Erschienen in: British Journal of Mathematical and Statistical Psychology ; 59 (2006), 1. - S. 195-213
https://doi.org/10.1348/000711005X67490

Analysing multitrait–multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample size is needed for valid results?

Fridtjof W. Nussbeck*, Michael Eid and Tanja Lischetzke
University of Geneva, Switzerland

Convergent and discriminant validity of psychological constructs can best be examined in the framework of multitrait–multimethod (MTMM) analysis. To gain information at the level of single items, MTMM models for categorical variables have to be applied. The CTC(M 1) model is presented as an example of an MTMM model for ordinal variables. Based on an empirical application of the CTC(M 1) model, a complex simulation study was conducted to examine the sample size requirements of the robust weighted least squares mean and variance adjusted χ^2 test of model fit (WLSMV estimator) implemented in Mplus. In particular, the simulation study analysed the χ^2 approximation, the parameter estimation bias, the standard error bias, and the reliability of the WLSMV estimator depending on the varying number of items per trait–method unit (ranging from 2 to 8) and varying sample sizes (250, 500, 750, and 1000 observations). The results showed that the WLSMV estimator provided a good – albeit slightly liberal – χ^2 approximation and stable and reliable parameter estimates for models of reasonable complexity (2–4 items) and small sample sizes (at least 250 observations). When more complex models with 5 or more items were analysed, larger sample sizes of at least 500 observations were needed. The most complex model with 9 trait–method units and 8 items (72 observed variables) requires sample sizes of at least 1000 observations.

1. Introduction

Determination of convergent and discriminant validity in present-day psychological research generally follows the multitrait–multimethod (MTMM) framework suggested by Campbell and Fiske (1959). Within this framework, convergent validity is usually confirmed by the convergence of different methods measuring the same trait. Discriminant validity is proven by a low correlation between different constructs. Over the past several years, structural equation modelling (SEM) has become one of the most

* *Correspondence should be addressed to Fridtjof W. Nussbeck, Faculty of Psychology and Educational Sciences, University of Geneva, 40 Boulevard du Pont d'Arve, CH-1211 Geneva 4, Switzerland (e-mail: fridtjof.nussbeck@pse.unige.ch).*

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-2-cssjfgu5kh5u4

important methodological tools for analysing MTMM data (e.g. Dumenci, 2000; Eid, Lischetzke, & Nussbeck, 2006; Marsh, 1989, 1993; Marsh & Grayson, 1995; Marsh & Hocevar, 1988; Widaman, 1985). In particular, SEM allows the separation of measurement error from method-specific effects and a test of the assumptions concerning the nature of trait and method influences. Applications of SEM to MTMM data typically refer to the analysis of metric outcome variables because traditional estimation and testing methods require metric response variables. However, psychological constructs are often measured by items with ordinal character, for instance, items with Likert-type response formats. Thus, to apply SEM to MTMM data, researchers typically aggregate their ordinal variables to test halves or test parcels, which are then assumed to be metric. Although this is a reasonable way to test trait- and method-specific influences, an unwanted side effect is the loss of information about convergent and discriminant validity at the level of single items, information that would be very useful for test construction. When establishing a new questionnaire, for example, a researcher might select only those items with high convergent and high discriminant validity coefficients.

It is well known that ordinal variables usually violate the distributional as well as the linear assumptions of SEM for metric response variables, in particular if they have only a small number of categories and if the variables are not symmetrically distributed. Yet, maximum likelihood (ML) estimation methods are often applied to the covariance matrix of ordinal variables; however, this strategy can cause two major problems (Schermelleh-Engel, Moosbrugger, & Müller, 2003): first, the ML parameter estimates – although consistent – are no longer efficient, and second, the type I error rate for model rejection increases. Such problems have led to the development of new methods for estimating and testing SEM models for ordinal variables such as the *weighted least squares* (WLS) estimator for analysing the polychoric covariance/correlation matrix (Bollen, 1989; Browne, 1984; Muthén, du Toit, & Spisic, in press; Satorra, 1989, 1992). However, because the weight matrix of this estimator increases rapidly with the number of indicator variables, very large sample sizes are necessary to obtain valid results. Consequently, Chou and Bentler (1995) recommend not using the WLS estimator when models are complex and the sample size is small.

As a result of this rather severe sample size requirement and the computational demands of the WLS estimator, Muthén *et al.* (in press) developed the robust WLS mean- and variance-adjusted χ^2 test of model fit (WLSMV estimator) as a less demanding alternative. In a first simulation study, they found that this estimator showed good statistical properties in terms of parameter estimation and testing the model fit, even when relatively small sample sizes ($N = 200$) and moderately complex models (10 parameters to be estimated; 3 latent variables and 12 indicators) were applied. Muthén *et al*. concluded that 'given the generality, statistical performance, and relative computational speed of this new approach, it provides a useful practical method for latent variable analysis with large models involving categorical outcomes'. Recently, Flora and Curran (2004) ran a simulation study replicating the results of Muthén *et al.* comparing the WLSMV estimator to the WLS estimator in more complex models of confirmatory factor analysis (2 latent variables and 5–20 indicators). The WLSMV estimator outperformed the WLS estimator in terms of its χ^2 approximation and smaller estimation biases. Hence, the WLSMV estimator seems to be promising for the analysis of more complex structural equation models for ordinal variables. However, we do not know whether this estimator is also appropriate when more complex models such as MTMM models are under consideration. In particular, open questions remain regarding

the condition under which the WLSMV-based goodness-of-fit test follows the theoretically expected χ^2 distribution, whether there is a bias in the parameter estimates, and whether the standard errors are appropriately estimated when more complex models such as MTMM models are analysed.

With the aim of pursuing this research question, we conducted a simulation study based on an application of an MTMM model to ordinal variables. Before the results of this simulation study are presented we will briefly describe how SEM can be applied to an MTMM analysis of ordinal variables by referring to an empirical application of the correlated trait–correlated method minus 1 (CTC(M 1)) model (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003).

2. SEM of MTMM models for ordinal variables

Previously developed structural equation MTMM models can easily be adapted to ordinal response variables. In MTMM models for metric variables, the observed variables are decomposed into trait, method, and error variables. In general, two classes of models can be distinguished, models in which there is only one observed indicator for each trait–method unit and models in which there are multiple indicators for each trait–method unit. The advantages of multiple indicator models include the more accurate separation of unsystematic measurement error from systematic method influences and the analysis of trait-specific method effects (Eid *et al.*, 2003). After considering the advantages offered by the various models, we selected a multiple indicator model as the basis of the simulation studies, a model which is also more complex than a single indicator model and, therefore, more appropriate for scrutinizing the appropriateness of the WLSMV estimator.

