UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Rose, N., et al. (2019). Model-Based Manifest and Latent Composite Scores in Structural Equation Models. *Collabra: Psychology*, 5(1): 9. DOI: https://doi.org/10.1525/collabra.143

ORIGINAL RESEARCH REPORT

Model-Based Manifest and Latent Composite Scores in Structural Equation Models

Norman Rose*, Wolfgang Wagner*, Axel Mayer† and Benjamin Nagengast*

Composite scores are commonly used in the social sciences as dependent and independent variables in statistical models. Typically, composite scores are computed prior to statistical analyses. In this paper, we demonstrate the construction of model-based composite scores that may serve as outcomes or predictors in structural equation models (SEMs). Model-based composite scores of manifest variables are useful in the presence of ignorable missing data, as full-information maximum likelihood estimation can be used for parameter estimation. Model-based composite scores of latent variables account for measurement error in the aggregated variables. We introduce the pseudo-indicator model (PIM) for the construction of three composite scores: (a) the sum score, (b) the weighted sum score, and (c) the average score of manifest and latent variables in SEM. The utility of manifest model-based composite scores in the case of missing values is shown by a simulation study. The use of multiple manifest and latent model-based composite scores in SEM is illustrated with data from motivation research.

Keywords: composite scores; index variables; latent variables; measurement model; structural equation model

Aggregates of variables, such as unweighted and weighted sum scores, are commonly used in the social and behavioral sciences. For example, total scores, mean scores, number-right scores, and proportion-correct scores are single scores formed by aggregating information from multiple variables. The same is true for many index variables, such as the *Economic Social and Cultural Status* (ESCS) index used in PISA (OECD, 2010), or numerous medical and prognostic indices, which have been developed in medicine and human health care (e.g., Feeny, Huguet, McFarland, & Kaplan, 2009; Gibson, 1981; Kirshner & Guyatt, 1985; Redelmeier & Lustig, 2001; Yourman, Lee, Schonberg, Widera, & Smith, 2012).

Another application of composite variables in psychometrics involves item parcels that are formed as sums or averages of item scores. Item parcels serve as indicators of latent variables in structural equation models (SEM). Despite the existence of controversial discussions about item parcels (T. D. Little, Cunningham, Shahar, & Widaman, 2002; T. D. Little, Rhemtulla, Gibson, & Schoemann, 2013), their use in measurement models is common practice in many areas of psychology.

However, index variables, scale scores, as well as item parcels are computed on observed data, a practice that can be problematic when some of the data are missing. This problem cannot be adequately addressed by standard full-information maximum likelihood (FIML) estimation in SEM because the composite variables are computed prior to the statistical analyses. Until recently, multiple imputation (MI) has therefore been the most popular method for overcoming the missing data problem in composite variables. Savalei and Rhemtulla (2017) proposed a two-stage maximum likelihood estimation method for dealing with missing data at the item level in a way that could avoid MI.

A potential problem involved in the use of composite scores comprised of manifest variables is the measurement error that occurs when some of the component variables are unreliable. A composite score of fallible measures is plagued by the accumulation of the error terms from each constituting variable (Raykov, 1998; Raykov, Marcoulides, & Li, 2017a, 2017b). This is not a problem when item parcels are used because the measurement error is explicitly modeled. However, statistical analyses with index variables or scale scores that do not account for the lack of reliability can lead to biased parameter estimates (e.g., attentuation of regression or correlation coefficients between the contaminated composite score and external variables; Spearman, 1904; Stefanski, 1985).

Model-based composite scores comprised of latent variables in multidimensional measurement models are promising for addressing various research questions. For example, multidimensional psychometric models are commonly used for assessing students' competencies (e.g., mathematics, reading comprehension, and science)

* Hector Research Institute of Education Sciences and Psychology, University of Tübingen, DE

† RWTH Aachen University, Institute of Psychology, DE

Corresponding author: Norman Rose (norman.rose@uni-tuebingen.de)

in educational research. A composite score of these latent proficiencies could serve as a student's school achievement index adjusted for measurement error.

Composite scores of latent variables have also been discussed in intelligence research as a viable alternative way to model the general ability factor (i.e., the *G*-factor) on the basis of the specific ability factors. In their overview, Conway and Kovacs (2015) argued that intelligence is better conceptualized as a set of different cognitive abilities, and that "... new models interpret *g* [i.e., the *G*-factor] as an emergent property and individual scores on *g* as an index of the collective performance on a battery of tests" (p. 2). A latent composite score may better represent this idea than a higher order factor. Similarly, Eid, Geiser, Koch, and Heene (2017) view a composite score of domain-specific intelligence factors for representing the *G*-factor as an alternative to bi-factor models or second-order factor models. Indeed, there are large conceptual differences between second-order factors and composite scores of latent variables. In second-order factor models, the first-order factors are considered fallible indicators of the second-order factor. The measurement model describes the stochastic dependency between second- and first-order factors by means of (linear) regressions. By contrast, composite scores and their constituent or component variables are not stochastically but functionally related. Hence, the values of a composite score are fully determined by the constituting first-order factors. However, the contribution of each component to the composite score as well as the correlation between the composite score and each component can be controlled by choosing appropriate weights. This is an advantage in comparison with higher order factors in many applications. Nevertheless, there is no simple recipe that can be applied to decide whether composites of latent variables or higher order factor models should be used. The choice of the model needs to be based on theoretical considerations regarding the specific research question and depends on the theoretical interpretation given to the higher order construct.

In the proposed model-based approach, the composite scores are not computed prior to the analysis but are defined as linear combinations of manifest and/or latent variables in SEM. The specified models can be estimated using standard estimation procedures and standard software for SEM. This allows the user to adjust the composite scores of manifest variables for ignorable missing data by applying common full information maximum likelihood (FIML) estimation (Arbuckle, 1996; Enders, 2001).

Here, we introduce the pseudo-indicator model (PIM), which allows unweighted and weighted sum scores and the average score of manifest and latent variables to be constructed. We will first outline the basic idea behind the construction of model-based composite scores in SEM. We then introduce the PIM for composite scores of manifest and latent variables. Composite scores typically serve as independent or dependent variables in more complex models. We subsequently outline the correct specification of the PIM in explanatory SEM with additional predictors, covariates, and outcomes. Using SEM with manifest model-based composite scores requires the inclusion of additional variables (i.e., the component variables of the composite score), which needs to be considered in the assessment of model fit. This is described before we demonstrate the value of model-based composite scores in the case of missing data by presenting a simulation study with ignorable missing data. We also present an application of two model-based composite scores with a real data example from Gaspard et al. (2015). A latent composite score of students' beliefs about the utility value of mathematics is the outcome in a multiple regression model that includes an index of socioeconomic status as a predictor. The latter is formed as a model-based composite score of manifest variables that suffer from a substantial amount of missing data. Finally, we discuss the advantages and disadvantages of the presented approach.

Constructing Model-Based Composite Scores

Let C be a composite variable defined as a linear combination of manifest or latent variables. Note that we use the term *manifest composite score* when the aggregated variables are manifest. Aggregates of latent variables are denoted by a *latent composite score*. So, the manifest composite score is given by

$$C = \sum_{q=1}^{Q} \gamma_q Y_q. \tag{1}$$

The variables $\boldsymbol{Y} = (Y_1, \ldots, Y_q, \ldots, Y_Q)^T$ that are aggregated into C are called constituent or component variables in the remainder of this paper. The corresponding weights are denoted by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q, \ldots, \gamma_Q)^T$. The latent composite score is obtained by replacing the manifest component variables in Equation 1 by Q latent component variables $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q, \ldots, \eta_Q)^T$. Linear combinations of manifest or latent variables can be constructed in SEM, because (a) the means and the variances of linear combinations of the component variables, and (b) the covariances between linear combinations and the component variables are fully implied by the mean and covariance structure of the component variables. For example, manifest and latent change variables are a widely used linear combination in longitudinal SEM (McArdle, 2009; Steyer, Eid, & Schwenkmezger, 1997; Steyer, Krambeer, & Hannover, 2004). The difference variable $\eta_2 - \eta_1$ in a latent change model is an example of a composite score according to Equation 1, with the weights $\gamma_1 = -1$ and $\gamma_2 = 1$. Latent difference variables have also been used to model method effects (Pohl & Steyer, 2010; Pohl, Steyer, & Kraus, 2008). Furthermore, Pohl and Steyer (2010) defined a common trait variable as the mean of true score variables in addition to method factors in order to account for method effects in longitudinal SEM. As Mayer, Steyer, and Mueller (2012) demonstrated, growth components in latent growth curve models can also be defined as weighted sums of latent state variables according to Equation 1. Finally, Mayer (2013) presented a comprehensive mathematical framework that includes true change models, latent growth curve models, and method effect models as special cases. In all these examples, composite variables of manifest or latent variables are defined in SEM. The contribution of this paper is to demonstrate how to construct model-based composite scores with arbitrary sets of weights.

The PIM exploits the fact that it is possible to find sets of linear combinations that have a one-to-one correspondence (i.e., bijective function) with their component variables. Hence, component variables can also be written as mathematical functions of the linear combinations they constitute. For example, the sum score $C = \sum_{q=1}^{Q} Y_q$ is a function of the component variables Y_q, with q = 1, ..., Q. In turn, each variable Y_q can be written as the difference between C and the reduced sum score $C^{(-q)} = \sum_{i \neq q} Y_i$, of the variables Y_i, with i = 1, ..., q – 1, q + 1, ..., Q. As noted previously, a difference variable is a special case of a weighted sum score. Therefore, each component variable Y_q can be written as

$$Y_q = w_1 C + w_2 C^{(-q)}, \tag{2}$$

with the weights w_1 = 1 and w_2 = –1 because

$$\begin{aligned} Y_q &= 1 \cdot C + (-1) \cdot C^{(-q)} \\ &= C^{(-q)} + Y_q - C^{(-q)} \\ &= Y_q. \end{aligned} \tag{3}$$

Variables in an SEM can be specified as weighted sums of other variables in commonly used SEM software packages by (a) fixing the path coefficients to the values of the weights, (b) fixing the intercept to zero, and (c) setting the regression residuals to zero (i.e., fixing the residual variance to zero). This general principle is used to construct model-based composite scores. The component variables are specified as weighted sums of the desired composite scores and the other component variables. However, composite scores are not manifest variables in the model. The technical trick is to define the composite score of manifest variables as a *pseudo-latent* variable in a measurement model with one of the manifest component variables as the indicator variable without a measurement residual. Latent composite scores are defined as pseudo higher order factors with one of the latent component variables as the indicator variable with a residual variances fixed to zero. The term *pseudo-latent* refers to the fact that model-based composite scores of manifest variables are not free from measurement error. Similarly, composite scores of latent variables are not higher order factors, although latent component variables are seemingly defined as indicators of the composite scores without a residual.

The model specification of the PIM is explained for three common composite scores: (a) the sum score, $C = \sum_{q=1}^{Q} Y_q$; (b) the average score, $C = Q^{-1} \sum_{q=1}^{Q} Y_q$; and (c) the weighted sum score, $C = \sum_{q=1}^{Q} \gamma_q Y_q$. Although the sum score and the average score are just special cases of the weighted sum score, we explain the specification of each composite score starting with the sum score as the simplest case. In the introduction of the PIM, we focus on manifest composite scores because the specification of latent composite scores is mathematically equivalent but requires a measurement model of the component variables. Peculiarities in the specification of latent composite scores are discussed after the general introduction of the PIM.

