International Journal of Statistics and Probability; Vol. 5, No. 5; September 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Construct-focused Configural Invariance for Measures Showing a Multi-dimensional Structure and Application to Exchange Test Data

Karl Schweizer[1], Xuezhu Ren[2] & Michael Schreiner[3]

[1] Department of Psychology, Goethe University Frankfurt, Frankfurt, Germany

[2] School of Education, Huazhong University of Science and Technology, Wuhan, China

[3] University of Education, Heidelberg, Germany

Correspondence: Karl Schweizer, Department of Psychology, Goethe University Frankfurt, Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt am Main, Germany. E-mail: K.Schweizer@psych.uni-frankfurt.de

Received: July 26, 2016 Accepted: August 15, 2016 Online Published: August 29, 2016

doi:10.5539/ijsp.v5n5p55 URL: http://dx.doi.org/10.5539/ijsp.v5n5p55

Abstract

The concept of construct-focused configural invariance is proposed for investigating the measurement invariance of measures that need to be represented by means of a multi-dimensional model with constrained discriminability. The major characteristic of this concept is the concentration of the investigation of invariance on the component representing the process or processes associated with the construct of interest. This concept enables the exclusion of other components accounting for irrelevant but systematic variance from an investigation. Three large sets of Exchange Test data that differed according to the applied version of the Exchange Test and the population from which the samples originated were investigated according to this concept. Two samples were university students and the third one internet users. It was possible to establish construct-focused configural invariance and corresponding metric invariance for Exchange Test.

Keywords: configural invariance, measurement invariance, confirmatory factor analysis, fixed-links modeling, impurity model, ceiling effect

1. Introduction

The concept of measurement invariance has been proposed and applied with the congeneric model of measurement in mind. This paper seeks to transfer this concept to complex models of measurement with constrained discriminability. Complex models of measurement tend to consider several sources of systematic variance and covariance besides the source associated with the construct which the measure is expected to represent. It is argued that the demonstration of construct-focused invariance may be sufficient for the generalization of structural validity. This type of invariance is demonstrated with respect to a cognitive test that has already been investigated by means of complex models of measurement with constrained discriminability in three large datasets.

Psychological measures representing psychological constructs are usually expected to apply to all members of the population equally well. However, the investigation of the psychometric properties of such measures is usually conducted on the basis of a restricted sample only, and the representativeness of the sample is rarely considered in the investigation. Therefore, the range of the applicability of such measures has always been and is the object of scrutiny and concerns (Widaman & Reise, 1997). It is well-known that there may be differences regarding gender, age or cultural background that may disadvantage the members of one group as compared to the members of other groups (French & Finch, 2016). Such differences are considered as differential item functioning in the framework of item response theory (Karami, 2012; Swaminathan & Rogers, 1990) and as lack of measurement invariance in the framework of the factor-analytic approach (Meredith, 1993). Nowadays the demonstration of the absence of differential item functioning or alternatively of measurement invariance is considered highly desirable.

The invariance analysis that is in the focus of this paper serves the comparison of parameter estimates obtained from different samples or different groups by means of factor-analytic methods. It provides the opportunity to detect variation in the functioning of the measure or of items that are part of a measure (for an overview see Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). The investigation of invariance is mostly conducted in a step-wise manner including the analysis of configural invariance, invariance of factor loadings that is also addressed as metric invariance, invariance of measurement intercepts that is also referred to as scalar invariance, invariance of error variances and partial invariance (Thompson, 2016). Configural invariance and metric invariance are especially important regarding the validity of a measure since they are concerning the representation of the construct by the measure. Configural invariance means that the pattern of factor loadings is retained unchanged in the investigated samples or groups and metric invariance the sizes of the factor loadings.

Invariance analysis is mostly conducted in assuming a unidimensional model of measurement that shows the characteristics of the congeneric model of measurement (Jöreskog, 1971). Within the framework of this model the observations included in the p×1 vector **x** [=(x_1 ,…, x_p)’] are described as follows

$$\mathbf{x} = \boldsymbol{\lambda}\xi + \boldsymbol{\delta} \tag{1}$$

where $\boldsymbol{\lambda}$ is the p×1 vector of factor loadings, $\boldsymbol{\xi}$ [=(ξ_{11})'] is the 1×1 vector of latent variables, and $\boldsymbol{\delta}$ [=(δ_1, …, δ_p)'] is the p×1 vectors of error components. Furthermore, there is the multi-dimensional version of this model of measurement that is occasionally considered:

$$\mathbf{x} = \boldsymbol{\Lambda}\xi + \boldsymbol{\delta} \tag{2}$$

It differs from the unidimensional version in that $\boldsymbol{\lambda}$ is replaced by the p×q matrix of factor loadings $\boldsymbol{\Lambda}$ and in that the 1×1 vector $\boldsymbol{\xi}$ of latent variables is extended to a q×1 vector. A major characteristic of this multi-dimensional version is that there are no cross-loadings although the latent variables may be allowed to correlate with each other. This means that the product of every pair of vectors $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ (i≠j) from the set of factor loading vectors {$\boldsymbol{\lambda}_1$, …, $\boldsymbol{\lambda}q$} that constitute $\boldsymbol{\Lambda}$ amounts to zero:

$$0 = \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{\lambda}_j \tag{3}$$