In general, three types of model can be distinguished that are appropriate for different modelling circumstances. The first and simplest model assumes that there is one latent variable for all indicators belonging to the same trait–method unit. The correlations between these latent variables represent the MTMM correlations of Campbell and Fiske's (1959) classical MTMM matrix, but, in contrast to Campbell and Fiske's approach, they are free of measurement error. In a second, more restricted type of model there is one second-order trait factor for all first-order factors representing the same trait but measured by different methods. This common trait factor represents the convergence of the different methods. Latent first-order residuals indicate method effects. In addition, it could be assumed that all first-order factors belonging to one method but different traits indicate a second-order method factor. Marsh and Byrne (1993), Marsh and Grayson (1995), and Marsh and Hocevar (1988) have developed and discussed MTMM models that belong to this class. These models are particularly appropriate if one is interested in a common latent trait variable. The third type of model is based on the idea of contrasting method and trait variables. The CTC(M 1) model developed by Eid *et al.* (2003) is an example of this type. In this model, which will be explained in detail below, the trait factor is defined as the true score of a standard method. This trait factor is used to predict the trait score measured by the other methods. Method-specific influences for non-standard methods are represented by the deviations of the (latent) trait values measured by a non-standard method from the values predicted by the trait variable of the standard method. Eid *et al.* (2006) give an overview of these different types of MTMM models and discuss the areas of application for which they are important.

All these models can easily be adapted to ordinal variables simply by replacing the metric observed variables by metric latent response variables that are assumed to underlie the observed ordinal variables. If an observed ordinal response variable is denoted by Y_{ijk}, where i is the index for the indicator (of the same trait–method unit), j represents the trait and k the method, and if Y^*_{ijk} is the latent continuous variable underlying each observed variable Y_{ijk}, then SEM for ordinal variables is based on the following measurement structure (Muthén, 1983; Takane & De Leeuw, 1987):

$$Y_{ijk} = \begin{cases} c_{ijk} \quad 1, \text{ for } \kappa_{(c_{ijk}\ \ 1)ijk} < Y^*_{ijk} \\ s, \qquad \text{for } \kappa_{sijk} < Y^*_{ijk} \leq \kappa_{(s+1)ijk}, \\ 0, \qquad \text{for } Y^*_{ijk} \leq \kappa_{1ijk} \end{cases}$$

with $s \in \{0, \ldots, c_{ijk} \quad 1\}$ indicating the observed categories. According to this model, the continuous latent variable Y^*_{ijk} is divided into c_{ijk} sections (observed categories) by the threshold parameters κ_{sijk}. The manifest variable Y_{ijk} equals s if a value of the latent variable Y^*_{ijk} lies between κ_{sijk} and $\kappa_{(s+1)ijk}$. If Y^*_{ijk} is smaller than the smallest threshold κ_{1ijk}, the observed value of Y_{ijk} will be 0. If Y^*_{ijk} is larger than the largest threshold the $\kappa_{(c_{ijk}\ \ 1)ijk}$, the manifest outcome Y_{ijk} will be $c_{ijk} \quad 1$. The underlying metric variable Y^*_{ijk} can now be decomposed into trait, method, and error variables according to the MTMM models for metrical variables.

Because simulation studies are extremely time-consuming we were unable to simulate data according to all available models. Instead, we had to choose one model that seemed to be appropriate for our data structure. Because we had structurally different raters (self-report, two peer reports) we chose the CTC(M 1) model which allows us to contrast the different methods (raters). However, as the other types of models are of comparable complexity we think that the results of this simulation study can also be transferred to other MTMM models for ordinal variables. The CTC(M 1) model will be briefly sketched in order to assist the reader in understanding the results of the study.

3. The CTC(M−l) model for ordinal variables

The CTC(M l) model was introduced in detail by Eid (2000) and extended to the analysis of multiple indicators by Eid *et al.* (2003). CT stands for *correlated trait* as all trait factors can be correlated in the model, C(M 1) stands for *correlated method* as all method factors can be correlated, and the ' 1' indicates that there are no method factors for the one method that is taken as the comparison standard. Before going any further, we will show how the confirmatory factor-analytic (CFA) CTC(M 1) model for ordinal variables can be formulated referring to the empirical application on which the simulation study is based. This empirical application comprises three *traits* representing the habitual mood level assessed by the trait version of the Multidimensional Mood Questionnaire (MMQ: Steyer, Schwenkmezger, Notz, & Eid, 1994): (1) pleasant–unpleasant; (2) awake–sleepy; (3) calm–restless. All three traits are measured by three different *methods*, namely, a self-report (SR) and two peer reports (PR A and PR B). Five hundred targets (SR) were required to bring two peer raters with them at a designated time for completing the questionnaires. The peers were arbitrarily assigned as peer A or peer B. Thus, each group (of three individuals) participated at the same time, although we separated them while they filled out the questionnaires to prevent the sharing of information. We also informed the participants that they would not be able to view the

other questionnaires from their group to prevent possible data manipulation (e.g. impression management, socially desirable response behaviour). All self-raters were students at the University of Trier and the Trier University of Applied Sciences (Germany).

The CTC(M 1) model for ordinal variables with two items per trait–method unit is depicted in Figure 1. This is equivalent to the figure presented by Eid *et al.* (2005) with the exception that the observed variables Y_{ijk} are replaced by the underlying variables Y^*_{ijk}.

First, we only consider the structure of the model for the first trait measuring the pleasant–unpleasant dimension (upper third of Figure 1). To apply the model, one method has to be chosen as the standard method. In this example the self-report (SR) has been chosen as the standard method (see the box at the top of Figure 1). The two items ('great' and 'happy') are indicators of a latent variable that represents the (error-free) trait of feeling pleasant or unpleasant by the self-report. As a consequence, the underlying variables belonging to the standard method of the first trait can be decomposed into a latent trait variable T_1 and a measurement error variable E_{ijk}:

$$Y^*_{i11} = \lambda_{\mathrm{T}i11} T_1 + E_{i11}.$$

The loading parameter of the first indicator is set to 1 for identification reasons. The factor loading of the second indicator can vary freely.