The Pseudo-Indicator Model (PIM)

The *pseudo-indicator* model received its name from the fact that one of the component variables Y_1, ..., Y_Q of the composite score C needs to be arbitrarily chosen as the pseudo-indicator variable Y_r, which is specified in the measurement model as the manifest indicator variable of the pseudo-latent composite score. The measurement residual of Y_r is fixed to zero. Q–1 additional paths from each component variable $Y_{q \neq r}$ to the pseudo-indicator Y_r are also specified. **Figure 1** shows the path diagram of a PIM with a composite score C of the three component

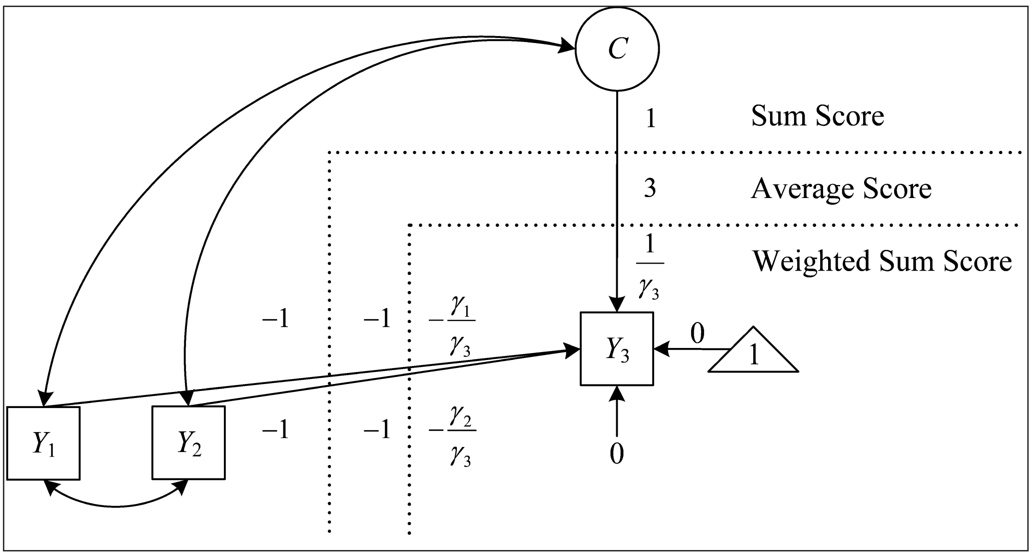

Figure 1: The path diagram shows the PIM with a manifest model-based composite score C (i.e., the sum score, the average score, or the weighted sum score) of three manifest variables Y_1 to Y_3. The third variable Y_3 has been chosen as the pseudo-indicator variable.

variables Y_1 to Y_3. The variable Y_3 has been chosen as the pseudo-indicator variable with the residual variance and the intercept fixed to zero. The path diagrams of the PIM is structurally equivalent for the sum score, the average score, or the weighted sum score. They only differ in the fixed values of the path coefficients and the factor loading of Y_r on C. In **Figure 1** the different values of the path coefficients ($Y_1 \rightarrow Y_3$ and $Y_2 \rightarrow Y_3$) and the factor loadings ($C \rightarrow Y_3$) are shown for the three composite scores (the sum score, the average score, and the weighted sum score) separated by dotted lines.

The PIM with the model-based sum score and the average score. In order to construct the sum score using the PIM, the factor loading of the pseudo-indicator variable Y_r must be fixed to one. The path coefficients of the Q–1 additional paths from each component variable $Y_{q \neq r}$ to the pseudo-indicator Y_r need to be set to minus one. In existing SEM software, this means that the user needs to specify a multiple regression of the pseudo-indicator variable Y_r on all $Y_{q \neq r}$ with all regression coefficients fixed to minus one. Finally, the residual variance and the intercept of the pseudo-indicator variable Y_r must be fixed to zero. The resulting equation for Y_r is

$$Y_r = C - \sum_{q \neq r} Y_q. \tag{4}$$

This equation is equivalent to Equations 2 and 3 and simply means that the component variable Y_r can be written as the sum score of all component variables minus the reduced sum score of only the variables $Y_{q \neq r}$ without Y_r. It is important that the means and variances of the component variables $Y_{q \neq r}$ as well as the covariances between all component variables $Y_{q \neq r}$ are freely estimated. Furthermore, the composite score C is correlated with its component variables, so all covariances $Cov(C, Y_{q \neq r})$ need to be estimated without restrictions. Note that the component variables $Y_{i \neq r}$ are formally exogenous variables in an SEM. Purely exogenous variables are considered fixed variables by default in many SEM packages such as lavaan (Rosseel, 2012) and M*plus* (i. e. M*plus* 7.2, Muthén & Muthén, 1998–2018). In the presence of missing data, the variables $Y_{q \neq r}$ have to be specified as stochastic variables to avoid listwise deletion of all cases with missing values in one or more variables $Y_{q \neq r}$ and to utilize FIML. In some software (e.g., M*plus*), this can be achieved by declaring the distributional parameters of exogenous variables (e.g., means and variances) or covariances with other variables to be freely estimable model parameters.

The specification of the PIM with the model-based average score instead of the sum score is obtained by fixing the factor loading of the pseudo-indicator variable Y_r to the number of the component variables (Q) of the composite score C. Apart from this, the model specification of the PIM with the sum score is identical to the specification with the average score.

The PIM with the model-based weighted sum score. The PIM also allows for model-based construction of the weighted sum score by adapting the value of the factor loading of the pseudo-indicator variable Y_r and the path coefficients of the (Q–1) paths of the component variables $Y_{i \neq r}$ on Y_r. The values can easily be derived from the equation of the weighted sum score by solving for the pseudo-indicator variable, which gives

$$\begin{aligned} C &= \sum_{q=1}^{Q} \gamma_q Y_q \\ &= \gamma_r Y_r + \sum_{q \neq r} \gamma_q Y_q \\ Y_r &= \frac{1}{\gamma_r} C + \sum_{q \neq r} -\frac{\gamma_q}{\gamma_r} Y_q. \end{aligned} \tag{5}$$

Hence, the loading of Y_r on C must be fixed to the inverse weight γ_r^{-1} of the pseudo-indicator variable Y_r. Furthermore, for each of the remaining component variables $Y_{q \neq r}$, the path coefficient (i.e., the regression coefficient) of the path from $Y_{q \neq r}$ to the pseudo-indicator variable Y_r needs to be fixed to the negative ratio of weights $-\gamma_q/\gamma_r$. The correctness of the specification of the model can also be demonstrated by rearranging and simplifying Equation 5 (shown in Appendix A). Again, the intercepts and residual variances of the pseudo-indicator variables are fixed to zero. The variances and means of the other component variables $Y_{q \neq r}$, along with their covariances with each other and with the weighted sum score C, are freely estimated. The specification of the path coefficients ($Y_q \rightarrow Y_r$) and the factor loadings ($C \rightarrow Y_r$) depending on the type of composite score is also summarized in **Table 1**.

The PIM with latent composite scores. Figure 2 shows the PIM with a latent composite score C of three latent component variables η_1 to η_3. In **Figure 2** the fixed values of the path coefficients ($\eta_1 \rightarrow \eta_3$ and $\eta_2 \rightarrow \eta_3$) and the factor loading ($C \rightarrow \eta_3$) are shown for each of the three latent composite scores (the sum score, the average score, and the weighted sum score) separated by dotted lines. This model is structurally equivalent to the PIM with the manifest composite score (cf. **Figure 1** and **Table 1**). The only difference is that each latent component variable η_q is a latent variable in a measurement model that is based on the manifest indicator variables X_{iq} (see Equation 6)

$$X_{iq} = \nu_i + \lambda_{iq}\eta_q + \varepsilon_i. \tag{6}$$

The measurement models of each component variable $\eta_{q \neq r}$ can be identified by the rules commonly used in confirmatory factor analysis. For example, at least one factor loading and one measurement intercept can be fixed (e.g., $\lambda_{iq} = 1$ and $\nu_i = 0$), whereas the mean and the

Table 1: Values of the Fixed Path Coefficients in the PIM for Constructing the Model-Based Sum Scores, the Average Scores, and the Weighted Sum Scores of Manifest or Latent Component Variables.

Paths	Composite score		
	Sum	Average	Weighted sum
$C \rightarrow Y_r$ or η_r	1	Q	γ_r^{-1}
$Y_{q \neq r} \rightarrow Y_r$ or $\eta_{q \neq r} \rightarrow \eta_r$	−1	−1	$-\gamma_q\gamma_r^{-1}$
$C \leftrightarrow Y_{q \neq r}$ or $\eta_{q \neq r}$	Free covariances $Cov(C, Y_{q \neq r})$ or $Cov(C, \eta_{q \neq r})$, no directed paths		

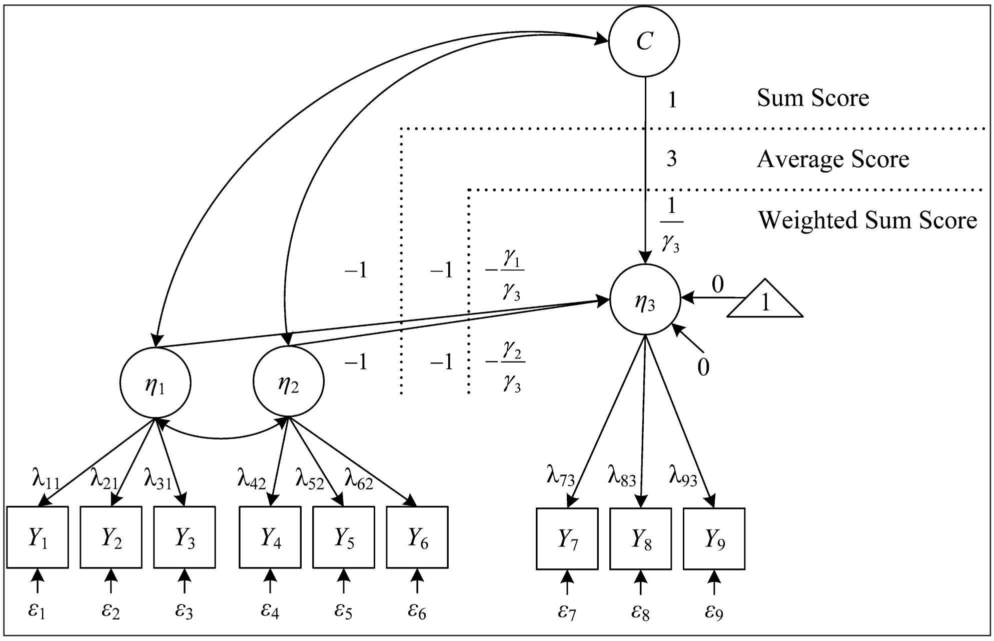

Figure 2: The path diagram shows the PIM with a latent model-based composite score C (i.e., the sum score, the average score, or the weighted sum score) of three latent variables η_1 to η_3. The third variable η_3 serves as the pseudo-indicator variable.

variance of η_q are freely estimated. Alternatively, the scale of the latent variables can be fixed by specifying the distributional parameters (e.g., $E(\eta_{q\neq r}) = 0$ and $Var(\eta_{q\neq r}) = 1$) and specifying the factor loadings and measurement intercepts as free estimands. Because the latent pseudo-indicator variable η_r is completely determined by the composite score C and the other component variables $\eta_{q\neq r}$, the mean and the variance of η_r cannot directly be set to arbitrary values. Therefore, the measurement model of η_r must be identified by fixing at least one loading λ_{ir} (e.g., $\lambda_{ir} = 1$). If all latent component variables have means set to zero, then the mean of the latent composite score must also be fixed to zero, and the measurement intercepts of all indicators of η_r are free estimands. If at least one of the latent component variables has a non zero mean, the mean of the latent composite score is also not fixed to zero but is freely estimated. In this case, the measurement intercept of at least one indicator variable of the latent pseudo-indicator η_r must be fixed (e.g., $\nu_{ir} = 0$).

