

Applied Dimensionality and Test Structure Assessment With the START-M Mathematics Test

Fabian Jasper

Johannes Gutenberg University Mainz, Germany

Abstract

This article presents a new approach to test both, the dimensionality and underlying structure of an achievement test. I focus on binary items as they are the most common format in the achievement test domain. A decision graph is presented that integrates modern methods of exploratory dimensionality assessment and confirmatory methods to confirm the underlying structure of the test. Conclusions are drawn concerning the gain in knowledge and the efforts needed to apply the methods from the point of view of a test developer. A further purpose of this article is to introduce the START-M which is a new German mathematics test for respondents aged 16+ years and serves to demonstrate my ideas with a dataset that is large enough ($N = 1,554$).

Keywords: dimensionality, START-M, assess, mathematics test, detect, dimtest, noharm, binary items

The dimensionality of an achievement test is a nontrivial aspect of test development and reanalysis. There are many ways to define dimensionality but I find the one by McDonald (1986) very useful, who defines the dimensionality of a test as the number of factors that are necessary to account for the relationships (correlations) between the items. The concept of simple structure stems from the early days of factor analysis and was introduced by Thurstone (1947), who suggested that the factors should be rotated such that each row shows at least one loading close to zero, each factor should have at least as many variables with near-zero loadings as there are factors, and if pairs of factors are regarded, variables should have loadings on only one factor. If these conditions are met, it is often easy to interpret the factor loading matrix because each variable tends to load on only one factor (e.g. a mathematics factor or a reading factor). But even when only some of the variables show salient loadings on multiple factors at the same time (i.e. a complex structure exists), the interpretation of the results and the detection of the correct number of dimensions of the test becomes quite difficult (Beauducel, 1997; Lorenzo-Seva, 2003, Zhang & Stout, 1999). I will present a framework that can help the practitioner to face dimensionality assessment in case of both simple and complex structures.

Even with correlated dimensions the tilt in profiles of respondents allows for astonishing prediction of development paths as Lubinski, Webb, Morelock and Benbow (2004) showed in a 10-year follow-up study. More than 60 years ago, it was McNemar (1946) who pointed out that only in case of unidimensionality, may people with the same score be regarded as quantitatively and, within limits, qualitatively similar. But how can one determine the dimensionality of a test?

Hattie (1984, 1985) offered a comprehensive overview concerning methods available to check the dimensionality of a test and came to the conclusion that most indices were inappropriate for dimensionality assessment. Some methods that failed to reliably assess the dimensionality of a test were

based on Cronbach's α , which strongly depends on the number of items in a test and is rather a measure of internal consistency (Green, Lissitz, & Mulaik, 1977). Cortina (1993) summarized the problems with this measure and concluded that it can only be recommended if one is already convinced that items belong together and form an unidimensional scale. Another line of research in this area stems from the idea to use linear factor analysis as a tool to detect the dimensionality of a test. The question if one applies a principal component or principal axis analysis is still under debate but seems to be less important because both methods tend to produce quite similar results (e.g. Thompson & Brown, 2001; Velicer & Jackson, 1990). A much bigger problem arises from the fact that both methods are not appropriate in case of dichotomous items (Green, 1981). Kubinger (2003) proposes the use of tetrachoric correlations in this case, which is an improvement but cannot really solve the problem because the assumptions (Gorsuch, 1983) of this transformation will seldom be met. Another approach that is even more difficult to justify is Rasch factor analysis. The idea behind this method is to apply a Rasch model to a dataset and to run a linear factor analysis based on the remaining residuals (Bond & Fox, 2007). The gain in information compared to the results of an ordinary Rasch analysis alone is quite low.

Almost 20 years after Hattie's tremendous work, Tate (2003) made another attempt and found out that the picture has changed. The small number of remaining failures in dimensionality assessment now rather reflected the assumptions and goals of the particular methods. However, I am interested in the practical conclusions and guidelines for test-developers and pursue the goal to foster the use of modern dimensionality assessment methods whenever they are appropriate. Many of the studies on dimensionality assessment I review and comment in this article are based on simulation data and do not consider the practical usability of the methods (e. g., Hattie, 1984; Nandakumar & Ackerman, 2004). Till now, there exists no approach that incorporates an easy to use decision graph, structural equation modeling, item response theory based methods, and a theory-driven analysis in one framework. Besides a new framework for dimensionality and test structure assessment, an important outcome of this study is a new model of mathematical ability that is theory-based and empirically tested.

Methods for Assessing the Dimensionality and Test Structure

In this section I offer a brief overview concerning four methods that are suitable for the purpose of dimensionality assessment and confirmatory testing of the underlying test structure. For each method, I point out where I see its particular strength and finally I integrate all of them into a decision graph.

DIMTEST

The nonparametric DIMTEST procedure tests if essential unidimensionality ($H_0: d = 0$) holds for a given set of items. According to Stout (1987, p. 597), a test (U_1, \dots, U_N) of length N is said to be essentially unidimensional if there exists a latent variable θ such that for all values of θ ,

$$\frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} |Cov(U_i, U_j | \theta)| \approx 0 \quad (1)$$

This means that after conditioning on the latent trait the sum of the remaining conditional covariances between item pairs (U_i, U_j) should be very small. In order to test the hypothesis of essential unidimensionality the test is divided into two partitions, an assessment test (AT) and a partitioning test (PT). The AT is supposed to be dimensional homogeneous and can either be user defined or extracted from the test by means of exploratory factor analysis which is an option in the DIMTEST program (Stout, 1987). Now, the standardized difference between an ordinary variance estimate σ_k^2 and an unidimensional variance estimate $\sigma_{U,k}^2$, both of the AT scores, is the key aspect of an (asymptotic normal distributed) test statistic called T . This difference is calculated over $k = 1..K$ subgroups and these subgroups are based on the scores of the participants on the potentially heterogeneous PT. The formula for the unidimensional variance estimate for a single group k and an AT with M items is (Stout, 1987, p. 594)

$$\sigma_{U,k}^2 = \sum_{i=1}^M \frac{\hat{p}_i^{(k)}(1-\hat{p}_i^{(k)})}{M^2} \quad (2)$$

Here $\hat{p}_i^{(k)}$ is the proportion of respondents with the same score k on the PT who got item i of the AT correct (this distinction is not made for σ_k^2). Now, in case the whole test is essentially unidimensional, it should not matter if one calculates the variance estimate separately for the K groups according to the PT. This leads to the test statistic (Stout, 1987, p. 594)

$$T = \frac{1}{\sqrt{K}} \sum_{k=1}^K \left(\frac{\sigma_k^2 - \sigma_{U,k}^2}{S_K} \right) \quad (3)$$

that is sensitive to deviations from essential unidimensionality. S_K is necessary to normalize the variance differences and not explained further here (see Stout, 1987, p. 594 eq. 8) just like a correction for statistical bias that does not foster understanding of the principles behind the DIMTEST method.