The same two items have been administered to the two peer raters (PR A and PR B). The latent trait variable of the self-ratings is used to predict the peer ratings. Deviations of the peer ratings from the values predicted by the self-ratings are represented by a method factor for each rater. A method factor M_{1k} captures systematic deviations of a peer report from the expected latent score based on the self-report for the first trait. For both peer reports the structural equations have to be extended to:

$$Y^*_{i1k} = \lambda_{\mathrm{T}i1k} T_1 + \lambda_{\mathrm{M}i1k} M_{1k} + E_{i1k},$$

where λ_{Mijk} is the factor loading on the method factor. The loading parameter of the first item ('great') on the method-specific factor is set to 1.

The same structure is applied to the second and third traits (awake–sleepy and calm–restless). It should be stressed here that the latent method-specific factors are conceived to be trait-specific, meaning that method-specific influences do not have to generalize perfectly across traits.

The linear decomposition of the underlying variables allows the definition of the consistency and method-specificity coefficients. The consistency coefficient represents that part of the true (error-free) variance of the underlying variable determined by the latent trait factor, which corresponds to the conceptualization of convergent validity as monotrait–heteromethod convergence according to Campbell and Fiske (1959):

$$CO(Y^*_{ijk}) = \frac{\lambda^2_{\mathrm{T}ijk}\mathrm{Var}(T_j)}{\lambda^2_{\mathrm{T}ijk}\mathrm{Var}(T_j) + \lambda^2_{\mathrm{M}ijk}\mathrm{Var}(M_{jk})}.$$

For the items of the standard method (the self-report), this value equals 1 because there is no method effect. The method specificity is the proportion of true variance of an indicator belonging to a non-standard method that is determined by the method factor:

$$MS(Y^*_{ijk}) = \frac{\lambda^2_{\mathrm{M}ijk}\mathrm{Var}(M_{jk})}{\lambda^2_{\mathrm{T}ijk}\mathrm{Var}(T_j) + \lambda^2_{\mathrm{M}ijk}\mathrm{Var}(M_{jk})}.$$

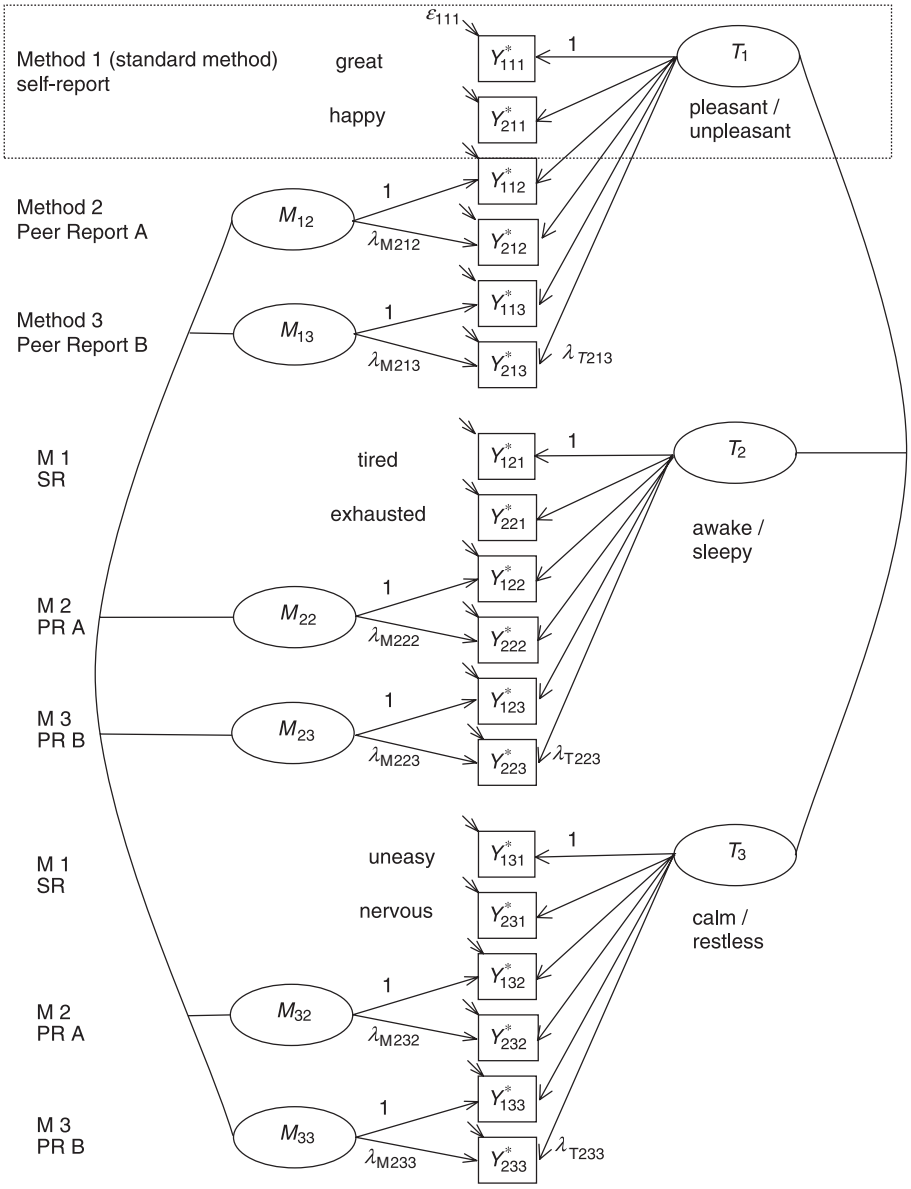

Figure 1. The multiple indicator CTC(M 1) model for ordinal variables of the empirical application. SR, self report; PR A, peer report A; PR B, peer report B; T_j, latent trait variable; M_{jk}, trait specific method factor; Y^*_{ijk}, latent item specific variable; ε_{ijk}, residual; λ_{Mijk}, λ_{Tijk}, factor loadings. Factor loadings are only depicted for one path for each factor but they are estimated for all variables. The loading of the first indicator is set to unity for all factors.