Model-Based Composite Scores in Explanatory Latent Variable Models

So far, the explanations for how to construct model-based composite scores were limited to SEM consisting merely of the composite score and its component variables. Such models allow the distributional parameters (i.e., means and variances) of manifest or latent composite scores to be estimated. In multiple-group SEM, the PIM can be used to test for mean differences in manifest or latent composite scores between groups. However, in most other applications, the composite score is usually part of a larger SEM with additional variables that are either predictors, covariates, or outcomes of composite scores. In such cases, the PIM is just a submodel in a more general explanatory SEM, which includes the model of substantive interest (i.e., the target model). In contrast to the variables in the target model, including the composite score, the component variables are typically not of theoretical interest. They serve only as auxiliary variables for constructing the composite score in SEM. An essential condition for the correct use of the PIM is that the component variables must be included without compromising the target model. Formally, this means that the theoretical model-implied variance-covariance structure and mean structure of the variables of the target model must remain unaffected by the inclusion of the component variables. This requires the correct specification of the covariances between the component variables and all variables from the target model such that the target model is preserved. The preservation of the model-implied variance-covariance structure and the model-implied means of the target model when using model-based composite scores is mandatory for constructing model-based composite scores in more complex SEM. This fact serves as a guiding principle for correct model specification. A second guiding principle results from the fact that a composite score does not have implications regarding (a) the variance-covariance structure and the means of its component variables or (b) the covariances between the component variables and other variables in the target model. Therefore, there is always a model that includes all variables that are included in the target model with the exception of the model-based composite score but with all the component variables of the composite score that must be equivalent to a model that additionally includes the model-based composite scores. Model equivalence in a statistical sense means (a) the same theoretical model-implied variance-covariance structure and mean structure, (b) identical values of the discrepancy function, degrees of freedom, χ^2-values, and (c) identical goodness-of-fit indices (Bentler & Satorra, 2010; T. Raykov & Penev, 1999). Based on these two guiding principles, the correct specifications of the

PIM within a more general PIM can be derived and can be summarized as follows:

- In an SEM with a PIM for constructing a manifest model-based composite score, the manifest component variables $Y_{q \neq r}$ have (a) free covariances with all manifest exogenous variables and (b) free covariances with the residuals of all endogenous manifest variables. Note that the model-based manifest composite scores are specified in SEM as pseudo-latent variables, which need to be treated like manifest variables. Therefore, the component variables also have (a) free covariances with exogenous model-based composite scores and (b) free covariances with the residuals of endogenous model-based composite scores. Illustration of how these rules are applied to SEM with a model-based composite score as an outcome, a predictor, and as an intermediate variable (e.g., mediation model) in lavaan and M*plus* are given in Sections 3.5, 3.6, and 3.8 of the PDF document RoseEtal2019suppl.pdf, which is part of the online supplemental material (https://osf.io/b6h7n/).[1]
- In an SEM with a PIM for constructing a latent model-based composite score, the latent component variables $\eta_{q \neq r}$ have (a) free covariances with all exogenous variables of the structural model and (b) free covariances with the residuals of all latent endogenous variables in the structural model. However, the covariances between the latent component variables $\eta_{q \neq r}$ and the measurement residuals of manifest indicators of the latent variables in the model must be zero. In the Sections 3.7 and 3.8 of the supplemental PDF document (RoseEtal2019suppl.pdf) the correct model specification of SEM with latent model-based composite scores as predictors, outcomes and intermediate variables (e.g., mediator) are presented.
- In an SEM with latent and manifest model-based composite scores, the covariances between the latent component variables $\eta_{q \neq r}$ of a latent composite score with the manifest component variables $Y_{q \neq r}$ of a manifest composite score are fixed to zero. However, the manifest component variables $Y_{q \neq r}$ have free covariances with the measurement errors of the manifest indicators of the latent component variables $\eta_{q \neq r}$ and η_r. The applied data example in this paper illustrates how manifest and latent composite scores are simultaneously included in SEM by means of the described rules.

Note that this set of rules ensures that the inclusion of model-based composite scores along with their component variables does not alter the target model or the model fit of the joint model consisting of the component variables and the target model. The manifest component variables are merely auxiliary variables with these specifications. However, some of the rules may be relaxed to simplify the model or to consider specific assumptions. For example, in an SEM consisting of a manifest model-based composite score C of the variables $Y_1, \ldots, Y_Q$ and one or more latent variables ξ measured with $X_1, \ldots, X_P$, the covariances between the measurement residuals of $X_1, \ldots, X_P$ and the component variables $Y_1, \ldots, Y_Q$ are free parameters according to the rules derived above. However, under the measurement invariance of ξ across all levels of $Y_1, \ldots, Y_Q$, conditional stochastic independence of all X_j and Y_q is implied. Hence, $Cov(X_j, Y_q | \eta) = 0$. In this case, it is sufficient to specify the covariances $Cov(\eta, Y_{q \neq r})$ as free parameters, whereas the covariances between the component variables $Y_1, \ldots, Y_Q$ and the measurement residuals of $X_1, \ldots, X_P$ are fixed to zero. However, the resulting model is no longer equivalent to a model without the composite score. The degrees of freedom and, therefore, the model fit statistics are different. There might be more applications that suggest that these rules should be relaxed for constructing model-based composite scores. Relaxing the rules should always be done with caution and only based on theoretical considerations to avoid unintended restrictions or misspecifications, which can lead to biased parameter estimation.

There may also be cases where some rules need to be relaxed for reasons of identification. An example is a PIM with a model-based composite score C that consists of both manifest (Y_q) and latent component variables (η_q measured with $X_1, \ldots, X_P$). According to the rules described above, the manifest component variables $Y_{q \neq r}$ have free covariances $Cov(Y_{q \neq r}, \varepsilon_i)$ with the measurement residuals ε_i of the manifest indicators X_i of the latent component variables. However, with this specification, the PIM is not identified. This model can be estimated only by fixing the covariances $Cov(Y_{q \neq r}, \varepsilon_i)$ of at least one manifest indicator X_i to zero, which again refers to the assumption of measurement invariance (i.e., conditional stochastic independence assumption of the manifest indicators X_i of the manifest component variables Y_q given the latent component variables η_q). Note that the restriction $Cov(Y_{q \neq r}, \varepsilon_i) = 0$ does not conflict with the first guiding rule for constructing model-based composite scores (i.e., the preservation of the target model) if the assumption of measurement invariance holds true. The specification of a PIM with a composite score of manifest and latent component variables is presented in Section 3.9 of the supplemental PDF document (RoseEtal2019suppl.pdf).

Assessing Model fit of SEM with Model-based composite scores

Assessing model fit is essential in application of SEM. As outlined above the χ^2-test, the RMSEA, and the test of close fit (H_0: RMSEA ≤ 0.05) is correct when latent or manifest model-based composite score are included in the model by means of the PIM. The same is true for Hoelter's critical N. Fortunately, all the remaining fit indices are also trustworthy for SEM with only latent model-based composite scores that are specified by means of the PIM. The commonly used fit indices are equal to that of an equivalent model with all latent component variables but without the latent model-based composite scores. This is implied by the guiding principles for the construction of the composite scores as outlined above. However, in the case of manifest model-based composite scores almost all absolute, incremental, and parsimony fit measures (except the χ^2-test, the RMSEA, the test of close fit (H_0: RMSEA ≤ 0.05), and Hoelter's critical N) are not trustworthy.

Correct Incremental Fit Indices for SEM with Manifest Model-based Composite Scores

The problem is essentially the same as in the case of including auxiliary missing-data-relevant variables in SEM (Graham, 2003). The manifest component variables are not part of the target model but are just auxiliary for constructing the model-based composite score. However they are also part of the baseline model that is used to compute the incremental fit indices. In the baseline model all variables including the component variables are assumed to be uncorrelated. This is inappropriate for evaluating the fit of the target model, which is just a sub-model of the joint model with the PIM. Therefore, the values of incremental fit indices, such as the Tucker-Lewis-Index (TLI, Tucker & Lewis, 1973), can be very low or even negative in SEM with manifest model-based composite scores. In order to obtain proper incremental fit measures an alternative baseline model M_0 needs to be estimated. M_0 comprises all manifest variables of the target model X_1, ..., X_P, the manifest component variables Y_1, ..., Y_Q, as well as the manifest model-based composite score C. Hence, the PIM needs to be specified within the baseline model M_0. The further specifications of M_0 are:

- According to the described rules for the specification of the PIM, the covariances between the manifest component variables $Y_{q \neq r}$ and the model-based composite score C have free covariances, and the residual variance of the pseudo-indicator variable Y_r is fixed to zero.
- All manifest component variables $Y_{q \neq r}$ have free covariances among each other.
- All manifest component variables $Y_{q \neq r}$ have free covariances with the manifest variables X_1, ..., X_P.
- The covariances between the manifest variables X_1, ..., X_P are fixed to zero.
- The covariances between the manifest variables X_1, ..., X_P and the model-based composite score C are fixed to zero.
- The variances of all manifest variables $Y_{q \neq r}$ and X_1, ..., X_P and the variance of the model-based composite score are freely estimated.

Let M_1 denote the analytical model that combines the PIM for constructing the composite score and the target-model. The χ^2-value and the degrees of freedom of M_1 are denoted by χ_1^2 and df_1. For computing correct incremental fit indices of the target model, which is part of M_1, such as the comparative fit index (CFI; Bentler, 1990), the TLI (Tucker & Lewis, 1973), the normed fit index (NFI; Bentler & Bonett, 1980), Bollen's incremental fit index (IFI; Bollen, 1989), the relative noncentrality index (RNI; McDonald & Marsh, 1990), the standard formulas of these indices can be used with the χ^2-values χ_0^2 and χ_1^2 and the degrees of freedom df_0 and df_1. In the case of no missing values and using the standard ML estimator, the resulting incremental fit indices are equal to those of the target model with the a priori computed composite score instead of the model-based composite score. This is not necessarily the case when adjusted χ^2-values are used that are computed with scaling correction factors obtained by robust estimation methods. The same is true when robust incremental fit indices are computed as proposed by (Brosseau-Liard & Savalei, 2014). The reason is that the scaling correction factors of M_0 with the model-based composite score and the target model with the a priori computed composite score can be different. To our experience these differences are negligible and do not restrict the usability of the fit indices.

Correct Residual-based Fit Indices for SEM with Manifest Model-based Composite Scores

The root mean residual (RMR) or the standardized root mean residual (SRMR) are popular and widely used absolute fit measures due to their good interpretability. Both fit indices must also be corrected in the case of overidentified SEM with model-based composite scores. The reason is that the inclusion of the manifest component variables of the composite score without restrictions regarding their variances and the covariances with other manifest variables increases the proportion of well-fitting model-implied variances and covariances. Therefore, the RMR and SRMR tend to be smaller than the RMR and SRMR of the target model only (without the manifest component variables). However, both fit indices can be computed using only the variance and covariances of the variables of the target model while ignoring all variances and covariances of the manifest component variables. The computation of these corrected RMR and SRMR requires not only the unrestricted estimates of the variances s_{ii} and covariances s_{ij} among all manifest variables X_1, ..., X_P of the target model but also the estimated variance s_c^2 of the model-based composite score C and the P unrestricted covariances s_{ci} between the manifest variables X_1, ..., X_P and the composite score C. If the manifest component variables Y_1, ..., Y_Q have missing values, then s_c^2 and s_{ci} of the unrestricted model can only be obtained by means of a saturated model that combines the PIM with model-based composite score with the saturated model of the variables X_1, ..., X_P. This model is denoted by M_2 in the remainder and requires the following specifications:

- According to the described rules for the specification of the PIM, the covariances between the manifest component variables $Y_{q \neq r}$ and the model-based composite score C have free covariances, and the residual variance of the pseudo-indicator variable Y_r is fixed to zero.
- All manifest component variables $Y_{q \neq r}$ have free covariances among each other.
- All manifest component variables $Y_{q \neq r}$ have free covariances with the manifest variables X_1, ..., X_P.
- The covariances between the manifest variables X_1, ..., X_P are free estimands.
- The covariances between the manifest variables X_1, ..., X_P and the model-based composite score C are freely estimated.
- The variances of all manifest variables $Y_{q \neq r}$ and X_1, ..., X_P and the variance of the model-based composite score are freely estimated.