Hattie, Krakowski, Roger, and Swaminathan (1996) evaluated the DIMTEST procedure and they concluded that the T statistic behaved reasonably robust and allows for a practical demarcation between one and multiple dimensions. In a study by Seraphine (2000), DIMTEST also proved to be a reliable method as long as no ceiling effects stemming from a mismatch between respondents ability and item difficulty were present. The procedure also provided good results for the detection of uni- and multidimensionality in a study by Tate (2003) that used monte carlo simulations as well as real datasets. The DIMTEST procedure developed in 1987 (Stout) has been improved in terms of bias correction and type of AT item selection in 1993 (Nandakumar & Stout) but the principles stated above remain valid.

The strength of the DIMTEST method lies in determining if the covariance structure of a test justifies to assume more than one dimension. This

is not trivial because other methods like linear factor analysis are hard to interpret in case of dichotomous items (e.g. Hattie, 1984, 1985). Besides, most of the common indexes of fit for the Rasch model do not directly aim at testing deviations from unidimensionality and their power is still under debate (Linacre, 1998; Verhelst, 2001). This means that there are good reasons to check if the test is indeed unidimensional before one applies the Rasch model.

Dimensionality Evaluation to Enumerate Contributing Traits (DETECT)

As pointed out before, DIMTEST can be used to answer the question if essential unidimensionality holds but does not give further information on the test structure. For this purpose, a procedure called DETECT was developed by Zhang and Stout (1999). DETECT offers a statistic $D(P)$ which was proved to be maximized at the correct partitioning of items into k partitions ($P=\{A_1, \dots, A_k\}$) under the condition of approximate simple structure. The formula for the DETECT index is (Gierl, Tan, & Wang, 2005, p. 6; Zhang & Stout, 1999, p. 218)

$$D(P) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq N} \delta_{ij} E[Cov(X_i, X_j | \Theta_{TT} = \theta)] \quad (4)$$

In this formula Θ_{TT} stands for a weighted test composite that is supposed to present the direction in which the test measures best. To clarify what is meant by direction take a look at the exemplary graphical representation of item vectors in Figure 1.

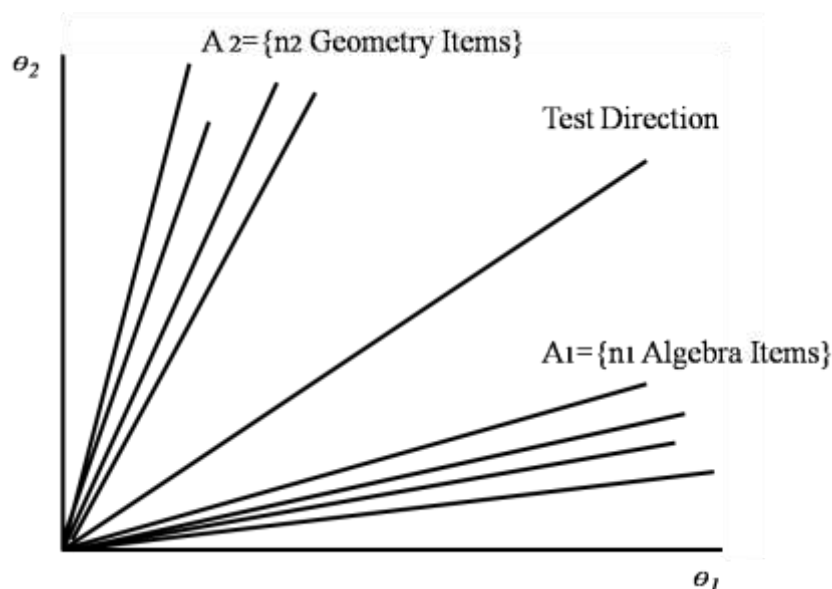


Figure 1. Item vector depiction according to Zhang and Stout (1999) in case of two underlying dimensions.

The direction that the whole test measures best is a result of vector products and Figure 1 is quite similar to a factor loading plot in case of a two factor solution. Now, according to Froehlich and Stout (2003) the conditional covariance between item pairs (X_i, X_j) that are supposed to measure the same

dimension are positive according to Equation 4 in case they are both on the same side of the line that represents the test direction in Figure 1. If two items are on different sides of the test direction vector the conditional covariance will be negative in sign (Froehlich & Stout, 2003). Moreover, it is defined that $\delta_{ij} = 1$ if the items X_i and X_j are in the same cluster of P and otherwise $\delta_{ij} = -1$. Besides the assignment of items to the k clusters, the DETECT index offers two indices. The above mentioned $D(P)$ and an index called r_{max} that indicates the amount of simple structure present in the data. According to Kim (1994), a $D(P) < .10$ indicates unidimensionality, values from .10 to .50 weak, from .51 to 1 moderate and above 1.0 strong multidimensionality. For the r_{max} index values above .80 indicate approximate simple structure and values below .80 complex structure (Zhang & Stout, 1999, p. 231).

It has to be made clear that algebra and geometry in the example of Figure 1 are just names given to the two item clusters. It is in no way guaranteed that all items on the right of the test direction vector are really algebra items. All that the DETECT index does is to maximize $D(P)$. That items which have positive conditional covariance and which are in the same item cluster according to DETECT look similar and require similar skills is what one wishes but this does not have to be the case. This could be compared to exploratory factor analysis or stepwise regression: DETECT will partition the test into k item clusters that maximize the DETECT index according to Equation 4 but the results do not have to be reasonable from the point of view of a test constructor. Even worse, the correctness of DETECT depends heavily on the amount of simple structure as Zhang and Stout (1999, p. 215) point out: "It is very important to note that DETECT is still informative when approximate simple structure fails to hold. In particular, it can still locate relatively dimensionally homogeneous clusters; however, there is no longer a unique "best" or "correct" partition to be found by DETECT because there will be little to no separation between some of the clusters found". Because simple structure is rare in the social sciences (Beauducel & Wittmann, 2005) the results of DETECT should be taken with care. Gierl, Leighton and Tan (2006) evaluated the accuracy of DETECT under non-optimal conditions and found out that DETECT works properly when up to 30% of the items are of complex structure, the intercorrelation of dimensions is less than .75 and the sample size is at least $N = 1000$. The method has already been applied in a few contexts, for example by Jang and Roussos (2007) to investigate the dimensionality of the TOEFL and by Gierl, Tan und Wang (2005) who analyzed a mathematics and reading SAT.

As pointed out, the results of DETECT can be difficult to interpret. Its main strength seems to be to identify the amount of simple structure in the data and to give the researcher a hint concerning the underlying test structure from a pure statistical point of view. It is not very realistic that DETECT will ever exactly confirm the expected test structure under the conditions in applied social sciences (hardly simple structure) but it can show the researcher if he is on the right track at all.

