden. Nachdem Fähigkeit und Verarbeitungsgeschwindigkeit modelliert wurden, zeigte sich, dass bei beiden Aufgabentypen nur Unterschiede in der Fähigkeit zur mentalen Rotation gefunden werden konnten, nicht jedoch in der entsprechenden

Verarbeitungsgeschwindigkeit. Zudem wurde eine negative Korrelation zwischen Verarbeitungsgeschwindigkeit und Fähigkeit gefunden, was mit den Ergebnissen, welche für andere Fähigkeitstests gefunden wurde, übereinstimmt.

Curriculum Vitae

Name:	Rudolf Debelak
Born:	06/02/1982 in Vienna
Nationality:	Austria
Languages:	German, English, French

Education:

1988-1992	Visit of the elementary school Piaristenvolksschule
1992-2000	Visit of the secondary school Bundesgymnasium 8
10/2000 - 09/ 2002	Diploma Study, Mathematics, University of Vienna
10/2000 – 07/2007	Diploma Study, Psychology, University of Vienna (Mag. rer. nat.)
Since 10/2007	Doctoral Study, Psychology, University of Vienna

Professional experience:

Since 05/2008	Psychologist in the research and development department, SCHUHFRIED GmbH

Peer-reviewed articles

Debelak, R., & Arendasy, M. (2012). An algorithm for clustering items and testing unidimensionality in Rasch measurement. *Educational and Psychological Measurement*, *72*, 375-387.

Debelak, R., & Tran, U.S. (2013). Principal component analysis of smoothed correlation matrices as a measure of dimensionality. *Educational and Psychological Measurement*, *73*, 63-77.

Debelak, R., Gittler, G., & Arendasy, M. (submitted). On gender differences in mental rotation processing speed. *Learning and Individual Differences*.

Rodewald, K., Bartolovic, M., Debelak, R., Aschenbrenner, S., Weisbrod, M., & Roesch-Ely, D. (2012). Eine Normierungsstudie eines modifizierten Trail Making Tests im deutschsprachigen Raum. *Zeitschrift für Neuropsychologie*, *23*, 37-48.

Scientific Presentations

Debelak, R. (2010). Chancen und Risiken von Testungen unter Zeitbeschränkung am Beispiel eines Raumvorstellungstests. Talk presented at the 48th conference meeting of the German psychological society, Sept 26-30 2010, Bremen, Germany.

Debelak, R. (2011). Modeling of Speed and Accuracy in Computer-Based Testing. Talk presented at the 12th European Congress of Psychology, July 4-8 2011, Istanbul, Turkey.

Debelak, R. (2011). Evaluation eines Algorithmus zur modellgeleiteten Itemgruppierung. Talk presented at the 10th meeting of the section on methods and evaluation of the German psychological society, Sept 21-23 2011, Bamberg, Germany.

Debelak, R. (2011). Zur Beziehung zwischen Bearbeitungsgeschwindigkeit und Leistungsvermögen bei Leistungstests. Talk presented at the 11th meeting of the section on differential psychology of the German psychological society, Sept 26-28 2011, Saarbrücken, Germany.

Debelak, R. (2012). Die Bearbeitungsgeschwindigkeit in Aufgaben zur Erfassung der Visualisierungsfähigkeit: Dimensionalität, Geschlechtsunterschiede und die Beziehung zum schlussfolgernden Denken. Talk presented at the 48th conference meeting of the German psychological society, Sept 23-29 2012, Bielefeld, Germany.

Debelak, R., & Egle, J. (2011). Inhibition – ein einheitliches Konstrukt? Überprüfung der Dimensionalität von fünf Aufgabenparadigmen. Talk presented at the 26th meeting of the German Neuropsychological Society, Sept 22-24 2011, Aachen, Germany.

Debelak, R., Egle, J., Sommer, M., & Kaller, C.P. (2012). Psychometric properties of a computerized version of the Tower of London: Using item response theory to evaluate its dimensionality and construct validity. Talk presented at the 2012 Meeting of the International Neuropsychological Society, Jun 27-30 2012, Oslo, Norway.

Vetter, M., & Debelak, R. (2012). Using non-linear regression to predict the driving fitness of post-injury patients. Talk presented at the 30. International Congress of Psychology, 22-27 July 2012, Cape Town, South Africa.