Appropriate RMR and SRMR values of SEM with manifest model-based composite scores can be computed based on the following Formulas (Equations 7 and 8):

$$RMR = \sqrt{\frac{\sum_{i=1}^{P}\sum_{j=1}^{i}\left(s_{ij}-\hat{\sigma}_{ij}\right)^2+\sum_{c=1}^{R}\left(s_c^2-\hat{\sigma}_c^2\right)+\sum_{c=1}^{R}\sum_{i=1}^{P}\left(s_{ic}-\hat{\sigma}_{ic}\right)^2}{\left(P+R\right)\left(P+R+1\right)/2}}. \tag{7}$$

$$SRMR = \sqrt{\frac{\sum_{i=1}^{P}\sum_{j=1}^{i}\left[\left(s_{ij}-\hat{\sigma}_{ij}\right)^2/\sqrt{s_{ii}s_{jj}}\right]+\sum_{c=1}^{R}\left[\left(s_c^2-\hat{\sigma}_c^2\right)/s_c^2\right]+\sum_{c=1}^{R}\sum_{i=1}^{P}\left[\left(s_{ic}-\hat{\sigma}_{ic}\right)^2/\sqrt{s_{ii}s_c^2}\right]}{\left(P+R\right)\left(P+R+1\right)/2}} \tag{8}$$

The model-implied variances $(\hat{\sigma}_{ii}, \hat{\sigma}_c^2)$ and covariances $(\hat{\sigma}_{ij}, \hat{\sigma}_{ic})$ are obtained by the model M_1. The Equations 7 and 8 do not differ from the standard formulas of the RMR and SRMR. The only difference is that not only the variances and covariances of the manifest variables X_1, ..., X_P are considered but also the variances of the model-based composite score and the covariances between X_1, ..., X_P and the model-based composite score. However, all variances and covariances of the manifest component variables Y_1, ..., Y_Q are excluded. Note that the two Equations 7 and 8 allow for the general case of $R \geq 1$ manifest model-based composite scores variables. It should also be noted that different versions of the SRMR exist. The corrected SRMR in Equation 8 is based on the definition of SRMR by (Hu & Bentler, 1999). There exist alternative versions of the SRMR, such as implemented in M*plus* (Asparouhov, 2018), that can also be computed based on the model M_2.

There are many more measures that are used for assessing model fit or for model selection, such as the goodness of fit index (GFI) and the adjusted GFI (AGFI) by Jöreskog and Sörbom (1981), or the parsimony fit indices proposed by James, Mulaik, and Brett (1982). These measures are also affected by the inclusion of the manifest component variables as auxiliary variables. The AIC, the BIC as well as other information criteria do increase due to the larger number of manifest variables and increased model complexity in SEM with manifest model-based composite scores. How to adjust these measures for using model-based composite scores needs to be addressed in future research.

Simulation Study

In order to demonstrate the utility of manifest model-based composite scores in the case of missing values, we conducted a small simulation study in which we considered a simple linear regression $E(C_Y|C_X) = \alpha + \beta C_X$. We focused on the point estimates of the regression coefficient β, with the true value $\beta = 2$, and the estimated standard errors $SE(\hat{\beta})$. The R syntax used for the simulation study is provided as a part of the supplemental material to this paper (https://osf.io/b6h7n/).

Design and Data Generation

In this example, the predictor variable C_X is the average score of the variables X_1 to X_4, each with a normal distribution $N(\mu_i, 1)$. The means of the variables were $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 4$, and $\mu_4 = 6$. The pairwise correlations of all X variables were set to 0.5. The outcome C_Y is also an average score of the observed variables Y_1 to Y_4. Each Variable Y_1 was simulated on the basis of the average score C_X, so that the true value $\beta = 2$ was implied. The implied correlations between the Y-variables ranged from $r = .23$ to $r = .67$. We studied the performance of model-based composite scores under different conditions. Three factors were systematically varied: (a) sample size, (b) overall proportion of missing data, and (c) the missing data mechanism. Two sample sizes were considered: a fairly small ($N = 200$) and a large sample size ($N = 2{,}000$). Three overall proportions of missing data were simulated: small (10%), moderate (25%), and large (50%) proportions of missing values in the component variables X_2, X_4, Y_2, and Y_4. Hence, each simulated case had at least two (X_1 and X_3) and a maximum of four observed X-variables (X_1 to X_4) that were averaged to determine C_X. Similarly, between two (Y_1 and Y_3) and four observed Y-variables (Y_1 to Y_4) were averaged to determine C_Y for each simulated case. Thus, only the observed values were aggregated into the average scores. We simulated two missing data mechanisms according to Rubin's classification of missing values (R. J. A. Little & Rubin, 2002; Rubin, 1976). First, we generated data with missing values that were missing completely at random (MCAR). We additionally simulated data with missing values that were missing at random (MAR). In the MAR condition, the probability that the component variables X_2, X_4, Y_2, and Y_4 were observed or missing was stochastically dependent on the other observed component variables X_1, X_3, Y_1, and Y_3. Specifically, the response probabilities of the four component variables X_2, X_4, Y_2, and Y_4 were determined by logit regression models, given by

$$\begin{aligned}
&P\left(X_2 \text{ not missing} \mid X_1\right) = F\left(c + z\left(X_1\right)\right)\\
&P\left(X_4 \text{ not missing} \mid X_1, X_3\right) = F\left(c + z\left(X_1 + X_3\right)\right)\\
&P\left(Y_2 \text{ not missing} \mid X_1, Y_1\right) = F\left(c + z\left(-X_1 + X_1Y_1\right)\right)\\
&P\left(Y_4 \text{ not missing} \mid X_3, Y_3\right) = F\left(c + z\left(-X_3 + X_3Y_3\right)\right).
\end{aligned} \tag{10}$$

The constant c was chosen such that the overall proportions of missing data were 10% ($c = 2.2$), 25% ($c = 1$), and 50% ($c = 0$). The operator z(.) stands for z-standardization and F(.) denotes the logistic distribution function. Hence, the linear combination of predictors in the logit regressions were standardized to have a mean of zero and a variance of one. This allowed us to avoid a situation in which missingness had too strong of a dependency on the missing related variables X_1 and X_3, especially when interactions were involved, and allowed us to have better control over the overall proportion of missing data. We compared two estimation procedures. First, the ordinary least squares (OLS) estimator, based on the regression of the manifest average scores C_Y and C_X, which were computed on the basis of the observed variables only. Therefore, we expected biased estimation of the regression coefficient β. Second, we used SEM with the model-based average scores C_Y and C_X as pseudo-latent variables defined in a PIM. The component variables X_1 and Y_1 were chosen as the pseudo-indicator variables of C_Y and C_X. FIML was used for parameter estimation. FIML was expected to yield unbiased parameter estimates under MCAR as well as MAR. The bias $E(\hat{\beta} - \beta)$, the relative bias $E(\hat{\beta} - \beta) / \beta$, the mean squared error $E[(\hat{\beta} - \beta)^2]$, and the coverage of the estimated regression coefficient were used as criteria. The coverage is the proportion of trials that yielded parameter estimates whose 95% confidence intervals included the true coefficient β. Therefore, 95% coverage rates are expected in the case of unbiased estimation of the parameter and its standard error. We additionally compared estimated standard errors and the empirical standard deviations $s(\hat{\beta})$ of the estimated regression coefficients to study potentially biased estimation of the standard errors.

Results

The results of the simulation study are summarized in **Table 2**. In all conditions, the estimated regression coefficient is biased when OLS with average scores of incomplete data was used such that the coverages were zero or close to zero in all conditions. Only in the case of N = 200 with 10% missing values that were MCAR, the coverage reached 0.222, which is far from being acceptable. The amount of bias depends on the overall proportion of missing values and the missing data mechanism. Under the simulated conditions the bias was higher when the missing data mechanism was MAR instead of MCAR and increased with the raising proportions of missing data. The relative bias ranged from –13.5% to –47.8%. Surprisingly, the regression coefficient was substantially underestimated even when the missing data mechanism was MCAR. However, missingness in the component variables is a source of construct-irrelevant variance. So the variance in the average scores of incomplete data is higher compared to the variance of the average scores of the complete data. Therefore, the correlation of two composite scores of incomplete data is lower than the correlation of the composite scores of complete data, which also leads to the underestimated regression coefficient. In our simulation the problem was increased by the mean differences between the component variables.

However, the bias vanished in all conditions in which model-based average scores were used with SEM. The bias, the relative bias, and the coverage of SEM with model-based composite scores were almost identical to the OLS with complete data. The MSE of $\hat{\beta}$ was slightly higher in the SEM with model-based average scores in small samples (N = 200) with 25% and 50% missing values in the component variables X_2, X_4, Y_2, and Y_4. This was due to larger standard errors resulting from the loss of information from missing data (see **Table 3**). In general, the empirical standard deviations $s(\hat{\beta})$ of the estimates $\hat{\beta}$ and the estimated standard errors in the SEM with model-based average scores were very close in all conditions, thus implying no systematic bias in the standard errors when model-based average scores were used in the SEM. The standard errors of the regression coefficients obtained with OLS based on average scores with incomplete data showed slightly higher deviations from the empirical variation of the estimates in some condition with small samples (N = 200). However, in light of the severely biased point estimates, this is negligible.

Applied Data Example

Sample

To illustrate the application of model-based composite scores to a real data example,[2] we present data from a study on gender differences in students' value beliefs about mathematics (Gaspard et al., 2015). The sample consisted of N = 1,868 German ninth-grade students from 82 classes in 25 academic-track secondary schools in the German state of Baden-Württemberg.

Models

According to Expectancy-Value Theory (Eccles et al., 1983), four value components can be distinguished: intrinsic value, attainment value, utility value, and cost. In our data example, we focus solely on the utility value component, which is the perceived usefulness of engagement and achievement for specific short- and long-term goals. As in Gaspard et al. (2015), the utility value component of mathematics consists of five subfacets that were measured with a total of 12 items Y_1–Y_{12}. The subfacets are: (a) utility for school (η_{sch}; usefulness of mathematics for one's present and future education), (b) utility for job (η_{job}; importance of mathematics for one's future career opportunities), (c) general utility of mathematics for future life (η_{fu}), (d) utility for daily life (η_{dl}; value of mathematics for one's daily routines and leisure time activities), and (e) social utility of mathematics knowledge for being accepted by peers (η_{soc}). All subfactes were positively correlated between $r(\eta_{job}, \eta_{soc})$ = .27 and $r(\eta_{fu}, \eta_{dl})$ = .86. In order to have a summary measure of students' beliefs about the utility value of mathematics, a sum score might be a viable choice. In order to account for measurement error in the utility dimensions, we formed a latent model-based sum score. It is important to note that the variances of the component variables can act like weights in linear combinations. That is, a variable with a larger variance contributes more to the variance of the composite score than the same rescaled variable with a lower variance. In order to avoid an unintended weighting due to differences in the variances of the latent variables η_q, we constructed a weighted sum score, $Utility = SD(\eta_{sch})^{-1}\eta_{sch} + SD(\eta_{job})^{-1}\eta_{job} + SD(\eta_{soc})^{-1}\eta_{soc} + SD(\eta_{fu})^{-1}\eta_{fu} + SD(\eta_{dl})^{-1}\eta_{dl}$, with the inverse standard deviations $SD(\eta_q)^{-1}$ as weights. This is conceptually equivalent to the unweighted sum score of standardized variables η_q with $Var(\eta_q) = 1$, for all $q = 1, ..., Q$.

Table 2: Bias, Relative Bias, MSE, and Coverage of the Estimated Regression Coefficients of the Simple Linear Regression $E(C_Y|C_X)$ Using OLS with Complete Data, OLS with Incomplete Data, and SEM with Model-Based Composite Scores for Incomplete Data.