Nonlinear Factor Analysis

Nonlinear factor analysis can be carried out by the program NOHARM where the acronym stands for the *normal ogive harmonic analysis robust method* (Fraser & McDonald, 1988). It uses a nonlinear factor analytic

approach developed by McDonald (1962) and offers an exploratory and confirmatory mode. The NOHARM model is defined as (McDonald, 1999)

$$P\{X_i = 1 \mid \Theta_1 = \theta_1, \dots, \Theta_k = \theta_k\} = \phi\{\beta_{i0} + \beta_{i1}\theta_1 + \dots + \beta_{ik}\theta_k\} \quad (5)$$

Here the probability of a correct answer to item i given the latent abilities Θ_k is a function of the item difficulty parameters β_{i0} , the item discrimination parameters β_{ik} and the normal ogive presented by ϕ . A special case of this very general model is the one parameter (normal ogive) Rasch model (Bond & Fox, 2007). This model would be written as follows

$$P\{X_i = 1 \mid \Theta_1 = \theta_1\} = \phi\{\beta_{i0} + \beta_{i1}\theta_1\} \quad (6)$$

Besides the nonlinear approach that solves the well known problem of difficulty factors (McDonald & Ahlawat, 1974), one of the great advantages of NOHARM lies in the existence of factor loadings that can be interpreted easily. The precise

formulation of the item discrimination parameters is $\beta_{ik} = \frac{\lambda_{ik}}{\sqrt{\Psi_i}}$ and in this

expression Ψ_i stands for unexplained residual variance of item i and λ_{ik} can be interpreted as a factor loading (the correlation of item i with factor k). Among many other parameters the NOHARM (Fraser & McDonald, 1988) output also encompasses these factor loadings which can be interpreted using the same well known guidelines as in ordinary linear factor analysis (Gorsuch, 1983). NOHARM has been subject of various studies, for example concerning its behavior when sample size is small (Champlain & Gessaroli, 1996) and its power in dimensionality and test structure detection and confirmation (Finch & Habing, 2005; Hattie, 1984, 1985; Tate, 2003).

The NOHARM-program offers two indices of fit, the $GFI = 1 - (Tr(Cres^2) / Tr(C^2))$ (called Tanaka index of fit in NOHARM) where $Cres$ stands for the residual covariance matrix and C for the sample covariance matrix (McDonald, 1999, S. 83) and the root mean square residuals (RMSR). Concerning the GFI and similar indexes of fit, especially in the context of structural equation modeling, there have been various attempts to establish guidelines (Beauducel & Wittmann, 2005; Hu & Bentler, 1999). Taking a GFI $\geq .90$ as an indicator for acceptable and a GFI $\geq .95$ as an indicator for good fit still seems an acceptable rule of thumb. For the RMSR, Fraser and McDonald (1988) recommend to regard a $RMSR \leq 4 \cdot 1 / \sqrt{N}$, where N is the sample size, as an indicator for a good model fit. I do not use the program CHIDIM by Champlain and Tang (1997) that further analyses the residual correlation matrix after fitting a NOHARM model because the somewhat ambiguous results concerning the resulting (not chi-square distributed) statistic presented by Tate (2003, p. 186) do not justify the application of this complicated (it is a Fortran 77 program) method in applied research.

Thus, the strength of NOHARM lies in the possibility to test precise theories concerning the test structure. It's results are easy to interpret, it avoids difficulty factors, and is an appropriate procedure for a confirmatory analysis of the test structure on item level. But there are also severe limitations of this method. When the test structure is complex, with salient loadings of items on multiple factors or models that include higher order factors, one cannot specify the model in NOHARM. In its confirmatory mode the design matrix of the

program does not permit to estimate the factor loadings for items that load on multiple factors. One can put constraints on those factor loadings which is often hard to justify theoretically. On the other hand, the exploratory mode does not permit to test precise theories of the test structure besides the number of factors and the inter-factor correlation.

Structural Equation Modeling

Since its beginning in the 1970s structural equation modeling (SEM) has become a well known technique and probably due to modern computers and software that is easy to use it has become an important tool in every test constructor's toolbox (Kaplan, 2000). In this article I focus on dichotomous items as pointed out before. Models that are rather simple like an ordinary model with for example four correlated dimensions can easily be tested with the NOHARM method which avoids problems of nonlinearity with dichotomous items. Thus I see two main advantages that a SEM approach has over the NOHARM method. First, NOHARM offers a very limited number of fit indexes, second more complex models like for example a Schmid-Leiman Model (Schmid & Leiman, 1957) cannot be tested easily with NOHARM.

Because dichotomous items lead to problems due to non-normality when using the common maximum likelihood method in SEM Satorra and Bentler (1994) developed a correction for the χ^2 based fit indexes and the standard errors (robust method in EQS). An even more elegant solution to this problem is offered by Muthén (1993) and is called weighted least squares with mean and variance adjusted (WLSMV). This method encompasses the advantages of asymptotic distribution free estimation methods without the extreme sample sizes necessary for those approaches. The WLSMV method yielded very accurate results for sample sizes as small as $N = 200$ (Muthén, du Toit, & Spisic, 1997). This has been confirmed by Beauducel and Herzberg (2006) who compared maximum likelihood and WLSMV estimations for confirmatory factor analyses with moderate item-factor loadings and differing numbers of answering categories (including dichotomous items). Unfortunately, it is only available in the MPLUS software by Muthén and Muthén (2007).

Another aspect concerning the use of SEM and WLSMV arises from the fact that if the polychoric correlations used by MPLUS for calculation purposes are very high the program states that the information from these variables can be used to create a new one. This is also what Kline (2005) suggests because a great amount of multicollinearity can cause severe problems in the estimation process. There are many different ways to combine items into mini-scales and a heated debate concerning the appropriateness a parceling is still going on (e.g. Little, Cunningham, Shahar, & Widaman, 2002). I only use difficulty based parceling here which is the most popular method in the domain of achievement tests as Bandalos and Finney (2001) point out. This means that within one assumed dimension the first parcel consists of the sum of the easiest item (lowest p -value) and the most difficult item (highest p -value). In this way, one cancels out random and systematic error by aggregating across these errors. This happens because the variance of the sum of two variables A and B , which I call AB is defined as follows (e.g. Hays, 1994)

$$\sigma^2_{AB} = \sigma^2_A + \sigma^2_B + 2\text{cov}_{A,B} \quad (7)$$

I suggest to use the WLSMV method and to decide whether to parcel or not based on the amount of multicollinearity in the data. Whenever problems such as warning messages or failures to converge occur due to multicollinearity I apply difficulty based parceling within the theoretical assumed dimensions of a test.

Integrated Point of View

I have presented some of the most advanced techniques to examine the dimensionality and underlying structure of psychological achievement tests in the previous sections. Based on their strengths and weaknesses I suggest the decision chart according to Figure 2 to determine the dimensionality and test structure of an achievement test.