However, the unidimensional representation of measures by means of vectors showing no cross-loadings presupposes specific properties: the items constituting a measure must show either a very high degree of purity regarding the representation of the construct or a high degree of variety so that the effects of auxiliary processes contributing to performance besides the main processes are negligible or averaged out. If there is a high degree of similarity among the items or trials serving as items instead of variety and various processes contribute to performance, a unidimensional representation cannot be good. Such similarity can be found in many cognitive measures capturing basic cognitive processes. Because of the similarity of items or trials performance is likely to reflect contributions of auxiliary processes besides the process or processes of interest. For example, according to Jensen (1982) four different processes contribute to the choice reaction time considered as indicator of mental speed; among them are motor processes that are not indicative of mental speed. Cognitive measures have repeatedly been denounced to show such kind of impurity (Miyake et al., 2000, Schweizer, 2007; Van Zomeren & Brouwer, 1994). Impurity can be due to similarity among the items or trials or due to the participants’ using different strategies in completing the items of the measure. There are reports of participants deviating from the expected processing strategy and this way changing the nature a measure (e.g., Patensko & Altman, 2010; Salnaitis, Baker, Holland, & Welsh, 2011). Impure measures lead to systematic variance that is irrelevant variance besides true variance. Such measures may show a high degree of reliability whereas the validity is likely to be low.

One way of dealing with such impurity is to represent it as part of a multi-dimensional model that separates impurity from the contribution of the cognitive process or processes of interest (Schweizer, 2007). In this case cross-loadings must be accepted. As a consequence Equation 3 cannot be retained. Furthermore, constrained factor loadings have to replace free factor loadings to assure that the observed variances and covariances are decomposed according to expectations that may reflect characteristics of items or treatment levels. The basic version of the impurity model assumes two latent variables for representing the core processes (=process or processes associated with the construct of interest) and the auxiliary processes that are typically assumed to contribute to each item equally. Accordingly the p×2 matrix of factor loadings $\boldsymbol{\Lambda}$ of an impurity model comprises 1s in the first column whereas constants reflecting the expectations regarding each item i (i=1, …, p) provide the contents of the second column. Since the use of letter “e” as symbol for this constant may be misleading, the word *expectation* is used:

$$\Lambda_{\mathrm{Impurity\_model}} = \begin{bmatrix} 1 & \mathrm{expectation}_{\mathrm{itm1}} \\ 1 & \mathrm{expectation}_{\mathrm{itm2}} \\ . & . \\ . & . \\ 1 & \mathrm{expectation}_{\mathrm{itm}p} \end{bmatrix} . \tag{4}$$

The constraint of factor loadings means the constraint of the discriminability of items or trials serving as observed variables of the model of measurement (Lucke, 2005). In assessment constraints also characterize the Rasch model (1960), the corresponding one-parameter item response model (Birnbaum, 1968), the linear logistic test model (Scheiblechner, 1972) and the tau-equivalent model (Graham, 2006; Lord & Novick, 1968; Schweizer, 2012). The constraint of factor loadings does not impair model fit in confirmatory factor analysis if the model is correct with respect to the data (Schweizer, Ren, Wang, & Zeller, 2015).

However, because of the constraints the matrix of factor loadings cannot play the major role in the invariance analysis of a measure, which it would play otherwise. In order to point out this argument in more detail and to develop an alternative, the model of the covariance matrix needs to be introduced. The model of the covariance matrix (Jöreskog, 1970) in confirmatory factor analysis is a general framework, according to which the pxp covariance matrix $\mathbf{\Sigma}$ is structured as follows:

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta} \tag{5}$$

where $\mathbf{\Lambda}$ is the pxq matrix of factor loadings, $\mathbf{\Phi}$ is the qxq matrix of the variances and covariances of the q latent variables and $\mathbf{\Theta}$ is the pxp diagonal matrix of error variances. The specification as impurity model requires the integration of the matrix of factor loadings according to Equation 4 into the model according to Equation 5:

$$\mathbf{\Sigma}_{\text{Impurity\_model}} = \mathbf{\Lambda}_{\text{Impurity\_model}}\mathbf{\Phi}\mathbf{\Lambda}'_{\text{Impurity\_model}} + \mathbf{\Theta} \quad . \tag{6}$$

In order to assure that the product of the matrix of factor loadings and the matrix of the variances and covariances of the q latent variables accounts for the systematic variance of the data, the parameters of $\mathbf{\Phi}$ need to be estimated. Since systematic variance and covariance cannot be at the same time pure and impure variance, overlap is excluded by setting the off-diagonal elements equal to zero so that

$$\mathbf{\Phi}_{\text{Impurity\_model}} = \text{diag}(\mathbf{\Phi}) \quad . \tag{7}$$

Estimates of the diagonal elements of this matrix, that are larger or smaller than one, lead after standardization to deviations of the factor loadings from the numbers assigned as constraints. Invariance analysis of measures represented by the impurity model must concentrate on this matrix instead on the matrix of factor loadings.

In the following section configural invariance and metric invariance are discussed with respect to models with free and constraint factor loadings. Furthermore, the concept of construct-focused configural invariance is introduced.

1.1 Measurement Invariance

1.1.1 Configural Invariance

Configural invariance means that the indicators originating from the same measure and serving as the manifest variables of the confirmatory factor model have factor loadings on the same latent variable in the application to the n datasets included in $\mathbf{X}$ (=$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$) (Thompson, 2016). For this purpose the confirmatory factor model is specified as multiple group model, and multiple group analysis is conducted. The datasets may originate from different samples of the same population or of different groups, as for example gender groups, age groups or ethnic groups. A good model fit suggests invariance with respect to the considered population or groups.