On the basis of these two coefficients, researchers can select items that are appropriate for different research questions. For example, in personality research one is usually interested in a high consistency (convergent validity) and a low method specificity of different ratings to ensure high convergent validity. Moreover, the idea of having several

correlation coefficients to determine the discriminant validity and the homogeneity of method influences is appealing. First, the correlations of the different trait variables represent the discriminant validity between the different traits (correlations on the right-hand side of Figure 1). These correlations should be quite low to justify the existence of different constructs. Second, researchers may be interested in the correlation between the different trait-specific method factors in order to determine the generalizability of method effects across traits (correlations between M_{12}, M_{22}, and M_{32}, and correlations between M_{13}, M_{23}, and M_{33} on the left-hand side of Figure 1). Empirical applications of the CTC(M 1) model (see Eid *et al.*, 2003; Trierweiler, Eid, & Lischetzke, 2002) show that these coefficients are typically far from 1, indicating only a small degree of generalizability. Third, the correlations between the different method factors belonging to the same trait (correlations between the method factors of peers A and B) are interesting. These correlations show whether the two peers share a common view of the target person which deviates from the target's own view. Fourth, for the two peers one could consider the correlations between the method factors of different traits in order to explore whether the overestimation of the first trait by peer A is associated with an overestimation of the second trait by peer B (e.g. the correlation between M_{12} and M_{23} shows whether the overestimation of pleasantness by A goes together with the overestimation of being awake by B). Finally, the model also allows the correlation of a method factor belonging to one trait and the trait factor of another trait (Eid *et al.*, 2003). This correlation is a measure of the discriminant validity corrected for method influences because a method factor is that part of a peer report that does not share anything with the self-report of that trait, whereas the trait factor is a pure self-report factor.

3.1. Identification of the model

Eid *et al.* (2003) showed that, in general, the multiple indicator CFA CTC(M 1) model is identified with three indicators for each trait–method unit, two traits, and three methods. With fewer than three indicators the models still work well if the following requirements are fulfilled: (1) one method factor has to be substantially correlated with at least one other method factor; and (2) one trait factor has to be substantially correlated with another trait, or the loadings of the non-standard methods on the trait factors have to be substantial. If requirement (1) is not fulfilled, the loading parameters belonging to the same method factors must be set equal to each other. If requirement (2) is not fulfilled, the loading parameters of the indicators of the standard method on the trait factors must be set to equal each other (for purposes of identification). Moreover, because there are no method factors for the standard method, the loading matrix is not rank-deficient. Therefore, the identification problems due to rank-deficient loading matrices in MTMM models (Grayson & Marsh, 1994) do not occur.

Because the factor-analytic version of the CTC(M 1) model for ordinal variables has the same structure as the CFA CTC(M 1) model for metric response variables, the conditions for identification are the same. It is additionally assumed that the distributions of the underlying variables Y^*_{ijk} follow the standard normal distribution.

4. Application of the CTC(M−1) model to empirical data

To illustrate the meaning of the CTC(M 1) model and to show its applicability, we report the results of the empirical application depicted in Figure 1. From the 500 groups participating in the study, 474 complete data sets (1422 participants) were suitable for analysis.

The model was analysed with the computer program Mplus (Muthén & Muthén, 2004) using the WLSMV estimator. It fits the data very well ($\chi^2(59, N = 474) = 70.57$, $p = .14$, CFI 1.00, RMSEA .02). Table 1 presents the standardized loading parameters of the underlying variables on the trait factor and on the method factors. The trait loadings range from $\lambda_{T122} = .14$ (trait loading of 'tired', peer A) to $\lambda_{Tij1} = .86$ (trait loadings of 'great' and 'uneasy', self), whereas the method factor loadings range from $\lambda_{M122} = .67$ (method loading of 'tired', peer A) to $\lambda_{M222} = .85$ (method loading 'exhausted', peer A).

The variances and correlations are reported in Table 2. The correlations have the following interpretations:

(1) Correlations among trait factors indicate discriminant validity at the level of the standard method. In the application presented here, discriminant validity is

Table 1. Standardized loading parameters of the CTC(M 1) model with two indicators and 474 observations

Item	Trait loading	Method factor A	Method factor B
Pleasant–unpleasant			
Self report			
Great	.86		
Happy	.82		
Peer report A			
Great	.29	.84	
Happy	.27	.73	
Peer report B			
Great	.19		.83
Happy	.27		.72
Awake–sleepy			
Self report			
Tired	.70		
Exhausted	.78		
Peer Report A			
Tired	.14	.67	
Exhausted	.19	.85	
Peer Report B			
Tired	.19		.68
Exhausted	.25		.70
Calm–restless			
Self report			
Uneasy	.86		
Nervous	.73		
Peer report A			
Uneasy	.21	.73	
Nervous	.20	.73	
Peer report B			
Uneasy	.29		.73
Nervous	.22		.70

Note. Blank cells indicate factor loadings fixed to zero by definition of the model. The original German items can be found in the Appendix.

Table 2. Variances (on the main diagonal) and correlations of the trait and method factors in the CTC(M 1) model with two indicators and 474 observations

	T_1	T_2	T_3	M_{112}	M_{113}	M_{122}	M_{123}	M_{132}	M_{133}
T_1	**.74**								
T_2	**.56**	**.48**							
T_3	**.49**	**.46**	**.73**						
M_{112}		.00	.02	**.71**					
M_{113}		.02	.10	**.37**	**.69**				
M_{122}	.01		.02	**.35**	.13	**.45**			
M_{123}	.11		.08	.12	**.46**	**.15**	**.46**		
M_{132}	.04	.00		**.45**	**.16**	**.33**	.06	**.53**	
M_{133}	.01	.03		.07	**.38**	.10	**.36**	**.20**	**.54**

Note. Parameters with absolute *t* values larger than 2 appear in bold. Empty cells indicate non admissible correlations that are set to 0. M_{ijk} denotes the trait specific method effect of peer A or peer B.

measured with respect to the self-ratings. High discriminant validity is given when these correlations are low, as is the case in the present application (all correlations are below $r = .56$).

(2) Correlations of method factors belonging to the same method but different traits represent the degree to which the method effects generalize across traits. In the present application, the largest correlation between the method factors belonging to the same method is $r = .45$. Thus the assumption of trait-specific method factors is empirically well founded.

(3) The correlation of a trait-specific method factor and the trait factor of another trait is a measure of the discriminant validity corrected for method influences. For example, the correlation of the method effect for peer A's judgement of the awake–sleepy dimension with the trait effect for the same peer's judgement of the calm–restless dimension is $r = .01$; thus the under- or overestimation of the awake–sleepy dimension by peer A does not depend on the trait score of the calm–restless dimension. All other correlations do not exceed $r = .11$; hence, no method effect can be significantly related to another trait in this application.