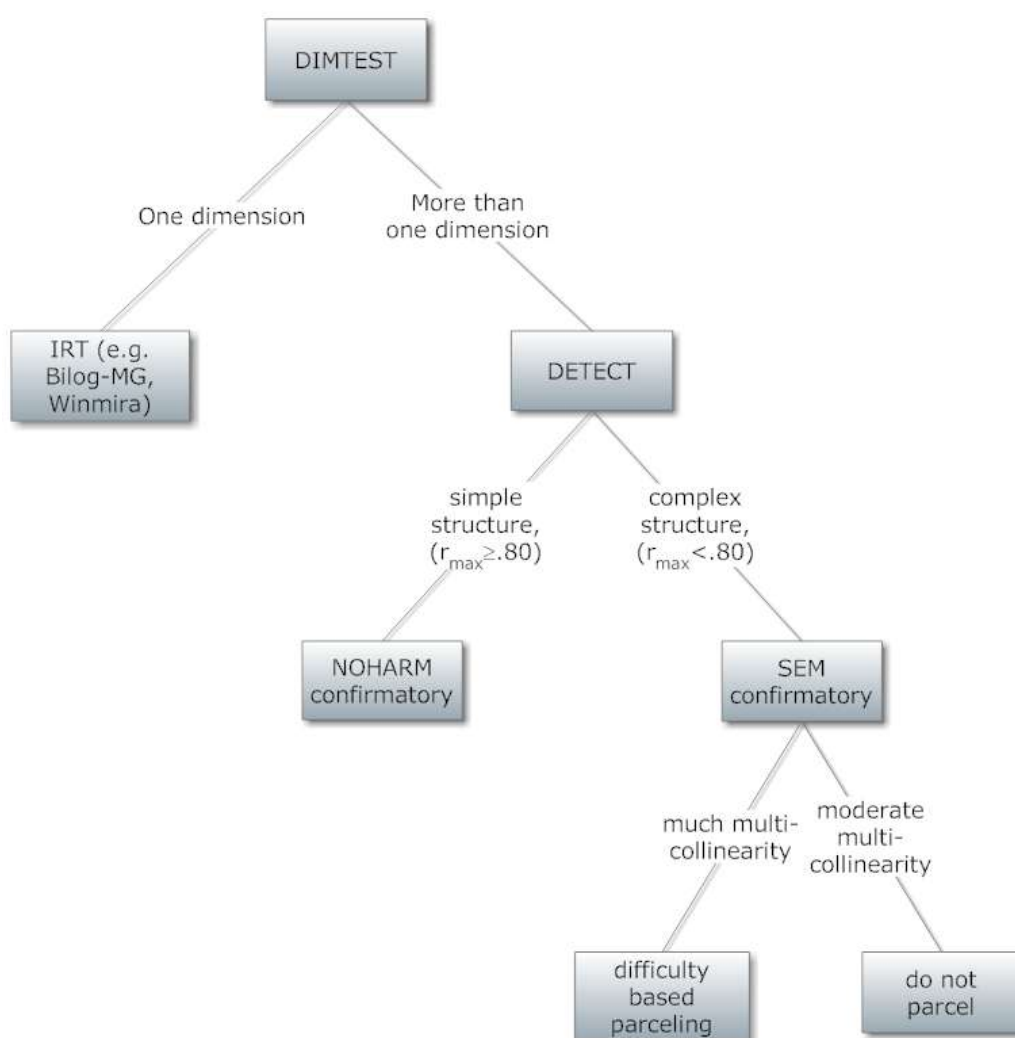


Figure 2. Decision graph for Dimensionality and Test Structure Determination.

In the very first step of the analysis one should test if the covariance structure of the dataset is appropriate for a multidimensional solution with the DIMTEST method (Stout, 1987). In case the data is unidimensional one should consider an one dimensional item response theory approach (e.g. Bond & Fox, 2007). In case the covariance structure of the dataset supports a multidimensional solution, the DETECT method (Zhang & Stout, 1999) can be used to identify if the underlying structure is predominantly complex or simple. Additionally, DETECT offers the researcher at least some hint concerning the appropriate number of dimensions and the amount of multidimensionality from a pure statistical point of view. The suggestions of the program can then be compared to the theoretically expected partitioning of the test. Depending on the t_{max} statistic and one's own theoretical assumptions one can then chose the NOHARM method or in case of a more sophisticated model the SEM approach. Of course it can happen that one expects a rather simple model structure but the data are not of that kind according to DETECT. In such cases the researcher should keep this information in mind and take a closer look at his own theory. There are many possible reasons for unexpected results. The underlying theory could be wrong, the respondents too good or too bad for the test (Fogarty & Stankov, 1995), or maybe the instrument does not measure the construct according to the theory. As will be pointed out in the next section there are different ways to translate a theory into model that do not have to contradict one another. Finally, in case one comes to the conclusion that the data is of complex structure one can chose between a parcel based or item based SEM approach based on the amount of multicollinearity in the data.

One reason to create the decision chart for this article is that other approaches I found (Jasper, 2009) were very likely to fail when the data is of complex structure. One example is the decision graph by Nandakumar and Ackerman (2004) which is similar to the one presented here. There are at least two important main differences: Firstly, they suggest to apply the DIMTEST procedure multiple times (on the DETECT clusters) until the remaining item clusters are essentially unidimensional. This procedure is likely to produce results that are hardly interpretable in case of an underlying complex structure. Secondly, they do not include SEM as a confirmatory tool in their chart although the SEM approach is without doubt very useful in case of a complex test structure. Thus, my decision chart relies more on an existing theory behind the test that one analyzes. In the following section I am going to apply the ideas behind the decision chart to an existing dataset.

The Real Data Application of the Mathematics Test

Mathematics Content Domain

Of course, dimensionality and test structure assessment are both senseless without theoretical grounds concerning the construct. At the beginning of a test construction one normally reviews literature that already deals with similar constructs to get some idea of the construct structure.

The test I am going to analyze focuses on the mathematics domain for pupils and adults beyond age 15 but one should not ignore the results of the PISA large scale assessments. There have been attempts to spread some light on the dimensionality and structure of the PISA instruments to assess mathematic literacy (OECD, 2005). Brunner (2006) reanalyzed the PISA 2000

achievement data using the Conquest software (Wu, Adams, Wilson, & Haldane, 2007) and found that a post-hoc partitioning of the dataset into four content domains (arithmetic, geometry, algebra, stochastic) offered the best fit measured by the AIC (Akaike, 1974) compared with a g-factor and a partitioning by kind of mathematical working behavior. However, the differences in fit were quite small and it was concluded that mathematic literacy seems to be rather unidimensional. That is something one could also expect from the SAT mathematics test that was reanalyzed by Gierl et al. (2005) who used NOHARM, DIMTEST and DETECT to determine the structure of the test. They found out that a confirmatory 4-factor solution (algebra, arithmetic, geometry, miscellaneous) offered the best fit to the data. The official TIMSS 2007 report (IEA, 2008) contains information about the degree to which four content dimensions (data and chance, geometry, algebra, numbers) should be part of the TIMSS questions but does not report correlations that one would find partitioning the instruments into these four domains. In another document that focuses on the cognitive domain in the TIMSS 2003 assessment (IEA, 2005) the medians (over all participating countries) of the latent correlations between application, reasoning and knowledge lie between $r = .81$ and $r = .95$.