The investigation of configural invariance requires that the same pattern of fixed and free factor loadings of the pxq matrix $\mathbf{\Lambda}$ is used for investigating each dataset. If there is only one measure without subscales, $\mathbf{\Lambda}$ reduces to the px1 vector $\boldsymbol{\lambda}$ with free factor loadings referred to as $\boldsymbol{\lambda}_{\text{free}}$. If there are subscales, $\mathbf{\Lambda}$ must be specified by fixing some elements to zero in such a way that there are no cross-loadings. The corresponding matrix of factor loadings is characterized as $\mathbf{\Lambda}_{\text{freefixed}}$. Configural invariance required that the fit results observed in multiple group analysis of $\mathbf{X}$ (=$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$) by means of the fit functions of the fit indices with lower and upper boundaries, $f_{\text{l-fit}}$ and $f_{\text{u-fit}}$, are smaller respectively larger than the cut-offs that mark the boundaries of the ranges of acceptable values:

$$f_{\text{l-fit}}\left(\mathbf{X}\middle|\mathbf{\Lambda}_{\text{freefixed}}\right) \le \text{cut-off} \quad \text{and} \quad f_{\text{u-fit}}\left(\mathbf{X}\middle|\mathbf{\Lambda}_{\text{freefixed}}\right) \ge \text{cut-off} \quad . \tag{8}$$

Equations 8 and some of the following Equations are written in using the notation of conditional probabilities since there are important dependencies on conditions and the outcomes of the considered indices vary between zero and one exactly or to a large extent.

In multi-dimensional models with constrained factor loadings, as for example the impurity model, all elements of the pxq matrix $\mathbf{\Lambda}$ are set equal to numbers based on expectations regarding the underlying processes. The adaptation of the model to data is achieved by estimating the variances and covariances included in the qxq matrix $\mathbf{\Phi}$; in most cases only the variances are estimated since a major aim is the separation of variance due to the process or processes of interest from other processes. Therefore, the further reasoning concentrates on diag($\mathbf{\Phi}$) while there is fixation of the elements of $\mathbf{\Lambda}$ that is now identified as $\mathbf{\Lambda}_{\text{fixed}}$. Accordingly, in multi-dimensional models with constrained factor loadings configural invariance means that

$$f_{\text{l-fit}}\left(\mathbf{X}\middle|\mathbf{\Lambda}_{\text{fixed}}, \text{diag}(\mathbf{\Phi})_{\text{free}}\right) \le \text{cut-off} \quad \text{and} \quad f_{\text{u-fit}}\left(\mathbf{X}\middle|\mathbf{\Lambda}_{\text{fixed}}, \text{diag}(\mathbf{\Phi})_{\text{free}}\right) \ge \text{cut-off} \ . \tag{9}$$

1.1.2 Construct-focused Configural Invariance

Since the main purpose of the construction of a measure is the achievement of a valid representation of the process or processes associated with the construct of interest, the focus must be on the process or processes of interests. However, there may be other processes that contribute to performance and, therefore, need to be represented in a multi-dimensional model of the measure in order to account for the whole systematic variance of the data appropriately, as is the case in the impurity model. In order to adapt the concept of measurement invariance to this specificity of multi-dimensional models with constrained factor loadings, another variant is proposed: construct-focused configural invariance. Construct-focused configural invariance denotes the configural invariance restricted to the core component that reflects the process or processes associated with the construct of interest.

Construct-focused configural invariance concentrates on the column of $\mathbf{\Lambda}_{\text{fixed}}$ that refers to the process or processes associated with the construct of interests. It is represented by a px1 vector $\boldsymbol{\lambda}$ and identified as $\boldsymbol{\lambda}_{\text{construct\_fixed}}$, and it requires the concentration on the corresponding part of the qxq matrix $\mathbf{\Phi}_{\text{free}}$ that is the parameter $\phi_{\text{construct\_free}}$. Construct-focused configural invariance with respect to the data of $\mathbf{X}$ (=$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$) requires that

$$f_{\text{l-fit}}\left(\mathbf{X}\middle|\boldsymbol{\lambda}_{\text{construct\_fixed}}, \phi_{\text{construct\_free}}\right) \leq \text{cut-off and } f_{\text{u-fit}}\left(\mathbf{X}\middle|\boldsymbol{\lambda}_{\text{construct\_fixed}}, \phi_{\text{construct\_free}}\right) \leq \text{cut-off} \quad . \tag{10}$$

The consideration of the concept of construct-focused configural invariance in combination with the impurity model enables the exclusion of changes regarding the auxiliary processes from influencing the outcome.

1.1.3 Metric Invariance

Metric invariance is given if the sizes of the factor loadings are the same in all samples or groups that are investigated (Thompson, 2016). It presupposes configural invariance and assures that the process or processes associated with the construct of interests are represented in exactly the same way in all samples or groups. It guarantees that the meaning of a theoretical concept is the same in different samples or groups. Lack of metric invariance indicates an implicit change of the representation with consequences for the meaning of the construct.

Metric invariance is usually investigated by comparing the chi-square result obtained in investigating configural invariance with the chi-square result achieved after constraining the free factor loadings of $\mathbf{\Lambda}_{\text{freefixed}}$ in such a way that they show the same size for all samples or groups. This means a modification of the multiple group model such that $\mathbf{\Lambda}_{\text{freefixed}}$ is replaced by $\mathbf{\Lambda}_{\text{samefixed}}$. In the application metric invariance with respect to the data of the n datasets of $\mathbf{X}$ (=$\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$) means that the constraint of the factor loadings to equal sizes does not lead to an impairment of model fit:

$$\chi^2_{\text{samefixed}} - \chi^2_{\text{freefixed}} \leq \chi^2_{\text{critical}} \tag{11}$$

where the subscripts *samefixed* and *freefixed* refer to $\mathbf{\Lambda}_{\text{samefixed}}$ and $\mathbf{\Lambda}_{\text{freefixed}}$ in corresponding order and the basic model is the congeneric model of measurement. For this investigation the χ^2_{critical} that marks the boundary of acceptable values has to be selected according to the requirements of the chi-square difference test.