Design			**Complete data**				**Incomplete data**							
			OLS				**OLS**				**SEM with model-based average scores**			
N	**MDM**	**Missing rate**	**Bias**	**Relative bias**	**MSE**	**Coverage**	**Bias**	**Relative bias**	**MSE**	**Coverage**	**Bias**	**Relative bias**	**MSE**	**Coverage**
200	MCAR	10%	0.001	0.000	0.005	0.953	–0.269	–0.135	0.083	0.222	0.002	0.001	0.006	0.957
200	MCAR	25%	0.000	0.000	0.006	0.934	–0.527	–0.264	0.290	0.005	0.001	0.001	0.007	0.948
200	MCAR	50%	–0.001	0.000	0.006	0.935	–0.637	–0.319	0.420	0.000	0.000	0.000	0.011	0.945
200	MAR	10%	0.002	0.001	0.005	0.945	–0.545	–0.273	0.306	0.000	0.006	0.003	0.007	0.940
200	MAR	25%	–0.003	–0.001	0.006	0.950	–0.836	–0.418	0.707	0.000	–0.002	–0.001	0.008	0.951
200	MAR	50%	0.000	0.000	0.006	0.954	–0.955	–0.478	0.920	0.000	0.005	0.003	0.012	0.933
2000	MCAR	10%	–0.002	–0.001	0.001	0.949	–0.273	–0.136	0.075	0.000	–0.001	–0.001	0.001	0.950
2000	MCAR	25%	0.001	0.001	0.001	0.960	–0.525	–0.262	0.276	0.000	0.002	0.001	0.001	0.950
2000	MCAR	50%	–0.001	0.000	0.001	0.940	–0.635	–0.318	0.405	0.000	0.000	0.000	0.001	0.942
2000	MAR	10%	0.001	0.000	0.001	0.944	–0.545	–0.273	0.298	0.000	0.001	0.001	0.001	0.935
2000	MAR	25%	0.000	0.000	0.001	0.964	–0.835	–0.418	0.698	0.000	0.000	0.000	0.001	0.945
2000	MAR	50%	–0.001	0.000	0.001	0.952	–0.950	–0.475	0.903	0.000	–0.001	0.000	0.001	0.950

Note: MDM = Missing data mechanism; OLS = Ordinary least squares; MSE = Mean squared error.

Table 3: Empirical Standard Deviations $s(\hat{\beta})$ of the Estimated Regression Coefficients, the Mean Estimated Standard Errors of the Regression Coefficients $\hat{\beta}$, and the Bias of the Estimated Standard Errors $SE(\hat{\beta})$.

Design			**Complete data**			**Incomplete data**					
			OLS			**OLS**			**SEM with model-based average scores**		
N	**MDM**	**Missing rate**	$s(\hat{\beta})$	$\overline{SE}(\hat{\beta})$	**Bias**	$s(\hat{\beta})$	$\overline{SE}(\hat{\beta})$	**Bias**	$s(\hat{\beta})$	$\overline{SE}(\hat{\beta})$	**Bias**
200	MCAR	10%	0,075	0,074	0,001	0.096	0.103	–0.007	0.078	0.077	0.001
200	MCAR	25%	0,076	0,078	–0,002	0.111	0.110	0.002	0.085	0.085	0.000
200	MCAR	50%	0,075	0,079	–0,004	0.123	0.119	0.004	0.101	0.103	–0.002
200	MAR	10%	0,075	0,074	0,001	0.085	0.095	–0.010	0.080	0.081	–0.001
200	MAR	25%	0,075	0,077	–0,001	0.086	0.090	–0.003	0.089	0.088	0.001
200	MAR	50%	0,075	0,075	0,000	0.091	0.091	0.001	0.103	0.109	–0.006
2000	MCAR	10%	0,024	0,023	0,000	0.030	0.033	–0.003	0.025	0.024	0.001
2000	MCAR	25%	0,024	0,023	0,000	0.035	0.035	0.000	0.027	0.026	0.001
2000	MCAR	50%	0,024	0,025	–0,001	0.039	0.037	0.002	0.032	0.033	–0.001
2000	MAR	10%	0,024	0,024	0,000	0.027	0.032	–0.005	0.025	0.026	–0.001
2000	MAR	25%	0,024	0,023	0,001	0.027	0.028	–0.001	0.028	0.028	0.000
2000	MAR	50%	0,024	0,024	0,000	0.029	0.028	0.000	0.032	0.033	–0.001

Note: MDM = Missing data mechanism; OLS = Ordinary least squares; MSE = Mean squared error.

The standard deviations $SD(\eta_q)$ were estimated with an equivalent initial model without the weighted latent sum score. One factor loading and one measurement intercept of each dimension η_q were fixed to one in order to identify the model. The resulting weights were $SD(\eta_{sch})^{-1} = 2.411$, $SD(\eta_{job})^{-1} = 1.814$, $SD(\eta_{soc})^{-1} = 1.857$, $SD(\eta_{fu})^{-1} = 1.501$, and $SD(\eta_{dl})^{-1} = 1.429$. In order to specify the latent composite score *Utility*, η_{fu} was chosen as the pseudo-indicator variable with the factor loading fixed to $\gamma_r^{-1} = SD(\eta_{fu}) = 0.666$ and the intercept fixed to zero. The path coefficients of the paths from the component variables $\eta_{q \neq r}$ to the reference variable η_r were fixed to the ratios of the weights $-\gamma_{q\neq r}\gamma_r^{-1}$ (i.e., $\eta_{sch} \rightarrow \eta_{fu} = -1.607$, $\eta_{job} \rightarrow \eta_{fu} = -1.209$, $\eta_{sco} \rightarrow \eta_{fu} = -1.237$, $\eta_{dl} \rightarrow \eta_{fu} = -0.952$). The measurement model of the five utility dimensions and the PIM with the latent composite score *Utility* can be seen on the right side of **Figure 3**. The latent utility index served as the outcome in a multiple latent regression on gender (*Sex;* 0 = female, 1 = male), grade in mathematics ($Grade_{math}$; ranging from 1 = best grade to 6 = worst grade), cognitive ability (*CA*), and socioeconomic status (see top left of **Figure 3**). Students' cognitive ability was assessed with a German test of verbal and figural reasoning, which is an indicator of fluid intelligence (Heller & Perleth, 2000). In our model, the socioeconomic status variable (SES_{comp}) is the model-based sum score (see bottom left of **Figure 3**) of eight variables: The Socio-Economic Index of Occupational Status (ISEI; Ganzeboom, Degraaf, Treiman, & Deleeuw, 1992) of the student's mother (X_1) and father (X_2), highest educational degree of the mother (X_3) and father (X_4), professional qualification of the mother (X_5) and father (X_6), the number of books at home (X_7), and the family's income (X_8). All SES indicators suffered from substantial proportions of missing values that ranged from 8.9% (X_2) to 46.7% (X_8). FIML was used for parameter estimation with missing data (R. J. A. Little & Rubin, 2002; Rubin, 1976). X_1 was used as the pseudo-indicator variable of the index variable SES_{comp}. All *SES* markers X_1–X_8 were *z*-standardized to avoid unintended differential weighting due to differences in the variances of the constituting variables. The resulting model with two model-based composite scores is graphically represented by the path diagram in **Figure 3**. As the number of covariances in this model was large (165 covariances), they are not shown in **Figure 3**. Here, we list only the covariances that were free estimands in this model according to the rules described previously: (a) all covariances of the manifest component variables $X_{q\neq r}$ of the SES composite score and the measurement residuals in the measurement model of the latent utility dimensions, (b) the covariances between the manifest component variables $X_{q\neq r}$ of the SES composite score and the manifest exogenous variables $Grade_{math}$, SES_{comp}, *Sex*, and *CA*, (c) the covariances between the manifest component variables $X_{q\neq r}$ and the SES composite score, (d) the covariances between the latent component variables $\eta_{q\neq r}$ and the exogenous variables $Grade_{math}$, SES_{comp}, *Sex*, and *CA*, (e) the covariances between the latent component variables $\eta_{q\neq r}$ and the residual of the latent composite score *Utility*, and (f) the covariances between the predictors $Grade_{math}$, SES_{comp}, *Sex*, and *CA*. We call this model Model A.

We also estimated a second version of the model with the additional assumption that the latent utility dimensions demonstrate measurement invariance (Model

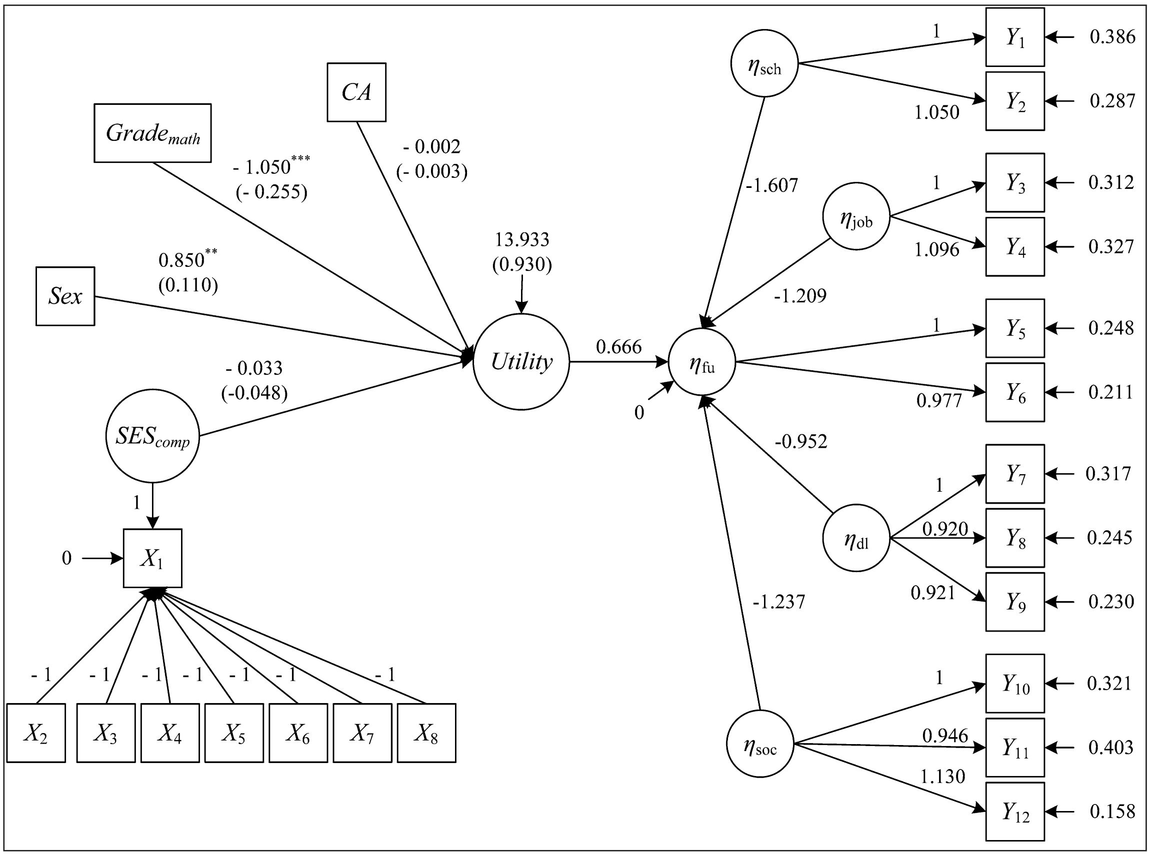

Figure 3: Path diagram of an SEM with the model-based latent weighted sum score *Utility* as the outcome. The predictor variable SES_{comp} is a manifest model-based sum score. Standardized coefficients are presented in parentheses.[3]

B). If measurement invariance holds true, the rules for constructing model-based composite scores can be relaxed. Specifically, the covariances between the manifest component variables $X_{q \neq r}$ of the SES composite score and the measurement residuals can be fixed to zero. Instead, the covariances between $X_{q \neq r}$ and the latent component variables $\eta_{q \neq r}$ and the residual of the latent composite score *Utility* were freely estimated. This model was more parsimonious. The degrees of freedom increased from $df = 72$ to $df = 121$ as the number of estimated covariances was reduced to 116. We applied a scaled χ^2 difference test (Satorra & Bentler, 2010) to compare the two models to test for measurement invariance.