Taking a look at commercial mathematics tests for personnel selection in Germany, one finds three mathematics tests that are up to date. The *Rechentest 9+* (RT 9+; Bremm & Kühn, 1992), the *Berufsrechentest* (BRT; Balser & Ringsdorf, 1986) and the *Mathematiktest für Lehre und Beruf* (MATLUB; Ibrahimovic & Bullheller, 2005). What they all have in common is that they are more or less based on German curricula and that, according to the manuals, they have no further theoretical base. The number of subtests varies from four (MATLUB), to seven (RT 9+) to eight (BRT) and there are either no factor analytic results on item or parcel level reported (e.g. MATLUB) or the results do not support the partitioning into subtests (BRT, RT9+).

Instrument

I applied the approach to the *START-M* mathematics test (Jasper & Wagener, in press) here. It is a paper pencil test that takes about one hour and aims at testing thoroughly the mathematical abilities of job starters aged 16 years and older. The theory behind the test stems from the idea that in modern intelligence diagnostics many instruments reliably measure at least three content domains which are figural, numerical and verbal intelligence. Examples are the *intelligence structure test* (Beauducel, Liepmann, Horn, & Brocke, 2008), the *wilde intelligence test* (Kersting, Althoff, & Jäger, 2008), and the *berlin intelligence test* (Jäger, Süß, & Beauducel, 1997). This finally led to the idea to create a mathematics test with four scales inspired by these three domains. The scales are called *mathematical literacy* (LIT) as a counterpart to verbal intelligence, *geometry and graphical functions* (GEO) referring to figural intelligence and finally the two scales *procedural computation* (PROC) and *complex computation* (COMPL) which refer to numerical intelligence (Jasper & Wagener, in press). The numerical intelligence facet led to two mathematics scales because there was a practical need to differentiate between rather simple calculations (PROC) which only require basic rule knowledge and a scale that contains more complex items that require a thoroughly understanding of what one is doing (Jasper & Wagener, in press).

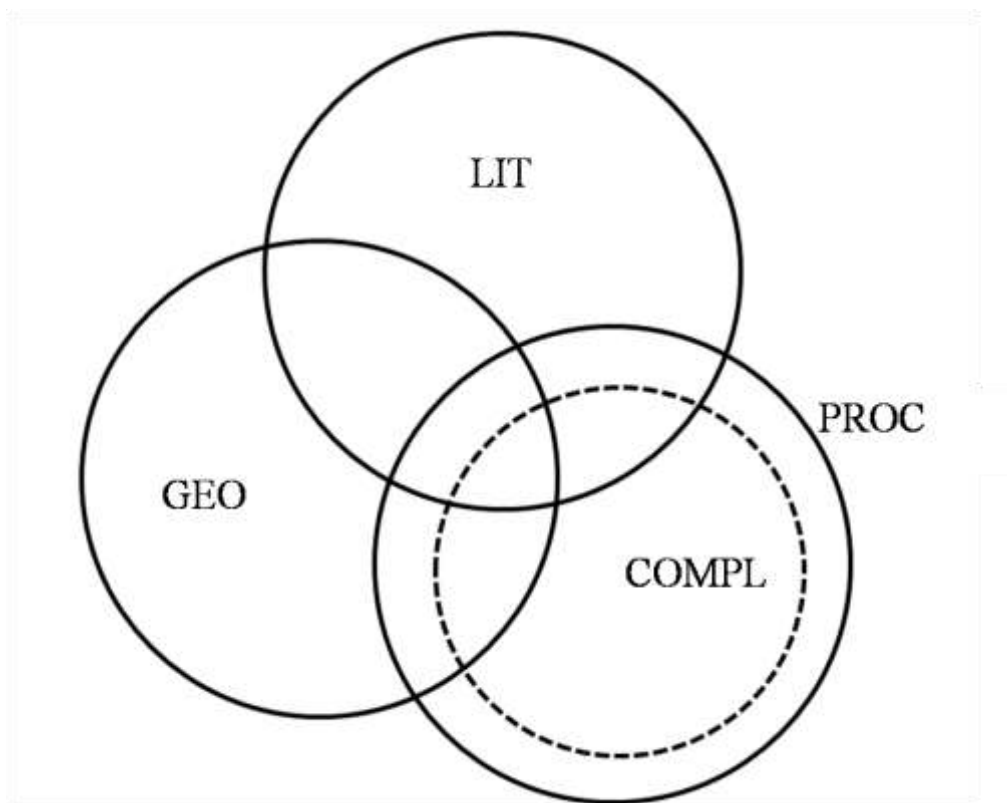


Figure 3. START-M model according to Jasper and Wagener (in press). GEO = Geometry and Graphical Functions, LIT = Mathematical Literacy, COMPL = Complex Calculations, PROC = Procedural Calculations.

Cronbach's α ranged from .82 for LIT to .88 for PROC indicating sufficient internal consistency. Because the scale intercorrelations reported are rather high ranging from $r(1,552) = .54, p < .01$ (GEO with LIT), to $r(1,552) = .74, p < .01$, (PROC with COMPL) one could raise doubt concerning the expected structure of the test which is depicted in Figure 3. The authors assume that the four dimensions are correlated and that PROC and COMPL are more similar to each other than the other dimensions.

Dataset

The dataset contains 1,554 respondents (67% male, 31% female, 2% unknown). They were collected in order to standardize the START-M (Jasper & Wagener, in press) between October 2008 and February 2009. About 59% of the pupils attended a school providing vocational education (Berufsschule) the rest attended continuative schools (33%) or were regular students of a high school (9%). The mean age was 20 years ranging from 14 to 41 years with $SD = 3.05$.

Results

In a first step, I applied the DIMTEST method to the dataset which yielded $T = 10.23 (N = 1,554, p < .01)$. This means that the data is not essentially

unidimensional according to the procedure. The DETECT program suggested an optimal partitioning of the test into four dimensions according to Table 1.

Table 1
Results of an exploratory DETECT solution (N = 1,554)

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Item	Scale	Item	Scale	Item	Scale	Item	Scale
1	GEO	2	GEO	5	GEO	13b	PROC
6a	GEO	3	GEO	13a	PROC	18	PROC
6b	GEO	4	GEO	13c	PROC	19	PROC
6c	GEO	7	GEO	15a	PROC	20a	PROC
10a	GEO	8	GEO	15b	PROC	21a	PROC
10b	GEO	9	GEO	15c	PROC	21b	PROC
10c	GEO	14a	PROC	15d	PROC	22a	PROC
11a	GEO	14b	PROC	16a	PROC	22b	PROC
11b	GEO	14c	PROC	16b	PROC	22c	PROC
11c	GEO	17a	PROC	16c	PROC	23a	PROC
11d	GEO	17b	PROC	20b	PROC	23b	PROC
12a	GEO	25b	LIT			23c	PROC
12b	GEO	25c	LIT			24a	PROC
		25d	LIT			24b	PROC
		26a	LIT			24c	PROC
		26b	LIT			24d	PROC
		26c	LIT			31a	COMPL
		27a	LIT			31b	COMPL
		27b	LIT			31c	COMPL
		27c	LIT			31d	COMPL
		27d	LIT			32a	COMPL
		27e	LIT			32b	COMPL
		28	LIT			34a	COMPL
		29	LIT			34b	COMPL
		30a	LIT			34c	COMPL
		30b	LIT			35	COMPL
		33	COMPL				

Note. GEO = Geometry and graphical functions, PROC = Procedural Computation, COMPL = Complex Computation, LIT = Mathematical Literacy.