In the application to multiple group models with contrained factor loadings $\mathbf{\Lambda}_{\text{fixed}}$, as for example the impurity model, diag($\mathbf{\Phi}$)$_{\text{free}}$ has to be replaced by diag($\mathbf{\Phi}$)$_{\text{samefree}}$:

$$\chi^2_{\text{samefree}} - \chi^2_{\text{free}} \leq \chi^2_{\text{critical}} \quad . \tag{12}$$

Furthermore, if it is the aim to establish construct-focused configural invariance, $\phi_{\text{construct\_free}}$ needs to be changed into $\phi_{\text{construct\_samefree}}$:

$$\chi^2_{\text{construct\_samefree}} - \chi^2_{\text{construct\_free}} \leq \chi^2_{\text{critical}} \quad . \tag{13}$$

Moreover, there is the possibility to investigate full configural invariance according to Equation 9 in the first step and to restrict the investigation of metric invariance in the second step to the major component, as is given by Equation 13.

1.2 Exchange Test Analysis

The Exchange Test (Schweizer, 1996) has been constructed as a computer-based measure of working memory capacity and has repeatedly been the subjects of attempts of purification or rather attempts of optimizing the representation of the construct by a confirmatory factor model. The construction of this test was mainly theory-driven and resulted into a set of rather homogeneous items. The original version of this Test includes five treatment levels that clearly differ according to their difficulty and the time necessary for completing an item. There is also a modified version that excludes the most difficult treatment level and employs graphically improved stimuli (Schreiner, Altmeyer, & Schweizer, 2012). This modified version can be applied online.

Because of concerns regarding the purity of cognitive tests (e.g., Jensen, 1982; Miyake et al., 2000, Van Zomeren & Brouwer, 1994), the representation of Exchange Test data by the impurity model mentioned in the first section is recommended. This model shows the structure of a fixed-links model (FLM) that is used for the theory-guided decomposition of variances and covariances (Schweizer, 2006a, 2006b, 2008) and is frequently considered in combination with experimentation. It is set up on the model of the covariance matrix that is basic to confirmatory factor analysis. Furthermore, it shares the mathematical foundation with early growth curve models (McArdle, 1986, 1989, 2009) and the tau-equivalent model of measurement (Lord & Novick, 1968) that is obvious in the constraint of factor loadings. The impurity model seeks to separate the process or processes of interest that provide the genuine contribution to working memory performance from the contributions of auxiliary processes, as for example perceptual and motor processes. These auxiliary processes are necessary for completing the items but are unrelated to working memory.

There is another problem regarding cognitive measures that is increasingly attracting attention in science: the use of a variety of strategies in completing the items of a measure (e.g., Geiser, Lehmann, & Eid, 2006; Perret, Bailleux, & Dauvier, 2011; Salnaitis et al., 2011; Schweizer, 2009). The use of strategies can become a problem if participants are expected to follow the prescribed strategy. In completing the Exchange Test it is necessary to follow the prescribed strategy; otherwise the test does no longer represent working memory capacity. However, since there is no feedback of the outcome of a trial, the participant does not learn that the selected alternative strategy leads in most cases to an incorrect result. In order to overcome this problem, the modeling of the most likely alternative strategy that is the simplicity strategy is recommended (Schweizer, Altmeyer, Ren, & Schreiner, 2015).

Furthermore, there is a practical problem of modeling cognitive data that is also concerning the modeling of Exchange Test data. It is the impairment of the distribution of very easy items that is known as ceiling effect. The ceiling effect does not only mean a distortion of the distribution but also a reduction of the variance (Hessling, Traxel, & Schmidt, 2004; Vogt, 2005). Furthermore, the ceiling effect creates skewness, and skewness is considered as a source of model misfit in confirmatory factor analysis (Fan & Hancock, 2012; West, Finch, & Curran, 1995). In order to overcome the adverse consequences of the ceiling effect on model fit, the modeling of this effect is proposed (Schweizer, 2016). The modeling of the effect concentrates on the reduction of variance and improves the model fit regarding Exchange Test data considerably.

It is a truly complex confirmatory factor model with constrained factor loadings that is finally achieved for representing working memory capacity by means of Exchange Test data. It includes a latent variable that can be expected to provide an especially pure representation of working memory capacity.

2. The Present Study

The study that is reported in the following sections was conducted in order to demonstrate that *construct-focused configural invariance* is a useful concept for the evaluation of the psychometric quality, especially of construct validity, of a psychological measure that requires the statistical control of distorting sources that are unrelated to the construct represented by the measure but also contribute to performance. The Exchange Test was selected for this study because of its representation by a multi-dimensional model taking various sources into consideration that was developed over the years. This model included one latent variable that was expected to represent working memory capacity whereas the other latent variables represented something else. The demonstration of invariance was highly desirable for this latent variable since multiple latent variables could be suspected to show variability instead of constancy over different samples. In contrast, invariance of the other latent variables was not considered to be necessary.