(4) The correlations between method factors of different methods indicate between-method associations that cannot be explained by the standard method. If the correlation between two method factors belonging to two different methods but one trait is significant, the non-standard methods have more in common with each other than can be explained by the standard method. Both non-standard methods differ in a similar manner from the standard method. In the first application, for example, both peer raters tend to over- or underestimate the trait values of the pleasant–unpleasant dimension ($r_{M_{112,113}} = .37$). Hence, both peers share a common view of the pleasant–unpleasant dimension of the self-rater which differs from the target's own view.

Table 3 presents the variance components. The consistency coefficients (*CO*) of the peer ratings range from .04 to .14, whereas the method-specificity coefficients (*MS*) range from .86 to .96. Thus, the major part of variance of the peer ratings is not shared with the self-ratings. As the consistency coefficient is a determination coefficient, the correlation

Table 3. Variance components in the CTC(M 1) model with two indicators and 474 observations

Item	Consistency	Method specificity	Latent correlation[a]
Pleasant–unpleasant dimension			
Self report			
Great	1.00		
Happy	1.00		
Peer report A			
Great	.10	.90	.32
Happy	.12	.88	.35
Peer report B			
Great	.05	.95	.22
Happy	.12	.88	.35
Awake–sleepy dimension			
Self report			
Tired	1.00		
Exhausted	1.00		
Peer report A			
Tired	.04	.96	.20
Exhausted	.05	.95	.22
Peer report B			
Tired	.07	.93	.26
Exhausted	.12	.88	.35
Calm–restless dimension			
Self report			
Uneasy	1.00		
Nervous	1.00		
Peer report A			
Uneasy	.07	.93	.26
Nervous	.07	.93	.26
Peer report B			
Uneasy	.14	.86	.37
Nervous	.09	.91	.30

[a] Latent correlation with the trait variable $\left(\sqrt{consistency}\right)$.

between the latent variables of the peer reports and the latent variables of the self-report can be computed by taking the square root of the consistency coefficient. These correlations range from $r = .20$ to $r = .37$. Values such as these can be expected for correlations between different ratings of emotion-related traits (McCrae, 1982; Stanton, Kirk, Cameron, & Danoff-Burg, 2000; Watson, Hubbard, & Wiese, 2000). For the items of the self-report (which was chosen as the standard method) there is no method factor; thus, all systematic variance of the self-report items is due to the latent trait variable. Therefore, no consistency or method-specificity coefficients must be computed.

The crucial question, then, as already stated above, is whether the WLSMV estimator (Muthén *et al.*, in press) applied in these examples is valid for applications of this complexity. More explicitly, can we trust its χ^2 approximation, and can we rely on the model parameters (loadings and covariances) and interpret them as we did? The validity of the WLSMV χ^2 test might be questionable because its χ^2 approximation is valid for infinite sample sizes and only approximately valid for large sample sizes. However, it remains an open question whether a sample size of about 500 is large enough for the

complexity of the CTC(M 1) model. To provide an answer regarding the sample size necessary to obtain valid results of the WLSMV χ^2 test statistic, we conducted a Monte Carlo simulation study.

5. Monte Carlo simulation

In order to scrutinize the sample size needed to obtain valid results when applying the WLSMV estimator, a Monte Carlo study (Efron, 1979, 1982) was conducted. In the recent literature there are three major criteria that must be applied when examining whether an estimator is solid and reliable (Flora & Curran, 2004; Muthén *et al.*, in press; Muthén & Muthén, 2002; Schermelleh-Engel *et al.*, 2003):

(1) To use the WLSMV estimator for testing a model appropriately, the test statistic (WLSMV χ^2 test) should follow the expected χ^2 distribution.
(2) There should be no bias in estimating the parameters of the model.
(3) The standard errors should be estimated appropriately to obtain valid results for testing the significance of single parameters of the model.

To compare our results with the results of the simulation study presented by Muthén *et al.* (in press) we adopted the following evaluation criteria (see also Muthén & Muthén, 2002):

(1) The χ^2 goodness-of-fit coefficient provided by Mplus should approximate the χ^2 distribution; χ^2 test rejection proportions at the 5% level should be less than .10.
(2) The parameter estimate biases (*peb*) of loading parameters, variances, and covariances should not exceed .10.
(3) The standard error biases (*seb*) of loading parameters, variances, and covariances should be smaller than .15.
(4) For at least 91% but no more than 98% of the replications, the true value of a parameter should fall within the boundaries of the observed 95% confidence interval (Muthén *et al.*, in press; Muthén & Muthén, 2002).

In Monte Carlo studies, data are usually generated from a population with well-known (hypothesized) parameter values. In our study, the polychoric correlation matrix and the threshold parameters of the empirically applied models were taken as the population correlation matrix and population threshold parameters. Additionally, all other model parameters (loading parameters, variances, and correlations) are considered as the population (true) parameters. We examined sample sizes of 250, 500, 750, and 1000 observations. We chose these modifications because they represent realistic sample sizes in the area of multimethod assessment.

Moreover, we improved the complexity of the models under consideration. As stated above, the MMQ consists of three scales comprising eight items each. Thus, we increased the number of items for each trait–method unit from 2 to 8 and ran 28 Monte Carlo simulations with four different sample sizes (250–1000) consisting of 500 replications each. The structural part of the model always remained the same, whereas the measurement part became more complex. We first analysed a base model using the empirical data we had gathered (474 observations) for each number of items per trait–method unit. Then we used the model-implied polychoric correlation matrix and

the threshold parameters obtained in each of the seven base applications as input for the simulation study. We chose this simulation design to make it most meaningful for the present research context.

As Mplus cannot complete replications if a category has zero frequency, some sparseness problems occurred especially concerning the first category of the observed variables in cases with extremely skewed distributions. In order to keep the number of replications at 500 to obtain comparable results for the WLSMV estimator for each simulation, we decided to collapse the first and second category where necessary.

The Monte Carlo option implemented in Mplus does not provide every single χ^2 value of every replication, but it provides a comparison of the expected probabilities of the χ^2 values with the observed ones[1]. If the observed and expected χ^2 probabilities equal each other approximately, the WLSMV estimator provides an adequate approximation of the χ^2 distribution. If the probabilities do not equal each other completely, at least the critical probabilities ($p = .01$, $p = .05$, and $p = .10$) should equal each other to gain adequate information about the goodness of fit.