However, the $D(P)$ value of .41 suggests an only weak amount of multidimensionality in the data (Kim, 1994). Because r_{max} takes the value of .72 the data can be assumed to be of a complex structure. The partitioning of the test suggested by DETECT is rather difficult to interpret. Cluster 1 contains only items of the GEO scale and Cluster 4 only items of the PROC and COMPL scale. Because PROC and COMPL are considered to be quite similar concerning their content both clusters could be judged to be conform to the theory behind the test. While Cluster 3 could be seen as reflecting the difference between the PROC and COMPL scale Cluster 2 contains all items of the LIT scale but also some items from all other domains. Because the test structure seems to be quite complex I continue with an SEM approach according to Figure 2. From the knowledge gained in the previous steps together with the model of Jasper and Wagener (in press) according to Figure 3 there are at least two types of models that can be judged to be conform to the theory of the test.

The easiest type of model that confirms to the theory is a correlated factor model with four or three content factors. In case of the four factor model this means that the items that belong to one scale only have loadings on one particular factor (i.e. all other loadings are fixed at zero). Moreover, each factor is correlated with the remaining three factors. While the four factor model consists of four factors according to the four scales of the test the three factor model contains a GEO, LIT and COMPL+PROC factor. This means that we treat the items that are supposed to measure COMPL and PROC as one scale. An example for a correlated factor model that consists of only two factors is shown on the left side of Figure 4.

Because of the complex structure and the rather unclear results from the DETECT method another type of model seems reasonable. I chose one that combines a mathematics g-factor and three to four content factors according to the scales that conforms to the theory behind the test. The most popular model that combines those two ideas is a so called Schmid-Leiman model (SLM; Schmid & Leiman, 1957).

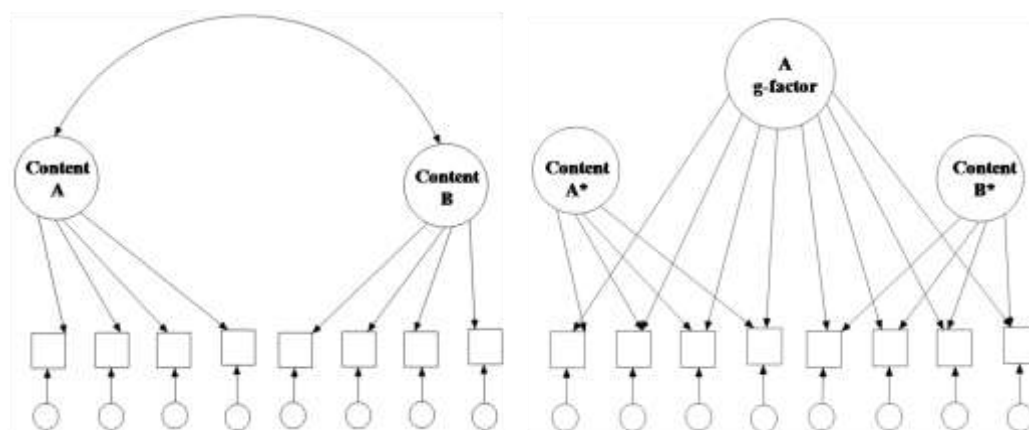


Figure 4. Example of a Schmid-Leiman Transformation (right side) of an Ordinary Model with Correlated Factors.

Figure 4 depicts a model of correlated content factors before (left) and after a Schmid-Leiman transformation. In this model a g-factor may account for as much common variance as possible among all the items (or parcels) and then the content factors account for the remaining group specific variances. These content factors are by definition uncorrelated. For the mathematics test, one can imagine the Schmid-Leiman g-factor as the general mathematical ability that is part of every item in the test. The four (or three) Schmid-Leiman group factors each represent the abilities that are specific to the particular scale and independent of the other scales and the g-factor. The SLMs that I propose are of the same type as the example in Figure 4. The only difference is that my models consist of three or four group factors instead of only two. The SLM has been applied in various areas of psychology, as for example the studies concerning the structure of the *16PF* (Chernyshenko & Stark, 2001), analyses of the *beck anxiety inventory* (Steer, 2009) or the *berlin intelligence structure model* (Beauducel & Kersting, 2002), to name just a few.

Because I wanted to compare those models with each other I decided to test all of them with the SEM approach (it would not be necessary for the ordinary correlated-factors models). Moreover, due to multicollinearity I

decided to parcel the items into mini-scales. A g-factor alone seemed to be rather unrealistic based on all previous results. To definitely rule out this option I also calculated a g-factor model and parceled the items according to the assumption of only one factor. The correlation between the four factors in the correlated factors model is shown in Table 2.

Table 2

Intercorrelations of the four latent factors reflecting the START-M scales (N = 1,554).

	Geometry and Graphical Functions	Procedural Calculations	Complex Calculations
Procedural Calculations	.72 (.63)		
Complex Calculations	.73 (.60)	.84 (.85)	
Mathematical Literacy	.58 (.47)	.68 (.69)	.71 (.69)

Note. All correlations are highly significant ($p < .01$). In parentheses: Correlation in case of ML-estimation, otherwise WLSMV based.

The highest correlation exists between procedural and complex calculations which confirms with the theory behind the START-M (see Figure 3, Jasper & Wagener, in press). Table 3 summarizes the results for all models I tested.

Table 3

Results of all SEMs of the START-M parcels (N = 1,554).

Model	χ^2	CFI	RMSEA	AIC ^a
Mathematics g-factor model	6,309.64*	.74	.117	9,8585
3 factors correlated	3,824.44*	.87	.096	9,6474
4 factors correlated	3,573.92*	.88	.092	9,6027
SLM, 3 factors	2,519.62*	.92	.073	9,4035
SLM, 4 factors	2,357.49*	.93	.070	9,3733

Note. ^aThe AIC is maximum likelihood based because there is no AIC available when using WLSMV. * $p < .01$. In all three factor models COMPL and PROC form one factor. SLM = Schmid-Leiman model.

The standardized factor loadings of the parcels in the correlated four-factor model were all above .40 and quite ordinary. However, as can be seen in Table 3 the SLMs yield the best fit in all fit indices. The best fit is offered by the four factor SLM. The standardized factor loadings are shown in Table 4 and as one can see some parcels yield zero or even negative loadings on their group factors (PROC and GEO), probably due to their loadings on the g-factor.