For the two versions of the model, we also present the equivalent models without the two model-based composite scores that can be used to check that the model is specified correctly. A model that is equivalent to Model A without the composite scores includes the manifest component variables X_1 to X_8 of the SES composite score and the measurement model of the five latent math utility dimensions. The variables X_1 to X_8 have free covariances with (a) each other, (b) the measurement residuals of the indicators of the latent math utility dimensions, and (b) the manifest covariates $Grade_{math}$, SES_{comp}, *Sex*, and *CA*. If the covariances of variables X_1 to X_8 with the measurement residuals are fixed to zero, whereas the covariances of X_1 to X_8 with the five latent utility variables are freely estimated, the result will be a model that is equivalent to Model B.

We used M*plus* 8 (Muthén & Muthén, 1998–2018) for the data analyses with robust ML estimation and Yuan and Bentler (1998) robust χ^2 statistics for nonnormal data. We also report the RMSEA, the corrected CFI, the corrected TLI, and the corrected SRMR as descriptive fit indices. The corrected fit measures are chosen because a manifest model-based composite score (SES_{comp}) is part of the model. The eight manifest component variables X_1 to X_8 are not part of the target model. They only serve as auxiliary variables in the PIM and lead to an improper baseline and unrestricted model for computing incremental and absolute fit indices of Model A and B. Therefore, we fitted the alternative baseline model (M_0) and the alternative unrestricted model (M_2) and calculated the corrected fit measures as described in the Section *Assessing Model fit of SEM with Model-based composite scores.* The hierarchical data structure in which students (Level 1) were nested in classes (Level 2), which were nested in schools (level 3), was taken into account by the sandwich estimator of the variance-covariance matrix (i.e., type = complex in M*plus* 8; Muthén & Satorra, 1995). The M*plus* input file is presented in Appendix B with additional explanations. We also present the R syntax for the models that can be estimated using the lavaan package, which yields the same point estimates. However, the multilevel structure of the data with missing values cannot be taken into account very easily in lavaan. Therefore, different standard errors and the model fit statistics are obtained in lavaan by ignoring the nested data structure.

Results

In our data example, there was a subsample of N = 650 participants with complete data. To check that the Models A and B were correctly specified, we also fit the two equivalent models to the data. The χ^2–values, the degrees of freedom, the p-values and the RMSEA of the equivalent models were identical to those of Model A or B. Furthermore, the model-implied variance-covariance structures and mean structures of the variables included in the two models were identical. Hence, the manifest component variables served only as auxiliary variables, and the construction of the latent sum score did not imply any additional restrictions. Although the χ^2 tests for both Models A and B were significant (Model A: χ^2 = 255.806, df = 72, p < .001; Model B: χ^2 = 308.706, df = 121, p < .001), the remaining descriptive Gof criteria indicated a good model fit for Model A (RMSEA = .037, CFI = 0.975, TLI = 0.959, SRMR = .027) and an even better fit for the more parsimonious Model B (RMSEA = .029, CFI = 0.975, TLI = 0.975, SRMR = .027). The lower value of the RMSEA for the more restrictive Model B as well as the nearly identical fit indices (CFI, TLI, and SRMR) were in line with the nonsignificant scaled χ^2-difference test of the two models (scaled χ^2-difference = 45.612, df = 49, p < .611). Hence, the hypothesis of measurement invariance did not need to be rejected. Therefore, we decided to present the results of Model B in the remainder of this section.

The model-implied variance of the latent utility index was s^2 = 14.982, which was exactly the weighted sum of the variances and two times the covariances of the five latent component variables.[4] Similarly, the model-implied variance of the manifest model-based SES_{comp} was s^2 = 31.899, which was also identical to the sum of the variance and two times the covariance of the manifest component variables. This also confirmed that the model was correctly specified. The proportion of variance that was explained in the latent math utility index was R^2 = 0.07, with gender (b = 0.850, SE = 0.244, p < .001) and math grade (b = –1.050, SE = 0.101, z = –10.343, p < .001) as significant predictors. The cognitive ability test (b = –0.002, SE = 0.022, p = .911) and the model-based manifest SES index were not significant (b = –0.033, SE = 0.019, p = .075) given the other predictors in the regression model. Thus, conditional on gender, students with better grades (i.e., lower values on the variable $Grade_{math}$) have on average higher beliefs about the utility value of mathematics. Furthermore, boys have on average higher utility value scores than girls given students' math grade.

Discussion

Composite scores are widely used in psychology, sociology, educational research, medical research, and many other areas in the social and behavioral sciences. In this paper, we demonstrated the model-based construction of composite scores of manifest and latent variables in SEM by means of the PIM. Model-based composite scores are not computed prior to the analyses but are constructed in the SEM. They are valuable for accounting for ignorable missing data in the constituting variables of the composite score because FIML can be used for parameter estimation (Arbuckle, 1996; Enders, 2001; Schafer & Graham, 2002). Model-based composite scores of latent variables are useful for adjusting for measurement errors in the components of the composite score. They can be used to model manifest and latent summary measures as independent or dependent variables or as covariates in SEM.

The PIM we introduced is quite flexible. As we demonstrated with our data example, multiple manifest and latent composite scores can be included as independent or dependent variables in SEM. Even dichotomous manifest variables can be used as component variables of manifest model-based composite scores. The PIM using ML estimation yields exactly the same model fit (χ^2, df, p-value, and RMSEA) as obtained from fitting the target model with the a prior computed composite scores of the binary variables. More examples that demonstrate how to use model-based composite scores in SEM are provided in the Supplemental Material of this article. There, we also demonstrate that manifest and latent variables can be jointly aggregated into model-based composite scores. This may be a viable choice if some of the component variables are free from measurement error, whereas others are not. Fortunately, the PIM can be used without limitations in existing extensions and variants of SEM, including multiple-group and mixture SEM and SEM for categorical variables, which are based on tetrachoric or polychoric correlations.

The focus of this article was on demonstrating how to construct model-based composite scores. We illustrated the adequacy and the value of using composite scores in SEM with only a few examples. For example, instead of using item parcels as indicators of latent variables, an item factor analysis with the single items as indicators can be used. Similarly, instead of latent composite scores, second-order factor models might be a viable choice. In our applied data example, we used the latent composite score of the five latent utility dimensions as a summary measure of students' beliefs about the utility value of mathematics. This is quite different from modeling math utility as a second-order factor. In contrast to latent composite scores, second-order factor models rest on assumptions about the variance-covariance structure of the first-order factors. Furthermore, the correlations between the first-order factors with the second-order factor can differ substantially from the correlations between the same first-order factors with a composite score. Similarly, the correlations between a second-order factor with external variables can be quite different from the correlations between latent composite scores with these external variables. This highlights the idea that higher order factors and composite scores of latent variables differ fundamentally in their meaning, interpretation, and underlying measurement perspective. Therefore, they should not be carelessly viewed as interchangeable. Composite scores of manifest and/or latent variables can be meaningful and valuable (e.g., average latent proficiency in multidimensional scholastic achievement tests) as independent or dependent variables. However, a (weighted) sum score or the average score of the six interest dimensions (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional) from Holland's RIASEC model, for example, hardly makes sense as the essential information is given by the individual interest profiles. In our view, there are no universal answers to the question of when to use composite scores and when not to. This issue is part of a more general debate about the theoretical status of manifest and latent variables and measurement in general (Bollen, 2002, 2011; Bollen & Bauldry, 2011; Borsboom, Mellenbergh, & Van Heerden, 2003; Grace & Bollen, 2008; Markus & Borsboom, 2013). With this paper, we did not aim to contribute to this very fundamental and in parts controversial and ongoing debate. Our contribution is more modest because we focused on examining the model-based construction of composite scores. With the proposed method, we provide a promising approach for whenever SEM is used with composite variables that cannot simply be computed prior to the analysis. However, researchers need to consider the adequacy and tenability of using manifest and latent composite scores by considering the particular study and the specific research question at hand.

There are some limitations and open questions. First, the approaches proposed here are limited to composite scores that are linear combinations of their constituting variables. There might be nonlinear composite scores that cannot be handled with the models proposed here. A disadvantage of the PIM is the increase in the complexity of the model that comes from including model-based composite scores in comparison with equivalent models with composite scores that are computed a priori. For example, the number of variables increases as the constituting variables remain part of the model in addition to the composite score. Furthermore, additional variables (e.g., the auxiliary factors) or additional paths that are not of substantial interest need to be specified. The correct specification of the covariance structure between the composite scores, their constituting variables, the auxiliary factors, and all covariates is essential for avoiding biased parameter estimation. Therefore, a thorough model specification is important. We recommend that a stepwise procedure be followed for model building. In a first step, an equivalent model should be estimated without the composite scores. In a second step, the composite scores can be included. The two models must be equivalent in terms of model fit. Deviations indicate misspecifications or convergence to local minima.

In our experience, the PIM can sometimes suffer from convergence problems, which can be solved by providing better starting values. We often found that higher starting values for the variances of composite scores are helpful because these variances are typically much larger than the variances of the component variables. This is another reason why we recommend the stepwise procedure. Parameter estimates of an equivalent initial model without the latent composite score can be used as starting values, or this can allow the user to derive good starting values for the SEM that includes the PIM in the second run.

Savalei and Rhemtulla (2017) showed that the two-stage maximum likelihood estimation method for ignorable missing data in manifest component variables of item parcels performed well in terms of unbiasedness, efficiency, and coverage. The performance of the PIM with manifest model-based composite scores estimated by FIML compared with the two-stage ML estimator or multiple imputation is an interesting open research question.

It is still unclear whether the proposed approach for model-based composite scores can be extended or adapted for nonlinear SEM with quadratic or cubic relations between latent variables and/or latent interaction effects (Klein & Moosbrugger, 2000; Klein & Muthén, 2007). Brandt (2008) proposed composite scores of latent variables as a promising approach in multidimensional IRT models. Whether the parameters of the PIM can be estimated with commonly used marginal ML or Bayesian estimation methods for multidimensional IRT models is also an interesting question for future research.

Data Accessibility Statement

The data of the applied data example used for illustration of model-based composite scores in this article as well as a simulated data set, and the accompanying documentation are available from the Open Science Framework (https://osf.io/b6h7n/).