Invariance was investigated with respect to three large datasets. The first dataset was collected by means of the original Exchange Test version in university students and the second dataset by means of the modified version also in university students. The third dataset originated from an internet sample. In using all available possibilities of addressing internet users, participation in this study was advertised. Internet users who agreed to participate were logged in via internet and received the modified Exchange Test version for completion.

2.1 Method

The first sample included 449 participants, the second sample 461participants and the third sample 418 participants. The third sample only included those internet users who tried all items of the Exchange Test. The mean ages of the first and second samples were 24.8 (SD=3.7) and 22.3 (SD=4.3) years of age in corresponding order. The first sample included 185 males and 264 females and the second sample 183 males and 278 females. Gender and age information was not available for the third sample since the internet users were not asked to provide this information to assure anonymity in a convincing way.

Basic data cleaning was conducted to exclude participants who did not really try to solve the items before investigating the data. Participants who were incorrect in more than 50 percent of the easiest items were eliminated since the probability of completing these items correctly was very high. Furthermore participants who committed more errors in the two easy treatment levels than in the two difficult treatment levels were excluded since the items of the two difficult treatment levels were considerably more difficult than the other ones. This cleaning reduced the first sample to 442 participants, the second one also to 442 participants and the third one to 378 participants.

The statistical investigation was conducted in two steps. The first step served the investigation of the individual datasets in order to find out how well the confirmatory factor models fitted to the individual datasets. In each case a set of four different models and, if necessary, modifications of these models served this investigation. The first model was a fixed-links model including one latent variable only. The factor loadings of the scores of the treatment levels on this latent variable were fixed to numbers according to the cognitive load that were 1, 2, 3, 4 and 5 for the original version and 1, 2, 3 and 4 for the modified version. This model was denoted *working memory model*. The second model was the *impurity model* that additionally included the latent variable for representing impurity. The constraints for the factor loadings of the impurity latent variable were 1, 1, 1, 1 and 1 respectively 1, 1, 1 and 1. The third model was referred to as *ceiling effect model* since it included weights for modeling the ceiling effect characterizing the scores of the first treatment level. The weights that were computed for the three datasets varied between 0.327 and 0.362. Finally there was the model including another latent variable for representing the simplicity strategy. It was addressed as *simplicity model*. The loadings on the simplicity latent variable were fixed according to the numbers proposed for the original Exchange Test version that were 0.5 for the indicators of the third and fourth levels and 1.0 for the fifth level. It turned out that the switch from the original version to the modified version required a change of the constraints for the representation of the simplicity strategy. The new constraints were 0.5 for the second treatment level, 1.0 for the third treatment level and 0.5 for the fourth treatment level.

In the second step the models for the three datasets were combined in order to achieve multiple group models. These multiple group models were specified to investigate different types of construct-focused configural invariance. Furthermore, metric invariance was investigated for the configurations showing configural invariance.

Since the Exchange Test yielded scores based on the outcomes of 12 binary items for each level, these scores were treated as continuous and parameter estimation was conducted using maximum likelihood estimation in LISREL (Jöreskog & Sörbom, 2006). For the evaluation of model-data fit the following fit statistics were used: chi-square, RMSEA, SRMR, GFI, NNFI and CFI. The results were compared with the following cut-offs: RMSEA .06, SRMR .08, GFI .95, NNFI .95, CFI .95 (see DiStefano, 2016). The comparisons of models applied to individual datasets were conducted by means of CFI and AIC. The chi-square difference test served the comparison of models with and without equality constraints.

2.2 Results

2.2.1 Results for the First Dataset

The fit statistics observed when applying the four models to the dataset collected by means of the original Exchange Test are presented in Table 1. Each row provides the results for one model. The chi-square, RMSEA, SRMR and AIC results showed a decrease from the first to fourth models. In contrast, the GFI, TLI and CFI results increased. The simplicity model yielded the largest number of fit statistics indicating a good model fit (SRMR, GFI, NNFI, CFI). Furthermore, RMSEA and AIC showed the lowest values. The RMSEA value of the fourth model indicated an acceptable model fit but not a bad or good one. Apparently, the consideration of impurity, ceiling effect and simplicity strategy was necessary for achieving a good or acceptable model fit. Furthermore, from each model to the next model there was a large reduction in chi-square and an improvement of model fit according to AIC. The CFI signified a better fit for the third model over the first and second models and a better fit for the fourth model over all the other models.

Table 1. Fit Results Observed in Investigating Exchange Test Data – Original Version (5 indicators) by means of Models Considering Working Memory, Impurity, Strategy Use and Weights (N = 442)

Model	χ^2	df	RMSEA	SRMR	GFI	NNFI	CFI	AIC
Working memory model (including working memory latent variable only)	107.07	9	0.152	0.130	0.92	0.85	0.87	113.07
Impurity model (including additionally impurity latent variable)	94.38	8	0.156	0.120	0.92	0.84	0.87	108.38
Ceiling effect model (including additionally weights for first indicator showing ceiling effect)	39.47	8	0.094	0.065	0.97	0.94	0.95	53.47

Simplicity model (including additionally a representation of simplicity strategy)	24.95	7	0.076	0.041	0.98	0.96	0.97	40.95

2.2.2 Results for the Second Dataset

The modified Exchange Test led to the fit statistics that are reported in Table 2. The results of the first row of this Table showed the worst model fit for the working memory model; it was even worse than the results for the corresponding model of Table 1. The consideration of impurity led to a considerable improvement of model fit, as is obvious from the second row of this Table. The inclusion of weights leading to the ceiling effect model and of the other latent variable for reaching the simplicity model both improved model fit also. The best fit was achieved for the simplicity model. Good model fit was indicated by the SRMR, GFI and CFI statistics, and the AIC reached the lowest value. The NNFI results indicated an acceptable model fit whereas the RMSEA was not good. An attempt was made to further improve the model fit by eliminating the impurity latent variable. However, it did not improve the model fit, as is obvious from the last row of this Table. While the RMSEA statistic became better, the CFI statistic worsened.