Table 4
Standardized factor loadings of the four factor Schmid-Leiman model

Parcel number ^a	Mathematics g-factor	Mathematical Literacy	Procedural calculations	Complex calculations	Geometry and graphical functions
1	.71	.53	-.20	.19	.15
2	.65	.54	-.07	.28	.17
3	.60	.68	-.23	.43	.33
4	.71	.31	.49	.35	.30
5	.61	.54	.45	.40	.32
6	.61	.35	.31		.62
7	.57	.44	.24		.78
8	.58		.31		.80
9	.70		.32		.32
10	.66		.33		.02
11	.83		.36		
12	.66		.05		
13	.75		.11		
14	.41		.42		
15	.40		.22		
16	.77				
17	.49				
18	.76				
19	.67				
20	.71				
21	.60				
22	.55				
23	.63				
24	.66				
25	.65				
26	.53				
27	.61				
28	.63				
29	.49				
30	.56				
31	.52				
32	.60				
33	.72				
34	.71				
35	.76				
36	.74				

Note. ^aFor the precise item to parcel allocation see Jasper (2009). Estimation method: WLSMV. The variances of all 5 factors were fixed at 1. All coefficients are highly significant ($p < .01$). $N = 1,554$.

Discussion

In this article I first presented some of the most up to date methods to determine the dimensionality and underlying structure of an achievement test. I suggested a decision graph and applied it to the analysis of a new mathematics test. The results show that a SLM with four group factors according to the scales of the test offers the best but still quite improvable fit to the data. It should be mentioned that it isn't so easy to judge the fit of a structural equation model because attempts to establish guidelines concerning overall fit indices (e. g., Hu & Bentler, 1999) have been the target of harsh critique (Marsh, Hau, & Wen, 2004).

One has to point out that this is the only mathematics test in the German language area for respondents of at least 16 years of age that is based on a theoretical model that has been tested and at least confirmed partially. However, further research is needed to test if the structural parameter estimates (and mean structures) also emerge in other populations (or sub groups of one population) and thus, measurement invariance holds (Kuljanin & Schmitt, 2008; Vandenberg & Lance, 2000).

The methods in this article helped to rule out a mathematics g-factor, showed us that the underlying structure is rather complex and that a multidimensional solution of four factors could be a promising solution. I presented a way to include these modern methods into the domain of applied dimensionality and test structure assessment which is quite necessary because without a broader application framework they are likely to produce results that are hardly interpretable. This also confirms to the results by Finch, Stage and Monahan (2008) who compared r_{max} and two other indices based on factor analysis. They concluded that none of the indices worked uniformly well in identifying the precise underlying item structure but that the r_{max} index might help in differentiating between simple and complex structures.

Another possibility to test the structure of an achievement test exists in form of the Multidimensional Random Coefficients Multinomial Logit Model (MRCML; Adams, Wilson, & Wang, 1997; Wu et al., 2007). This is a very broad generalization of the Rasch model to multiple (correlated) latent traits. It should be mentioned here because the MRCML has been used extensively within all TIMSS and PISA studies up to date and also offers the possibility to test complex models like the SLM. Unfortunately, with more than two factors the computing time becomes extremely large (Tate, 2003, p. 167). There are workarounds for this problem which include the estimation of the ability estimation with Monte Carlo methods (Wu et al., 2007). Nevertheless, it is not clear how the output of the program should be interpreted because it only offers overall fit indices for every item and the fit of the SLMs cannot be compared to the correlated factor models (Jasper, 2009).

The DIMTEST and DETECT methods require some extra effort and are not easy to understand from the start. The manual for the computer programs consists of some statistical articles (e.g. Stout, 1987; Zhang & Stout, 1999) with many (complicated) formulas and one article that applies the methods (Jang & Roussos, 2007) and is not very helpful to someone who just wants to assess the dimensionality of his test. The situation with the NOHARM program is slightly better because the manual that comes together with the software (McDonald & Fraser, 2003) gives some examples concerning

the interpretation of the program output. Probably Blinkhorn was right when he concluded that „much (most?) of what is still current in the theory of individual psychological differences has its origin in the work of a relatively small number of psychologists active between 1920 and 1950 ... With the fading from the scene of the Grand Old Men, contributors to test theory are much more rarely themselves test constructors.” (Blinkhorn, 1997, p. 177). This article should enable test developers to use and understand the DETECT, DIMTEST and NOHARM methods without further reading (although I recommend further reading).

The authors of the START-M (Jasper & Wagener, in press) suggest to interpret the four scale values as well as the sum of all items in the manual. Based on the results this is a reasonable practice. It is quite interesting that there are only two up to date mathematics tests (MATLUB, START-M) for adults in Germany although mathematical ability is a very important competence for daily living (IEA, 2005) and will be the focus of the upcoming PISA 2012 study. One reason is probably that mathematics test developers have to keep both, the curricula and the theoretically expected factor structure in mind at the same time. This problem becomes especially difficult in Germany because there are 16 different curricula for mathematics in the country, one for each province.

A very difficult question is to judge whether the decision chart fosters a too exploratory analysis of the test structure. As the Nobel-winning economist Ronald Coase once said “If you torture the data long enough, nature will always confess” (Coase, 1994, p. 27). I do think that the proposed model does not contradict the initial theory the authors propose and that it is most important to point out how one came to the final model (here, the four factor SLM). The decision chart is intended to help researchers that have some very precise ideas concerning the construct their test is supposed to measure. This should help to rule out a too exploratory approach that could create arbitrary results.

It would be quite interesting for future research to test the reliability and validity of the decision chart with simulated data. One could first create datasets with an either complex or simple underlying structure (e.g. like the one of the START-M) and varying numbers of latent traits (factors). It is very unrealistic that real life test developers simply apply methods like DIMTEST or DETECT and do not take their theoretical considerations into account. Therefore, in the next step, subjects who are familiar with the application of the methods are given some theoretical information concerning the construct (e.g. a g-factor, a speed and power factor etc.). Finally, the subjects apply the decision graph presented in this article and the outcomes of their analyses are compared with the true structure.

Although I focused on the mathematics content domain, the decision chart could be valuable in other research areas as well and I encourage to use it. There are no reasons why the described procedures couldn't be applied to a personality test or even tests from the domain of clinical psychology.