To determine the parameter estimate biases as well as the standard error biases, the mean value of the parameter estimates, the mean value of the standard errors, and the standard deviation of the parameter estimates have to be considered. Mplus provides these values. The parameter estimate bias is defined as (Eid, 1995; Muthén *et al.*, in press):

$$peb = \frac{\bar{m}_{\mathrm{P}} \quad e_{\mathrm{P}}}{e_{\mathrm{P}}},$$

where *peb* represents the parameter estimate bias, $\bar{m}_{\mathrm{P}}$ denotes the mean value of the parameter estimates over the replications, and e_{P} displays the theoretically expected parameter value (sometimes referred to as true value) which is exactly the parameter value of the empirical application. The standard error bias is defined quite similarly (Eid, 1995; Muthén *et al.*, in press):

$$seb = \frac{\bar{m}_{\mathrm{SE}} \quad SD_{\mathrm{P}}}{SD_{\mathrm{P}}},$$

where *seb* is the standard error bias, $\bar{m}_{\mathrm{SE}}$ displays the mean value of the standard errors, and SD_{P} represents the standard deviation of the parameter estimates.

Finally, Mplus calculates the proportion of replications for which the theoretically expected parameter falls into the observed 95% confidence interval. We will first present the results of the Monte Carlo study for the application consisting of two items per trait–method unit.

Table 4 shows the comparison of the expected and observed probabilities of the χ^2 values of the first simulation run. The similarity of the expected χ^2 probabilities to the expected ones indicates that the WLSMV estimator provides a good approximation of the χ^2 distribution. For smaller probabilities it tends to slightly underestimate the χ^2 probability, whereas for higher probabilities it seems to slightly overestimate the χ^2 probability. Increasing the number of observations produces smaller deviations of the observed from the expected probabilities. For example, the simulation for 1000

[1] *This procedure is perfectly reasonable because the degrees of freedom of a model are partly estimated by relying on data properties.*

Table 4. Comparison of the expected and observed probabilities of the WLSMV based goodness of fit test for the MMQ models consisting of two items per trait–method unit in the Monte Carlo simulations with 250, 500, 750, and 1000 observations

	Sample sizes			
Expected values	250	500	750	1000
.99	1.00	1.00	1.00	1.00
.98	.99	.99	.99	.98
.95	.97	.96	.96	.95
.90	.95	.91	.90	.90
.80	.86	.83	.80	.79
.70	.75	.72	.72	.70
.50	.52	.51	.53	.49
.30	.25	.29	.31	.30
.20	.15	.19	.19	.18
.10	.06	.09	.09	.08
.05	.02	.02	.04	.04
.02	.01	.01	.02	.01
.01	.00	.01	.01	.01

observations results in a nearly perfect approximation of the χ^2 distribution. Because there are 474 observations (participants) in our application, the performance of the WLSMV estimator for this application can best be evaluated by the simulation with 500 observations. The χ^2 distribution of the simulation shows only very small deviations of the observed from the expected probabilities. For this sample size the WLSMV estimator provides a slightly liberal but acceptable approximation of the χ^2 distribution. Given the χ^2 probability of $p = .14$ in our application, at least 10% of the simulated χ^2 values fall below this value. Regarding the critical expected probability of .05 for simulations consisting of models with 250 and 500 observations, the observed probability is smaller (.02); for 750 and 1000 observations it is very close (.04). Hence, the goodness-of-fit statistic based on the WLSMV estimator rejects a slightly smaller proportion of applications than would be expected by the theoretical χ^2 distribution.

Tables 5 and 6 present the results of the Monte Carlo simulations with respect to the free model parameters. The *true value* is the theoretically expected parameter value (based on the empirical application), the *simulated value* is the mean of all parameter estimates over the simulated data sets, *peb* represents the parameter estimation bias, Est. SD is the standard deviation of the parameter estimates, SE represents the average standard errors of the parameter estimates, SE Dev. is the standard deviation of the standard errors, *seb* is the standard error estimation bias, and Coverage represents the proportion of replications for which the 95% confidence interval includes the true parameter. The results of the simulation study showed very similar values of the evaluation criteria for each type of parameter (loading parameters, variances, and covariances) within each simulation condition. We thus decided to present the mean values of all parameters belonging to the same type of parameter to make Tables 5 and 6 most meaningful and comprehensible. The first line in Table 5 thus presents the mean value of all true values for all estimated loading parameters, the mean of their simulated values, of their *peb*-values, of their standard deviation, of their average standard errors,

Table 5. Resu ts of the Monte Car o s mu at on study for the oad ng parameters of the MMQ mode w th two nd cators per tra t–method un t: means of the expected va ues and of the parameter est mates (s mu ated va ue), parameter est mat on b ases (*peb*), standard dev at ons (Est. SD), means of the standard errors (SE), standard error dev at ons (SE Dev.), standard error b ases (*seb*) and coverage of the 95% confidence nterva (coverage)

Type of parameter	Mean expected va ue	Mean s mu ated va ue	Mean *peb*	Mean Est. SD	Mean SE	Mean SE Dev.	Mean *seb*	Mean coverage
250 observat ons								
Tra t se f	0.98	1.00	0.02	0.13	0.13	0.03	−0.03	0.96
Tra t A + B	0.28	0.28	0.00	0.11	0.10	0.01	−0.05	0.94
Method A + B	1.00	1.02	0.02	0.23	0.20	0.09	−0.12	0.94
500 observat ons								
Tra t se f	0.98	0.99	0.01	0.09	0.09	0.02	0.00	0.96
Tra t A + B	0.28	0.28	0.00	0.07	0.07	0.01	−0.02	0.95
Method A + B	1.00	1.00	0.01	0.15	0.14	0.03	−0.06	0.94
750 observat ons								
Tra t se f	0.98	0.99	0.01	0.08	0.08	0.01	−0.01	0.96
Tra t A + B	0.28	0.28	0.00	0.06	0.06	0.00	−0.01	0.95
Method A + B	1.00	1.00	0.01	0.12	0.11	0.02	−0.02	0.95
1000 observat ons								
Tra t se f	0.98	0.99	0.01	0.07	0.07	0.01	−0.03	0.95
Tra t A + B	0.28	0.28	0.00	0.05	0.05	0.00	−0.01	0.95
Method A + B	1.00	1.00	0.01	0.10	0.10	0.02	−0.03	0.95

Note. Tra t se f: tra t oad ng parameters of the se f-report. Tra t A + B: tra t oad ng parameters of peer parameters of peer report A and peer report B. Method A + B: method oad ng parameters of peer report A and peer report B.