Table 2. Fit Results Observed in Investigating Exchange Test Data – Modified Version (4 indicators) and Standardized Condition by means of Models Considering Working Memory, Impurity, Strategy Use and Weights (N = 442)

Model	χ^2	df	RMSEA	SRMR	GFI	NNFI	CFI	AIC
Working memory model (including working memory latent variable only)	123.92	5	0.232	0.177	0.88	0.56	0.63	133.92
Impurity model (including working memory and impurity latent variables)	81.08	4	0.209	0.151	0.92	0.60	0.73	93.08
Ceiling effect model (including additionally weights for first indicator showing ceiling effect)	28.64	4	0.118	0.083	0.97	0.86	0.91	40.64
Simplicity model (including additionally a representation of simplicity strategy)	16.58	3	0.101	0.052	0.98	0.91	0.96	30.58
Simplicity model plus without impurity latent variable	20.55	4	0.097	0.063	0.98	0.92	0.94	32.55

2.2.3 Results for the Third Dataset

The investigation of the internet data by means of the modified Exchange Test yielded the fit statistics reported in Table 3. The fit results for the working memory model reported in the first row were far from indicating a good model fit. Including the impurity latent variable improved the model fit to a considerable degree and also the inclusion of the weights. However, the simplicity strategy latent variable was additionally necessary for achieving a good model fit that was an excellent fit of the model to the data. All fit statistics of this model were well below respectively above the corresponding cut-offs, and the AIC reached its lowest value. Furthermore, since the chi-square was close to the expected value, even the normed chi-square defined as ratio of chi-square and degreed of freedom was good.

In sum, in each dataset the simplicity model provided the best account of the data that were investigated. Furthermore, in each dataset the succession of models was associated with successive decreases of the chi-square. After scaling the variances of the latent variables under the assumption of average factor loadings of 0.3 (Schweizer, 2011) the working memory latent variable accounted for 45.3 percent of the variance at the latent level in the first dataset, for 46.3 percent in the second dataset and for 32.0 percent in the third dataset. The smallest percentage in the third dataset indicated that internet data included more systematic but irrelevant variance as data collected under standardized conditions.

Table 3. Fit Results Observed in Investigating Exchange Test Data – Modified Version (4 indicators) and Free Internet Condition by means of Models Considering Working Memory, Impurity, Strategy Use and Weights (N = 378)

Model	χ^2	df	RMSEA	SRMR	GFI	NNFI	CFI	AIC
Working memory model (including working memory latent variable only)	121.49	5	0.249	0.191	0.86	0.63	0.69	131.49
Impurity model (including working memory and impurity latent variables)	79.44	4	0.224	0.166	0.91	0.66	0.78	99.44
Ceiling effect model (including additionally weights for first indicator showing ceiling effect)	26.82	4	0.123	0.083	0.97	0.90	0.94	38.82
Simplicity model (including additionally a representation of simplicity strategy)	3.65	3	0.024	0.036	1.00	1.00	1.00	17.65

2.2.4 Results Regarding Measurement Invariance

In order to be able to conduct multiple group analysis, the indicator of the fifth treatment level was eliminated from the models according to the original Exchange Test. The fit results obtained in invariance analysis are provided in Table 4. The first to seventh rows give the results for the models assuming construct-focused or full configural invariance: (1) the model assuming construct-focused configural invariance with respect to the working memory latent variable of the first row; (2) the model assuming construct-focused configural invariance with respect to the working memory and impurity latent variables of the second row; (3) the model assuming construct-focused configural invariance with respect to the working memory and impurity latent variables of the third row whereat the latter was restricted to the second and third dataset; (4) the model assuming construct-focused configural invariance with respect to the working memory and simplicity latent variables of the fourth row; (5) the model assuming construct-focused configural invariance with respect to the working memory and simplicity latent variables of the fifth row whereat the latter was restricted to the second and third dataset; (6) the model assuming full configural invariance of the sixth row; (7) the model assuming full configural invariance that was, however, only partial with respect to the impurity and simplicity latent variables of the seventh row. All SRMR, GFI, NNFI and CFI statistics reported for these models indicated a good model fit. Only the RMSEA statistics did not reach the cut-offs but signified an acceptable degree of model fit. The model assuming construct-focused configural invariance with respect to the working memory and simplicity latent variables (4) showed the best fit according to the combination of CFI and AIC.

The investigation of metric invariance required the comparison of the models with equality constraints with the model that showed no equality constraint. The fit results for the model with no equality constraint are included in the last row of Table 4. All comparisons yielded insignificant results: (1) model assuming construct-focused configural invariance with respect to the working memory latent variable [$\chi^2_{critical}$ (2) = 0.25, ns]; (2) model assuming construct-focused configural invariance with respect to the working memory and impurity latent variables [$\chi^2_{critical}$ (4) = 8.85, ns]; (3) model assuming construct-focused configural invariance with respect to the working memory and impurity latent variables whereat the latter was restricted to the second and third dataset [$\chi^2_{critical}$ (3) = 2.07, ns]; (4) model assuming construct-focused configural invariance with respect to the working memory and simplicity latent variables [$\chi^2_{critical}$ (4) = 1.25, ns]; (5) model assuming construct-focused configural invariance with respect to the working memory and simplicity latent variables whereat the latter was restricted to the second and third dataset [$\chi^2_{critical}$ (3) = 0.38, ns]; (6) model assuming full configural invariance [$\chi^2_{critical}$ (6) = 12.22, ns]; (7) the model assuming full configural invariance that was, however, only partial with respect to the impurity and simplicity latent variables [$\chi^2_{critical}$ (4) = 2.79, ns].