References

- Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.
- Balser, H., Ringsdorf, O., & Traxler, A. (1986). *Berufsbezogener Rechentest [Work related calculation test]*. Weinheim: Beltz.
- Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: New developments and techniques* (pp. 269-296). Mahwah, NJ: LEA.
- Beauducel, A. (1997). Transformations-matrix-search and identification (trasid): A new method for oblique rotation to simple structure. *Methods of Psychological Research Online, 2*, 113-138.
- Beauducel, A., & Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin model of intelligence structure. *European Journal of Psychological Assessment, 18*, 97-112.
- Beauducel, A., Liepmann, D., Horn, S., & Brocke, B. (2008). *IST - Intelligence structure test* (1th UK Ed.). Göttingen: Hogrefe.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least square estimation in confirmatory factor analysis. *Structural Equation Modeling, 13*, 186-203
- Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75.
- Blinkhorn, S. F. (1997). Past imperfect, future conditional: fifty years of test theory. *British Journal of Mathematical and Statistical Psychology, 50*, 175-185.
- Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: LEA.
- Bremm, M. H., & Kühn, R. (1992). *Rechentest RT 9+ [Calculation test 9+]*. Weinheim: Beltz.
- Brunner, M. (2006). *Mathematische Schülerleistung: Struktur, Schulformunterschiede und Validität. [Mathematical performance of pupils: Structure, school type differences and validity]*. (Unpublished doctoral dissertation). Humboldt University, Berlin.
- Champlain, A. D., & Gessaroli, M. E. (1996, April). *Assessing the dimensionality of item response matrices with small sample sizes and short test lengths*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.
- Champlain, A. D., & Tang, K. L. (1997). CHDIM: a fortran program for assessing the dimensionality of binary item responses based on McDonald's nonlinear factor analytic model. *Educational and Psychological Measurement, 57*, 174-178.
- Chernyshenko, O. S., & Stark, S. (2001). Investigating the hierarchical factor structure of the fifth-edition of the 16PF: An application of the Schmid-Leiman orthogonalization procedure. *Educational and Psychological Measurement, 61*, 290-302.

- Coase, R. H. (1994). *Essays on economics and economists*. Chicago, IL: University Press.
- Cortina, J. M. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of applied psychology, 78*, 98-104.
- Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement, 42*, 149-169.
- Finch, W. H., Stage, K. A. & Monahan, P. (2008). Comparison of factor simplicity. Indices for dichotomous data: DETECT R, Bentler's Simplicity index and the loading simplicity index. *Applied Measurement in Education, 21*, 41-64.
- Fogarty, G. J., & Stankov, L. (1995). Challenging the law of diminishing returns. *Intelligence, 21*, 157-174.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: least squares item factor analysis. *Multivariate Behavior Research, 23*, 267-269.
- Fraser, C., & McDonald, R. P. (2003). NOHARM. A windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Retrieved from <http://people.niagaracollege.ca/cfraser/download/nhmanw/nhman.html>
- Froehlich, A. G., & Stout, W. F. (2003). *A new bias correction method for the DIMTEST procedure*. (Unpublished Manuscript). Iowa State University.
- Green, S. B. (1981). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7*, 139-147.
- Green, S. B., Lissitz, R. W. & Mulaik, S. A. (1977). Limitations of Coefficient Alpha as an index of test unidimensionality. *Educational and psychological measurement, 37*, 827.
- Gierl, M. J., Tan, X., & Wang, C. (2005). Identifying content and cognitive Dimensions on the SAT. *College Board Research Report, 11*, 1-31.
- Gierl, M. J., Leighton, P., & Tan, X. (2006). Evaluation DETECT classification accuracy and consistency when data display complex structure. *Journal of Educational Measurement, 43*, 265-289.
- Gorsuch, R. L. (1983). *Factor Analysis*. Hillsdale, NJ: LEA.
- Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.
- Hattie, J. (1985). Methodology Review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.
- Hattie, J., Krakowski, K., Roger, H., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1-14.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, FL: Harcourt.
- Hu, L-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.
- Ibrahimovic, N., & Bulheller, S. (2005). *Mathematiktest. Grundkenntnisse für Ausbildung und Beruf [Math test. Base knowledge for education and work]*. Frankfurt: Harcourt.
- IEA. (2005). *TIMSS 2003: International report on achievement in the mathematics cognitive domains*. Boston, MA: TIMSS & PIRLS International Study Center.

- IEA. (2008). *TIMSS 2007: International mathematics report*. Boston, MA: TIMSS & PIRLS International Study Center.
- Jäger, A. O., Süß, H-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test*. Göttingen: Hogrefe.
- Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance based nonparametric approach. *Journal of Educational Measurement, 44*, 1-21.
- Jasper, F. (2009). *Zur Psychometrie der Mathematik am Ende der Sekundarstufe I [Psychometrics in the math domain at the age of 16+]*. (Unpublished doctoral dissertation). University of Mannheim, Germany.
- Kersting, M., Althoff, K., & Jäger, A. O. (2008). *Wilde-Intelligenz-Test 2*. Göttingen: Hogrefe.
- Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science, 45*, 106-110. Jasper, F., & Wagener, D. (in press). *START-M: Mathematik. Testbatterie für Berufseinsteiger [START-M: Mathematics. Test battery for job starters]*. Göttingen: Hogrefe.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.
- Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.
- Kline, R. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford press.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173.
- Linacre, J. M. (1998) Rasch first or factor first? *Rasch Measurement Transactions, 11*, 603.
- Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika, 68*, 49-60.
- Lubinski, D., Webb, R. M., Morelock, M., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86*, 718-729.
- Marsh, H. W., Hau, K-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.
- McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika, 27*, 397-415.
- McDonald, R. P. (1986). Describing the elephant: Structure and function in multivariate data. *Psychometrika, 4*, 513-534.
- McDonald, R. P. (1999). *Test Theory. A unified treatment*. Mahwah, NJ: LEA.
- McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99
- McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin, 43*, 289-374.
- Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-243). Newbury Park, CA: Sage.

- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, B., du Toit, S. H-C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.
- Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 93-106). Thousand Oak, CA: SAGE.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, 18, 41-68.
- OECD. (2005). *PISA 2003 technical report*. Retrieved from <http://www.oecd.org/dataoecd/49/60/35188570.pdf>
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: sage.
- Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 53-61.
- Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18, 210-222.
- Seraphine, A. E. (2000). The performance of dimtest when latent trait and item difficulty distributions differ. *Applied Psychological Measurement*, 24, 82-94.
- Steer, R. A. (2009). Amount of general factor saturation in the Beck Anxiety Inventory responses of outpatients with anxiety disorders. *Journal of Psychopathological Behavior Assessment*, 31, 112-118.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52, 589-617.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159-203.
- Thompson, B. & Vidal-Brown, S. A. (2001, February). *Principle components versus principle axis factors: when will we ever learn?* Annual meeting of the southwest educational research association. New Orleans.
- Thurstone, L. L. (1947). *Multiple-Factor Analysis*. Chicago, IL: University of Chicago Press.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4-70.
- Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis. Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, 25, 1-28.
- Verhelst, N. (2001). Testing the unidimensionality assumptions of the Rasch model. *Methods of Psychological Research Online*, 6, 231-271.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *Acer Conquest Version 2.0*. Melbourne: ACER.

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

Author Note

Fabian Jasper, Department of Clinical Psychology, Johannes Gutenberg University Mainz.

I would like to thank Dr. Dietrich Wagener (University of Mannheim) and Prof. Dr. Liepmann (FU Berlin) for their great support.

Correspondence concerning this article should be addressed to Fabian Jasper, Department of Clinical Psychology, Johannes Gutenberg University, Mainz, 55122, Germany. E-mail: fabian.jasper@gmx.net