Table 6. Resu ts of the Monte Car o s mu at on study for the var ances and covar ances of the MMQ mode w th two nd cators per tra t–method un t: means of the expected va ues and of the parameter est mates (s mu ated va ue), parameter est mat on b ases (*peb*), standard dev at ons (Est. SD), means of the standard errors (SE), standard error dev at ons (SE Dev.), standard error b ases (*seb*), and coverage of the 95% confidence nterva (coverage)

Type of parameter	Mean expected va ue	Mean s mu ated va ue	Mean *peb*	Mean Est. SD	Mean SE	Mean SE Dev.	Mean *seb*	Mean coverage
250 observat ons								
Var (Tra t)	0.65	0.65	0.00	0.10	0.09	0.02	0.02	0.95
Var (Method)	0.56	0.57	0.02	0.13	0.12	0.04	0.08	0.95
Cor (Tra t–tra t)	0.32	0.32	0.01	0.06	0.06	0.00	−0.04	0.94
Cor (Tra t–method)	−0.01	−0.01	0.11	0.05	0.05	0.01	−0.05	0.94
Cor (Method–method)	0.13	0.13	−0.01	0.06	0.05	0.01	−0.07	0.93
500 observat ons								
Var (Tra t)	0.65	0.65	0.00	0.07	0.07	0.01	0.00	0.95
Var (Method)	0.56	0.57	0.01	0.09	0.08	0.01	−0.05	0.94
Cor (Tra t–tra t)	0.32	0.32	0.00	0.04	0.04	0.00	−0.01	0.95
Cor (Tra t–method)	−0.01	−0.01	0.16	0.04	0.04	0.00	−0.03	0.94
Cor (Method–method)	0.13	0.13	−0.01	0.04	0.04	0.00	−0.03	0.94
750 observat ons								
Var (Tra t)	0.65	0.65	0.00	0.06	0.06	0.01	0.01	0.95
Var (Method)	0.56	0.57	0.01	0.07	0.07	0.01	−0.03	0.95
Cor (Tra t–tra t)	0.32	0.32	0.00	0.03	0.03	0.00	−0.01	0.94
Cor (Tra t–method)	−0.01	−0.01	0.06	0.03	0.03	0.00	−0.01	0.95
Cor (Method–method)	0.13	0.13	0.00	0.03	0.03	0.00	−0.01	0.95
1000 observat ons								
Var (Tra t)	0.65	0.65	−0.00	0.05	0.05	0.00	−0.01	0.95
Var (Method)	0.56	0.57	0.00	0.06	0.06	0.01	−0.04	0.94
Cor (Tra t–tra t)	0.32	0.32	0.00	0.03	0.03	0.00	−0.01	0.95
Cor (Tra t–method)	−0.01	−0.01	0.14	0.03	0.03	0.00	−0.01	0.95
Cor (Method–method)	0.13	0.13	0.00	0.03	0.03	0.00	−0.02	0.94

of their standard deviation of standard errors, of their *seb*-values and of their coverages.[2]

As can be seen in Table 5, the loading parameter estimation biases are smaller than .02 for all kinds of parameters, and the standard error estimation biases are smaller than .12. Between 94% and 96% of the replications included the theoretically expected values within the observed 95% confidence interval. Hence, the WLSMV estimator provides very stable loading parameter estimates throughout this simulation study.

The average parameter estimation biases for the variances and covariances are quite similar (see Table 6). All but three biases do not exceed values larger than .06. The large mean parameter estimation biases (*peb* .11, *peb* .16 and *peb* .14) are mainly due to small deviations of the average parameter estimation from the true value, whereby the true value is very close to 0 (.005 in all cases). Therefore, even a small deviation produces large parameter estimation biases.

In sum, for this special situation of two indicators for each trait–method unit (18 observed variables) and 500 observations, the WLSMV estimator is a very stable estimator of model parameters combined with an acceptable estimation of the χ^2 goodness-of-fit index.

The findings of all other simulations point in the same direction as the results of the first simulation. Figure 2 shows a graphical comparison of the expected and observed critical χ^2 probabilities. The *x*-axis represents all simulation conditions ranging from 2 indicators and 250 observations (2/250) to 8 indicators and 1000 observations (8/1000). On the *y*-axis the observed χ^2 quantiles are depicted. The four graphs represent the theoretically expected quantiles for $p = .01$, $p = .02$, $p = .05$, and $p = .10$. For a perfect match one would expect four horizontal lines on the four corresponding observed quantile levels. In general, the χ^2 approximation becomes more and more biased with an increasing number of indicators. Increasing the sample size always leads to better approximations. Accepting a deviation of 3% at the $p = .05$ level, the WLSMV estimator yields valid results for 2, 3 and 4 items per trait–method unit (18, 27, and 36 observed variables) with at least 250 observations, for 5 and 6 items (45 and 54 observed variables) with 500 observations, for 7 items (63 observed variables) with 750 observations, and for 8 items (72 observed variables) with 1000 observations.

Throughout all simulations the same results concerning the estimations biases were found: the estimations of the loading parameters, variances, and correlations are very stable.[3] Only a few parameter estimation biases exceed the critical value of .10. As already described above, these large parameter estimation biases only occur for correlations between trait and method factors and they are due to a small deviation of the observed mean parameter estimates from their true values, which are very close to 0. All other parameter estimation biases are very small ($|peb| \leq .04$ in all cases). Standard error biases are negligibly small as well ($|seb| \leq .08$ in all cases). Thus, the WLSMV estimator yields good parameter estimation properties even for the very complex models with 8 items per trait–method unit (72 observed variables).

[2] *Detailed results for every single parameter can be obtained from the first author.*

[3] *Complete results of all empirical applications with increasing number of items and simulation studies can be obtained from the first author. They are omitted here due to space limitations.*

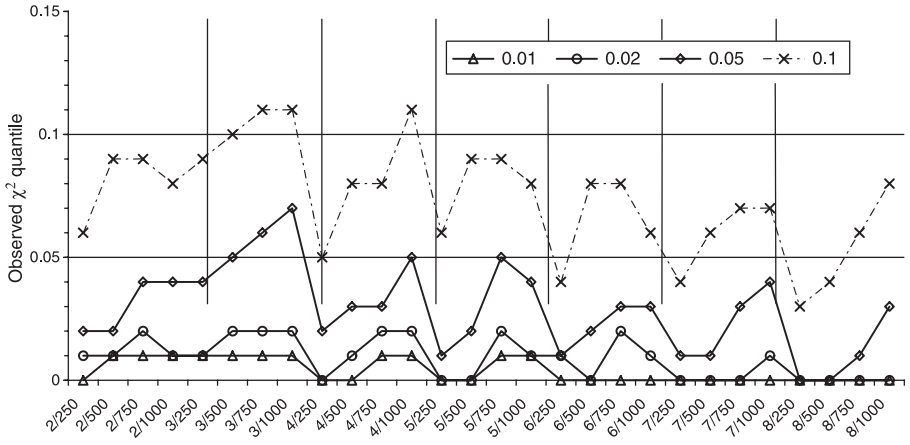

Figure 2. Comparison of the theoretically expected χ^2 probability ($p = .01$, $p = .02$, $p = .05$ and $p = .10$) with the observed probabilities depending on the number of items per trait–method unit and sample size (e.g. 2/250 indicates 2 items per trait–method unit and 250 observations).