Table 4. Fit Results Observed in Investigating Restricted Configural Invariance of Exchange Test Data by means of Models Considering Working Memory, Impurity, Strategy Use and Weights (N = 442, 442, 378)

Equality constraints of	χ^2	df	RMSEA	SRMR	GFI	NNFI	CFI	AIC
Working memory	37.22	11	0.075	0.028	1.00	0.96	0.98	75.22
Working memory and impurity	45.82	13	0.078	0.059	0.99	0.96	0.97	79.82
Working memory and restricted (2+3)[1] impurity	39.04	12	0.073	0.039	0.99	0.96	0.97	75.04
Working memory and simplicity strategy	38.01	13	0.068	0.029	1.00	0.97	0.98	72.01
Working memory and restricted (2+3)[1] simplicity strategy	37.35	12	0.071	0.027	1.00	0.96	0.98	73.35
All latent variables constrained	49.19	15	0.074	0.031	0.99	0.96	0.97	79.19
Working memory and restricted (2+3)[1] in other all latent variables	39.76	13	0.070	0.045	0.99	0.96	0.97	73.76
No equality constraints	36.97	9	0.086	0.027	1.00	0.95	0.97	78.97

[1] The equality constraint is restricted to the second and third dataset.

3. Discussion

Measurement invariance is an important concept that is especially useful for extending the structural validity of a measure to a set of samples or groups whereat the groups may refer to different populations. It has been developed on the basis of the congeneric model of measurement. However, the focus on the congeneric model of measurement prevents the application to measures that require multi-dimensional models of measurement with constrained discriminability to achieve the statistical control of sources giving rise to irrelevant systematic variance. The extension of the concept of measurement invariance to such models requires some modifications. These modifications must bring about the concentration of invariance analysis on the essential part of a model of measurement. It means the replacement of full configural invariance by construct-focused configural invariance. In a way this replacement means the return to the original aim of invariance analysis since invariance analysis on the basis of the congeneric model is always intended to concentrate on the essential part of a measure.

In the application to Exchange Test data the usefulness of the extension of the concept of measurement invariance to multi-dimensional models with constrained discriminability and statistical control of irrelevant systematic variance is demonstrated. The results even seem to indicate that the Exchange Test data show full configural and metric invariance. However, in evaluating the results it must be taken into consideration that there has been a change of the representation of the simplicity strategy in modeling the data obtained by the modified version of Exchange Test as compared to the original version. The change from the original to modified versions is not expected to cause an alteration of the representation of working memory capacity by means of Exchange Test. But the switch from the original to modified version may exert an influence on the selection of alternative strategies in completing the items if participants deviate from the prescribed strategy. Furthermore, there were indications that the auxiliary processes contributed differently to performance in the measurement with the original and modified versions of Exchange Test. These observations give reason for calling full configural and metric invariance into question. In contrast, there is reason to expect construct-focused invariances, and the results provide evidence in favor of the expectation.

References

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

DiStefano, C. (2016). Examining fit with structural equation models. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction. Standards and recent advances* (pp. 166-193). Göttingen, Berne: Hogrefe.

Fan, W., & Hancock, G. R. (2012). Robust means modeling: an alternative for hypothesis testing of independent means under variance heterogeneity and nonnormality. *Journal of Educational and Behavioral Statistics, 37*, 137-156. http://dx.doi.org/10.3102/1076998610396897

French, B. F., & Finch, W. H. (2016). Detecting differential item functioning. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction. Standards and recent advances* (pp. 197-217). Göttingen, Berne: Hogrefe.

Geiser, C., Lehmann, W., & Eid, M. (2006). Separating "Rotators" From "Nonrotators" in the Mental Rotations Test: A Multigroup Latent Class Analysis. *Multivariate Behavioral Research, 41*, 261-293. http://dx.doi.org/10.1207/s15327906mbr4103_2

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66*, 930-944. http://dx.doi.org/10.1177/0013164406288165

Hessling, R. M., Traxel, N. M., & Schmidt, T. J. (2004). Ceiling effect. In Lewis-Beck, M. S., A. Bryman, & Liao, T. F. (Eds.), *The SAGE encyclopedia of social science research methods* (p. 107). Thousand Oaks: Sage.

Jensen, A. R. (1982). The chronometry of intelligence. In R. J. Sternberg (Ed.), *Advances in research on intelligence*, Vol. 1 (pp. 93132). Hillsdale, NJ: Erlbaum.

Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika, 57*, 239-257. http://dx.doi.org/10.2307/2334833

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133. http://dx.doi.org/10.1007/BF02291393

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80*. Lincolnwood, IL: Scientific Software International Inc.

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment, 11*, 59-76.