Appendix A

In this Appendix, the validity of the model equations of the pseudo-indicator variable Y_r is shown for a PIM with the weighted sum score $C = \sum_{q=1}^{Q} \gamma_q Y_q$ in order to demonstrate correct model specification. This is achieved by rearranging and simplifying the last line of Equation 5, which is the model equation of the pseudo-indicator variable Y_r, which yields:

$$\begin{aligned} Y_r &= \frac{1}{\gamma_r} C + \sum\nolimits_{q \neq r} -\frac{\gamma_q}{\gamma_r} Y_i \\ &= \frac{1}{\gamma_r}\left(\sum\nolimits_{q=1}^{Q} \gamma_q Y_q\right) + \sum\nolimits_{q \neq r} -\frac{\gamma_q}{\gamma_r} Y_i \\ &= \sum\nolimits_{q=1}^{Q} \frac{\gamma_q}{\gamma_r} Y_q + \sum\nolimits_{i \neq r} -\frac{\gamma_q}{\gamma_r} Y_q \\ &= \sum\nolimits_{q \neq r} \frac{\gamma_q}{\gamma_r} Y_q - \sum\nolimits_{q \neq r} \frac{\gamma_q}{\gamma_r} Y_q + \frac{\gamma_r}{\gamma_r} Y_r \\ &= Y_r. \end{aligned} \tag{11}$$

Appendix B

This Appendix presents the specification of the model for the data example in M*plus*. We fit two versions of the model. Model A included the manifest component variables without assuming measurement invariance. Model B was more restrictive because measurement invariance was assumed. Here, we first present the M*plus* input file for Model A:

```
1  DATA:  file = GaspardEtal2015mathUtility.dat;
2  VARIABLE: names are id school class sex math ca
3         isei_m isei_d edu_m edu_d voc_m voc_d book income y1–y12;
4         usevariables are sex math ca isei_m isei_d edu_m edu_d
5         voc_m voc_d book income y1–y12;
6         Missing are all (–99); cluster = class; stratification = school;
7  ANALYSIS: TYPE = COMPLEX; ESTIMATOR = MLR;
8  MODEL: u_sch BY y1 y2; u_job BY y3 y4; u_fut BY y5 y6;
9              u_dl  BY y7–y9; u_soc BY y10–y12;
10 ! Identification
11 [y1@0 y3@0 y5@0 y7@0 y10@0]
12 ! Free means and intercepts
13 [SUM u_sch u_job u_dl u_soc];
14 ! Covariances between component variables and composite score (residual)
15 u_sch u_job u_dl u_soc WITH SUM;
16 ! Fix the factor loading of the pseudo-indicator variable ‘u_fut’
17 SUM BY u_fut@0.66633;
18 u_fut@0; [u_fut@0];
19 ! Fix the path coefficients of the remaining component variables
20 u_fut ON u_sch@-1.606672 u_job@-1.208522
21               u_dl @-0.951904 u_soc@-1.237349;
22 ! PIM with manifest composite score SES !
23 SES BY isei_m@1;
24 isei_m ON isei_d@-1 edu_m@-1 edu_d@-1 voc_m@-1
25               voc_d@-1 book@-1 income@-1;
26 [isei_m@0]; isei_m@0; [SES];
27 ! Covariances between component variables and composite score (residual)
28 isei_d edu_m edu_d voc_m voc_d book income SES WITH
29 isei_d edu_m edu_d voc_m voc_d book income SES;
30 ! Latent regression model
31 SUM ON SES ca sex math;
32 ! Additional covariances
33 isei_d edu_m edu_d voc_m voc_d book income WITH ca sex math y1–y12;
34 isei_d edu_m edu_d voc_m voc_d book income WITH
35 u_sch@0 u_job@0 u_dl@0 u_soc@0;
36 u_sch u_job u_dl u_soc WITH ca sex math SES;
37 ca sex math WITH ca sex math;
38 ! Starting values for variances
39 u_sch*1; u_job* 1; u_dl*1; u_soc*1; SUM*15;
```

In Lines 2 to 5, the variables in the data set are named. The variables id, school, and class are ID variables for students, classes, and schools. The classes and schools are required to identify the cluster variables in Line 6 with the cluster and stratification statement. In order to correct for the hierarchical structure of the data, FIML with robust standard errors (ESTIMATOR = MLR) is used in combination with TYPE = COMPLEX (see Line 7). The variables y1 to y12 are the 12 manifest variables in the measurement model of a student's mathematics utility. The eight indicators of *SES* are: (a) The socio-economic index of occupational status of the student's mother (isei_m) and father (isei_d), the highest educational degree of the mother (edu_m) and father (edu_d), the professional qualification of the mother (voc_m) and father (voc_d), the number of books at home (book), and the family's income (income). The student's grade in mathematics (math), gender (sex), and cognitive ability (ca) are predictors in the regression model. All missing data in the data set were coded to –99 (see Line 6).

The MODEL command starts with the specification of the measurement model of the five latent mathematics utility dimensions η_{fu} (u_fut), η_{sch} (u_sch), η_{job} (u_job), η_{dl} (u_dl), and η_{soc} (u_soc), which are measured by the variables Y_1 to Y_{12} (see Lines 8–9). The model was identified by fixing the factor loading of the first indicator per latent dimension to one (by default) and the measurement intercept of the same indicators to zero (see Line 11). The means of the four utility dimensions, with the exception of the pseudo-indicator variable, are freely estimated (see Line 13). The variable η_{fu} (u_fut) was chosen as the pseudo-indicator of the latent weighted sum score (SUM). The variable "SUM" denotes the overall utility index variable defined as $Utility = SD(\eta_{sch})^{-1}\eta_{sch} + SD(\eta_{job})^{-1}\eta_{job} + SD(\eta_{soc})^{-1}\eta_{soc} + SD(\eta_{fu})^{-1}\eta_{fu} + SD(\eta_{dl})^{-1}\eta_{dl}$. Hence, the weights are equal to the inverse standard deviations of the five latent utility dimensions. This is equivalent to an unweighted sum score of standardized latent component variables that have the same variance $Var(\eta_q) = 1$. The standard deviations of the latent component variables were estimated in an equivalent model without the latent composite score. Alternatively, the unweighted sum score can be specified in a preliminary model. The estimated variances of the latent component variables can be obtained from the model-implied variance-covariance matrix of the latent variables, which can be found in the TECH 4 in M*plus* or can be requested by the inspect() function of the lavaan package in R with the argument what = "cov.lv". The estimated standard deviations obtained in the first step are $SD(\eta_{sch}) = 0.415$, $SD(\eta_{job}) = 0.551$, $SD(\eta_{soc}) = 0.539$, $SD(\eta_{fu}) = 0.666$, and $SD(\eta_{dl}) = 0.700$. The weights γ_q are simply the inverse values of the standard deviations: $SD(\eta_{sch})^{-1} = 2.411$, $SD(\eta_{job})^{-1} = 1.814$, $SD(\eta_{soc})^{-1} = 1.857$, $SD(\eta_{fu})^{-1} = 1.501$, and $SD(\eta_{dl})^{-1} = 1.429$. Finally, the factor loading and the path coefficients have to be computed according to Figure 1 and 2 or Table 1. The factor loading of the latent pseudo-indicator variable η_{fu} was fixed to $\gamma_r^{-1} = SD(\eta_{fu}) = 0.666$ (see Line 18), and the residual variance and intercept were fixed to zero (see Line 18). The path coefficients of the paths $\eta_q \rightarrow \eta_r$ from the remaining component variables to the pseudo-indicator variable are equal to the ratios $-\gamma_{q\neq r}\,\gamma_r^{-1}$ of the weights. Accordingly, the paths were fixed to the computed values: $\eta_{sch} \rightarrow \eta_{fu} = -1.607$, $\eta_{job} \rightarrow \eta_{fu} = -1.209$, $\eta_{sco} \rightarrow \eta_{fu} = -1.237$, $\eta_{dl} \rightarrow \eta_{fu} = -0.952$ (lines 20–21). A composite score has free covariances with its component variables. As the latent weighted sum score is an endogenous variable in this data example, the latent component variables, with the exception of the pseudo-indicator variable, are allowed to covary with the residual of the latent composite score (see Line 15).

As an index of students' socioeconomic status, a manifest model-based sum score (SES) is formed by creating a second PIM (see Lines 23–26) using the socio-economic index of occupational status of the student's mother (isei_m) as the pseudo-indicator variable with the factor loading set to one and the intercept and the residual variance fixed to zero (see Line 26). The covariances between the component variables and (a) the SES composite score (see Lines 28–29), (b) the manifest predictors of the latent sum score, and (c) the manifest indicators of the five math utility dimensions (see Line 33) are freely estimated. The latent regression of the latent composite score (SUM) on the SES composite score, the student's grade in mathematics (math), gender (sex), and cognitive ability (ca) is specified in Line 31. According to the rules for constructing the PIM, additional covariances are allowed in order to avoid unintended restrictions that may lead to biased parameter estimates. Specifically, free covariances between the predictor variables (math, sex, and ca) and the latent math utility dimensions with the exception of the pseudo-indicator variable are allowed (see Lines 33). However, covariances between the latent math utility dimensions and the manifest component variables of the SES composite score are fixed to zero (see Lines 34–35). Finally, starting values are provided in Line 39; these account for the strong differences in the variances of the latent component variables and the latent composite score. This is not mandatory, but in our experience, appropriate starting values may prevent potential convergence problems. In this version of the model (Model A), the manifest component variables of the SES are purely auxiliary variables. The model even allows for potential measurement invariance in the measurement model of student's math utility depending on the manifest SES markers. As described in the test, we opted for a more parsimonious version of the model (denoted by Model B) with the assumption of measurement invariance. The M*plus* Syntax of Model B is obtained by fixing the covariances between the manifest component variables of the SES composite score and the measurement residuals of the manifest math utility indicators to zero (delete y1–y12 in Line 33). Instead, the covariances between the manifest component variables of the SES composite score and the latent math utility dimensions as well as the residual of the latent composite score SUM are specified as freely estimable parameters (delete the Lines 34–35) and add "isei_d edu_m edu_d voc_m voc_d book income WITH SUM" to the MODEL command.

In the following, we describe how to specify Model A in R. In order to fit the model using the lavaan package in R, the model needs to be syntactically correctly described in a character string. The character string needs to be a named object in R, which needs to be passed as the model argument to the sem() or lavaan() function. We arbitrarily chose the name "pim.mod" for the string that contains the model specification of Model A. Note that comments start with a "#" at the beginning of a line:

```
pim.mod <- '
# measurement model of the latent component variables
  u_sch = ~ y1 + y2;
  u_job = ~ y3 + y4;
  u_fut = ~ y5 + y6;
  u_dl = ~  y7 + y8 + y9;
  u_soc = ~ y10 + y11 + y12
# fix the measurement intercepts of one indicator per latent component variable to zero
  y1 + y3 + y5 + y7 + y10 ~ 0*1
# free the means of the latent component variables with the exception of the pseudo-indicator
# variable
  u_sch +  u_job + u_dl + u_soc ~ 1
# PIM with u_fut as the pseudo-indicator variable
  SUM = ~ 0.66633*u_fut
  u_fut ~ (–1.606672)*u_sch + (–1.208522)*u_job + (–0.951904)*u_dl + (–1.237349)*u_soc
# fix the residual variance and the intercept of the pseudo-indicator variable to zero
  u_fut ~~ 0*u_fut
  u_fut ~ 0*1
# free the mean of the latent composite score
  SUM ~ 1
# PIM with isei_m = pseudo indicator variable
  SES = ~ 1*isei_m
  isei_m ~ (–1)*isei_d + (–1)*edu_m + (–1)*edu_d +
          (–1)*voc_m + (–1)*voc_d + (–1)*book + (–1)*income
# fix the residual variance and the intercept of the pseudo-indicator variable to zero
  isei_m ~~ 0*isei_m
  isei_m ~ 0*1
# free the mean of the ses composite score
  SES ~ 1
# latent regression of the latent math utility index (SUM) on predictors including the SES
# composite score
  SUM ~ SES + sex + math + ca
# specify additional covariances according to the described rules
isei_d + edu_m + edu_d + voc_m + voc_d + book ~~ income
isei_d + edu_m + edu_d + voc_m + voc_d ~~ book
isei_d + edu_m + edu_d + voc_m ~~ voc_d
isei_d + edu_m + edu_d ~~ voc_m
isei_d + edu_m ~~ edu_d
isei_d ~~ edu_m
isei_d + edu_m + edu_d + voc_m + voc_d + book + income ~~ SES + sex + math + ca +
                                                           y1 + y2 + y3 + y4 + y5 + y6 +
                                                          y7 + y8 + y9 + y10 + y11 + y12
sex + math + ca ~~ SES
```

```
44 sex + math ~~ ca
45 sex ~~ math
46 u_sch + u_job + u_dl + u_soc ~~ SUM + SES + sex + math + ca
47 # starting values
48 SUM ~~ start(15)*SUM'
49 # run the model with the sem() function from the lavaan package
50 pim.mod.out <- sem(pim.mod, estimator = "mlr", missing = "fiml", data = mathutility_dat)
51 summary(pim.mod.out, fit = T, standardized = T, rsquare = T)
```

Note that the model specification ends at line 48. Lines 50 – 51 show the R code that calls the sem() function from the lavaan package to initiate model estimation. The analysis of only complete cases is the default setting in lavaan. Hence, unless otherwise specified, missing data are treated with listwise deletion. In order to utilize FIML for parameter estimation, the "missing" argument in the sem function needs to be specified as missing = "fiml". We also specified the argument estimator = "mlr" in line 50, which means that robust ML estimation is used with mean and variance adjusted χ^2-statistics, and robust standard errors are used.