6. Discussion

The Monte Carlo simulations showed that the χ^2 approximation of the WLSMV estimator implemented in Mplus (Muthén & Muthén, 2004) generally works very well. Even with relatively small sample sizes and many observed variables it provides reliable parameter estimates and a good – albeit slightly liberal – χ^2 approximation. If the number of observations is increased, even large models can be solidly tested. In most cases, parameter estimation biases and standard error biases are negligibly small. No less than 91% and no more than 98% of the replications result in models in which the theoretically expected values fall into the observed 95% confidence interval (except for the trait–method correlations in the model with eight indicators). These results are in line with the findings reported by Muthén *et al.* (in press) and Flora and Curran (2004). Muthén *et al.* reported stable and adequate results of the WLSMV estimator for sample sizes exceeding 400 when there are skewed distributions. Flora and Curran replicated their results, finding valid estimates for 200 observations. In our simulation study we increased the number of observed variables and examined more complex models with ordinal response variables. As reported, the WLSMV estimator provided very stable results and seems to be a useful, practical and adequate method for the analysis of large models with many observed ordinal response variables. For the classical CTC(M 1) model for ordinal variables with three traits and three methods, at least 250 observations are needed when there are 2–4 indicators, 500 for 5 and 6 indicators, 750 for 7 indicators and at least 1000 for 8 indicators.

The WLSMV estimator provides adequate results of complex MTMM analyses. These models are the basis for the development of psychological questionnaires and tests because they provide estimations of the consistency and method-specificity coefficients for every single item. Moreover, they allow researchers to examine the convergent and discriminant validity for latent trait and method effects. In spite of these very positive findings, we are aware that at this point these results apply only to the types of data that existed in our application. Additional research is needed to examine the performance of the WLSMV estimator when other types of data are encountered and in which other MTMM models seem to be more appropriate.

In addition to the excellent psychometric properties of the WLSMV estimator (respecting the categorical character of a scale), its computational properties (stable and reliable results examining small samples) make it a very promising and useful estimator in multimethod analysis, leading us to conclude that multimethod item and test analyses can and, of course, should be conducted using the WLSMV estimator.

Acknowledgement

This research has been funded by grant Ei 379/5 1 from the Deutsche Forschungsgemeinschaft.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Browne, M. W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62 83.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin*, *56*, 81 105.

Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37 55). Thousand Oaks, CA: Sage.

Dumenci, L. (2000). Multitrait multimethod analysis. In S. D. Brown & H. E. A. Tinsley (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 583 611). San Diego, CA: Academic Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1 26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: SIAM.

Eid, M. (1995). *Modelle der Messung von Personen in Situationen* [Models for measuring persons in situations]. Weinheim: Psychologie Verlags Union.

Eid, M. (2000). A multitrait multimethod model with minimal assumptions. *Psychometrika*, *65*, 241 261.

Eid, M, Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology*. Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait specific method effects in multitrait multimethod models: A multiple indicator CT C(M 1) model. *Psychological Methods*, *8*, 38 60.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative models of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466 491.

Grayson, D. A., & Marsh, H. W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis: Multitrait multimethod models. *Psychometrika*, *59*, 121 134.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*, 335 361.

Marsh, H. W. (1993). Multitrait multimethod analyses: Inferring each trait method combination with multiple indicators. *Applied Measurement in Education*, *6*, 49 81.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory factor analysis of multitrait multimethod self concept data: Between group and within group invariance constraints. *Multivariate Behavioral Research*, *28*, 313 349.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177 198). Thousand Oaks, CA: Sage.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait multimethod analyses: Application of second order confirmatory factor analysis. *Journal of Applied Psychology*, *73*, 107 117.

McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self reports and ratings. *Journal of Personality and Social Psychology*, *43*, 293 303.

Muthén, B. O. (1983). Latent variable structural equation modeling with categorical variables. *Journal of Econometrics*, *49*, 22 45.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (in press). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599 620.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide*. Los Angeles: Author.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*, 131 151.

Satorra, A. (1992). Asymptotoic robust inferences in the analysis of mean and covariance structures. In P. V. Marsden (Ed.), *Sociological methodology 1992* (pp. 249 278). Oxford: Blackwell.

Schermelleh Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness of fit measures. *Methods of Psychological Research*, *8*, 23 74.

Stanton, A. L., Kirk, S. B., Cameron, C. L., & Danoff Burg, S. (2000). Coping through emotional approach: Scale construction and validation. *Journal of Personality and Social Psychology*, *78*, 1150 1169.

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1994). Testtheoretische Analysen des Mehrdimensionalen Befindlichkeitsfragebogen (MDBF) [Theoretical analysis of the Multi dimensional Mood Questionnaire (MMQ)]. *Diagnostica*, *40*, 320 328.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393 408.

Trierweiler, L. I., Eid, M., & Lischetzke, T. (2002). The structure of emotional expressivity: Each emotion counts. *Journal of Personality and Social Psychology*, *82*, 1023 1040.

Watson, D., Hubbard, B., & Wiese, D. (2000). Self other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*, 546 558.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait multimethod data. *Applied Psychological Measurement*, *9*, 1 26.

Appendix

Translation of the items of the German scale into English

pleasant/unpleasant	awake/sleepy	calm/restless
zufrieden–contented	ausgeruht–rested	ruhelos–restless
schlecht–awful	schlapp–worn out	gelassen–serene
gut–great	müde–tired	unruhig–uneasy
unwohl–uncomfortable	munter–energetic	entspannt–relaxed
wohl–good	schläfrig–sleepy	ausgeglichen–at ease
unglücklich–unhappy	wach–alert	angespannt–tense
unzufrieden–discontented	frisch–fresh	nervös–nervous
glücklich–happy	ermattet–exhausted	ruhig–calm