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

Lucke, J. F. (2005). The α and ω of congeneric test theory: an extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement, 29*, 65–81. http://dx.doi.org/10.1177/ 0146621604270882

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics, 16*, 163-200. http://dx.doi.org/10.1007/BF01065485

McArdle, J. J. (1989). Structural modeling experiments using multiple growth functions. In P. Ackerman, R. Kanfer, & R. Cudeck (Eds.), *Learning and individual differences: ability, motivation and methodology* (pp. 71-117). Hillsdale, NJ: Erlbaum.

McArdle, J. J. (2009). Latent variable modeling of differences and change with longitudinal data. *Annual Review of Psychology, 60*, 577-605. http://dx.doi.org/10.1146/annurev.psych.60.110707.163612

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

http://dx.doi.org/ 10.1007/BF02294825

Miyake, A., Friedman, N. P., Emerson, M. J.,Witzki, A. H., Howerter, A., & Wager, T. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100. http://dx.doi.org/10.1006/cogp.1999.0734

Patsenko, E. G., & Altman, E. M. (2010). How planful is routine behavior? A selective-attention model of performance in the Tower of Hanoi. *Journal of Experimental Psychology: General, 139*, 95-116. http://dx.doi.org/10.1037/a0018268

Perret, P., Bailleux, C., & Dauvier, B. (2011). The influence of relational complexity and strategy selection on children's reasoning in the Latin Square Task. *Cognitive Development, 26*, 127-141. http://dx.doi.org/10.1016/j.cogdev.2010.12.003

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Salnaitis, C. L., Baker, C. A., Holland, J., & Welsh, M. (2011). Differentiating Tower of Hanoi performance: Interactive effects of psychopathic tendencies, impulsive response styles, and modality. *Applied Neuropsychology, 18*, 37-46. http://dx.doi.org/10.1080/09084282.2010.523381

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Aufgaben [Learning and solving complex cognitive problems]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19*, 476-505.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222. http://dx.doi.org/10.1016/j.hrmr.2008.03.003

Schreiner, M., Altmeyer, M., & Schweizer, K. (2012). The Web Version of Exchange Test: Description and Psychometric Properties. *European Journal of Psychological Assessment, 28*, 181-189. http://dx.doi.org/10.1027/1015-5759/a000125

Schweizer, K. (1996). The speed-accuracy transition due to task complexity. *Intelligence, 22,* 115-128. http://dx.doi.org/10.1016/S0160-2896(96)90012-4

Schweizer, K. (2006a). The fixed-links model in combination with the polynomial function as tool for investigating choice reaction time data. *Structural Equation Modeling, 13*, 403-419. http://dx.doi.org/10.1207/s15328007sem1303_4

Schweizer, K. (2006b). The fixed-links model for investigating the effects of general and specific processes on intelligence. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2,* 149-160. http://dx.doi.org/10.1027/1614-2241.2.4.149

Schweizer, K. (2007). Investigating the relationship of working memory tasks and fluid intelligence tests by means of the fixed-links model in considering the impurity problem. *Intelligence, 35*, 591-604. http://dx.doi.org/10.1016/j.intell.2006.11.004

Schweizer, K. (2008). Investigating experimental effects within the framework of structural equation modeling: an example with effects on both error scores and reaction times. *Structural Equation Modeling, 15*, 327-345. http://dx.doi.org/10.1080/10705510801922621

Schweizer, K. (2009). Fixed-links models for investigating experimental effects combined with processing strategies in repeated measures designs: a cognitive task as example. *British Journal of Mathematical and Statistical Psychology, 62*, 217-232. http://dx.doi.org/10.1348/000711007X268558

Schweizer, K. (2011). Scaling variances of latent variables by standardizing loadings: applications to working memory and the position effect. *Multivariate Behavioral Research, 46*, 938-955. http://dx.doi.org/10.1080/00273171.2011.625312

Schweizer, K. (2012). A weighted version of the tau-equivalent model of measurement for items with ordered response categories. *International Journal of Statistics and Probability, 1*, 151-163. http://dx.doi.org/10.5539/ ijsp.v1n2p151

Schweizer, K. (2016). A confirmatory factor model for the investigation of cognitive data showing a ceiling effect: an example. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. Douglas & M. Wiberg, (Eds.), *Quantitative Psychology Research*. Heidelberg: Springer.

Schweizer, K., Altmeyer, M., Ren, X., & Schreiner' M. (2015). Models for the detection of deviations from the expected processing strategy in completing the items of cognitive measures. *Multivariate Behavioral Research, 50*, 544-554. http://dx.doi.org/10.1080/00273171.2015.1032397

Schweizer, K., Ren, X., Wang, T., & Zeller, F. (2015). Does the constraint of factor loadings impair model fit and accuracy in parameter estimation? *International Journal of Statistics and Probability, 4*, 40-50. http://dx.doi.org/10.5539/ijsp.v4n4p40

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370. http://dx.doi.org/10.1111/ j.1745-3984.1990.tb00754.x

Thompson, M. S. (2016). Assessing measurement invariance of scales using multiple-group structural equation modeling. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction. Standards and recent advances* (pp. 166–193). Göttingen, Berne: Hogrefe.

Van Zomeren, A. H., & Brouwer, W. H. (1994). Clinical neuropsychology of attention. New York, NY: Oxford University Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organisational research. *Organisational Research Methods, 16*, 4-70. http://dx.doi.org/10.1177/109442810031002

Vogt, W. P. (2005). *Dictionary of statistics & methodology: a nontechnical guide for the social sciences* (3rd ed.). Thousand Oaks: Sage.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: problems and remedies. In R. H. Hoyle. (ed.), *Structural equation modeling: concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological intruments: applications in the substance domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Copyrights