In order to specify Model B with assumption of measurement invariance for the five latent utility dimensions regarding the manifest SES variables, Lines 40–42 need to be replaced by the following code:

```
isei_d + edu_m + edu_d + voc_m + voc_d + book + income ~~ SES + sex + math + ca + SUM +
                                                          u_sch + u_job + u_dl + u_soc
```

As a result, the covariances between the manifest component variables of the SES composite score and the residual variances of the manifest indicators of the five math utility dimensions are fixed to zero by default. Instead, the covariances between the manifest component variables of the SES and (a) the residual of the latent composite score and (b) the latent components variables of the latent composite score are specified as free estimands, with the exception of the two pseudo-indicator variables.

Notes

[1] The specification of various SEM with one or more manifest and/or latent composite scores in R and M*plus* is described in the PDF file 'RoseEtal2019suppl.pdf'. How to check for correct model specification by means of model comparisons is also explained for each example. The PDF and the associated data set ('RoseEtal2019suppl.dat') can be downloaded form: https://osf.io/b6h7n/.

[2] The data set 'GaspardEtal2015mathUtility.dat' can be downloaded from: https://osf.io/b6h7n/.

[3] In order to preserve readability and clarity, covariances and intercepts are not shown in the path diagram in Figure 3.

[4] Generally the variance of the weighted sum score is $Var(\sum_{q=1}^{Q}\gamma_q Y_q) = \sum_{q=1}^{Q}\gamma_q^2 Var(Y_q) + 2\sum_{q=1}^{Q-1}\sum_{m=q+1}^{Q}\gamma_q\gamma_m Cov(Y_q, Y_m)$.

Additional File

The Additional File for this article can be found as follows:

- Supplemental material for this article, including (a) simulated and real data, (b) the R-code of the simulation study, and (c) a PDF document with executable example of a wide range of SEM with manifest and latent model-based composite scores, is available from the Open Science Framework (https://osf.io/b6h7n/).

Acknowledgements

We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions, which helped to greatly improve and clarify the manuscript. The collection of the real data set was funded in part by German Research Foundation Grant TR 553/7-1 awarded to Ulrich Trautwein, Oliver Lüdtke, and Benjamin Nagengast.

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Norman Rose: Theoretical and mathematical derivation of the model, implementation in existing software (R, M*plus*), planning and conducting the simulation study, conducting statistical analyses, preparing supplemental material, drafting of the manuscript, approval of the final version to be published, agreeing to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Wolfgang Wagner: Substantial contribution to the theoretical development of the proposed models, reviewing the manuscript critically for important intellectual content at all stages of the manuscript preparation and revision, approval of the final version to be published, agreeing to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Axel Mayer: Substantial contributions to theoretical derivations and conceptualization of model-based composite scores, contributions and advice in the simulation study, reviewing the manuscript critically for important intellectual content at all stages of the manuscript preparation and revision, approval of the final version to be published, agreeing to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Benjamin Nagengast: Substantial contributions to theoretical integration of manifest and latent model-based composite scores, contributions and advice in the statistical analyses and description of the applied data example, reviewing the manuscript critically for important intellectual content at all stages of the manuscript preparation and revision, approval of the final version to be published, agreeing to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In: Marcoulides, G. A., & Schumacker, R. E. (Eds.), *Advanced structural equation modeling: Issues and techniques*, 243–277. Mahwah, NJ: Erlbaum.

Asparouhov, T. (2018). SRMR in M*plus*. Retrieved from: http://www.statmodel.com.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. DOI: https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & **Bonett, D. G.** (1980). Significance tests and goodness of fit in the analysis of covariance-structures. *Psychological Bulletin, 88*(3), 588–606. DOI: https://doi.org/10.1037/0033-2909.88.3.588

Bentler, P. M., & **Satorra, A.** (2010). Testing model nesting and equivalence. *Psychological Methods, 15*(2), 111–123. DOI: https://doi.org/10.1037/a0019625

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303–316. DOI: https://doi.org/10.1177/0049124189017003004

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*(1), 605–634. DOI: https://doi.org/10.1146/annurev.psych.53.100901.135239

Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *Mis Quarterly, 35*(2), 359–372. DOI: https://doi.org/10.2307/23044047

Bollen, K. A., & **Bauldry, S.** (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*(3), 265–284. DOI: https://doi.org/10.1037/a0024448

Borsboom, D., Mellenbergh, G. J., & **Van Heerden, J.** (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203–219. DOI: https://doi.org/10.1037/0033-295X.110.2.203

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. *IERI monograph series: Issues and methodologies in large-scale assessments, 1*, 51–69.

Brosseau-Liard, P. E., & **Savalei, V.** (2014). Adjusting incremental fit indices for nonnormality. *Multivariate Behavioral Research, 49*(5), 460–470. DOI: https://doi.org/10.1080/00273171.2014.933697

Conway, A. R. A., & **Kovacs, K.** (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*, 419–426. DOI: https://doi.org/10.1002/wcs.1356

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & **Midgley, C.** (1983). Expectancies, values, and academic behaviors. In: Spence, J. T. (Ed.), *Achievement and achievement motives*, 74–146. San Francisco, CA: Freeman.

Eid, M., Geiser, C., Koch, T., & **Heene, M.** (2017). Anomalous results in g-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541–562. DOI: https://doi.org/10.1037/met0000083

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8*(1), 128–141. DOI: https://doi.org/10.1207/S15328007SEM0801_7

Feeny, D., Huguet, N., McFarland, B. H., & **Kaplan, M. S.** (2009). The construct validity of the health utilities index mark 3 in assessing mental health in population health surveys. *Quality of Life Research, 18*(4), 519–526. DOI: https://doi.org/10.1007/s11136-009-9457-3

Ganzeboom, H. B. G., Degraaf, P. M., Treiman, D. J., & **Deleeuw, J.** (1992). A standard international socioeconomic index of occupational-status. *Social Science Research, 21*(1), 1–56. DOI: https://doi.org/10.1016/0049-089X(92)90017-B

Gaspard, H., Dicke, A. L., Flunger, B., Schreier, B., Hafner, I., Trautwein, U., & **Nagengast, B.** (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology, 107*(3), 663–677. DOI: https://doi.org/10.1037/edu0000003

Gibson, G. (1981). Indices of severity for emergency medical evaluative studies: Reliability, validity, and data requirements. *International Journal of Health Services, 11*(4), 597–622. DOI: https://doi.org/10.2190/6YFT-J9RA-YKR1-D70X

Grace, J. B., & **Bollen, K. A.** (2008). Representing general theoretical concepts in structural equation models: The role of composite variables. *Environmental and Ecological Statistics, 15*(2), 191–213. DOI: https://doi.org/10.1007/s10651-007-0047-7

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(1), 80–100. DOI: https://doi.org/10.1207/S15328007SEM1001_4

Heller, K. A., & **Perleth, C.** (2000). *Kognitiver Fähigkeitstest für 4.–12. Klassen, Revision (KFT 4–12+ R)*. Göttingen: Hogrefe.

Hu, L., & **Bentler, P. M.** (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. DOI: https://doi.org/10.1080/10705519909540118

James, L. R., Mulaik, S. A., & **Brett, J. M.** (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills: Sage.

Jöreskog, K. G., & **Sörbom, D.** (1981). *Lisrel V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago, IL: National Educational Resources.

Kirshner, B., & **Guyatt, G.** (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38*(1), 27–36. DOI: https://doi.org/10.1016/0021-9681(85)90005-0

Klein, A. G., & **Moosbrugger, H.** (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457–474. DOI: https://doi.org/10.1007/BF02296338

Klein, A. G., & **Muthén, B. O.** (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research, 42*(4), 647–673. DOI: https://doi.org/10.1080/00273170701710205

Little, R. J. A., & **Rubin, D. B.** (2002). *Statistical analysis with missing data*. New York, NY: Wiley. DOI: https://doi.org/10.1002/9781119013563

Little, T. D., Cunningham, W. A., Shahar, G., & **Widaman, K. F.** (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173. DOI: https://doi.org/10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K., & **Schoemann, A. M.** (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18*(3), 285–300. DOI: https://doi.org/10.1037/a0033266

Markus, K. A., & **Borsboom, D.** (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Mayer, A. (2013). *Contributions to latent variables in structural equation models and to causal mediation models*. (Unpublished doctoral dissertation). Jena: Friedrich-Schiller-University.

Mayer, A., Steyer, R., & **Mueller, H.** (2012). A general approach to defining latent growth components. *Structural Equation Modeling, 19*(4), 513–533. DOI: https://doi.org/10.1080/10705511.2012.713242

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605. DOI: https://doi.org/10.1146/annurev.psych.60.110707.163612

McDonald, R. P., & **Marsh, H. W.** (1990). Choosing a multivariate model – noncentrality and goodness of fit. *Psychological Bulletin, 107*(2), 247–255. DOI: https://doi.org/10.1037/0033-2909.107.2.247

Muthén, B. O., & **Muthén, L. K.** (1998–2018). *Mplus user's guide (eights edition)* (7th ed.). Los Angeles, CA: Muthén and Muthén.

Muthén, B. O., & **Satorra, A.** (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316. DOI: https://doi.org/10.2307/271070

OECD. (2010). *PISA 2009 results: Overcoming social background – equity in learning opportunities and outcomes, II*. Paris: OECD.

Pohl, S., & **Steyer, R.** (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45*(1), 45–72. DOI: https://doi.org/10.1080/00273170903504729

Pohl, S., Steyer, R., & **Kraus, K.** (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(1), 41–63. DOI: https://doi.org/10.1111/j.1467-985X.2007.00517.x

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*(4), 375–385. DOI: https://doi.org/10.1177/014662169802200407

Raykov, T., Marcoulides, G. A., & **Li, T.** (2017a). On the fallibility of principal components in research. *Educational and Psychological Measurement, 77*(1), 165–178. DOI: https://doi.org/10.1177/0013164416629714

Raykov, T., Marcoulides, G. A., & **Li, T.** (2017b). On the unlikely case of an error-free principal component from a set of fallible measures. *Educational and Psychological Measurement*. DOI: https://doi.org/10.1177/0013164416686147

Raykov, T., & **Penev, S.** (1999). On structural equation model equivalence. *Multivariate Behavioral Research, 34*(2), 199–244. DOI: https://doi.org/10.1207/S15327906Mb340204

Redelmeier, D. A., & **Lustig, A. J.** (2001). Prognostic indices in clinical practice. *JAMA, 285*(23), 3024–3025. DOI: https://doi.org/10.1001/jama.285.23.3024

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (beta). *Journal of Statistical Software, 48*(2), 1–36. DOI: https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. DOI: https://doi.org/10.1093/biomet/63.3.581

Satorra, A., & **Bentler, P. M.** (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243–248. DOI: https://doi.org/10.1007/s11336-009-9135-y

Savalei, V., & **Rhemtulla, M.** (2017). Normal theory gls estimator for missing data: An application to item-level missing data and a comparison to two-stage ml. *Frontiers in Psychology, 8*. DOI: https://doi.org/10.3389/fpsyg.2017.00767

Schafer, J. L., & **Graham, J. W.** (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. DOI: https://doi.org/10.1037/1082-989X.7.2.147

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology, 15*(1), 72–101. DOI: https://doi.org/10.2307/1412159

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika, 72*(3), 583–592. DOI: https://doi.org/10.1093/biomet/72.3.583

Steyer, R., Eid, M., & **Schwenkmezger, P.** (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2*(1), 21–33.

Steyer, R., Krambeer, S., & **Hannover, W.** (2004). Modeling latent trait-change. In: Van Montfort, K., & Oud, J. (Eds.), *Recent developments on structural equation models: Theory and applications*, *19*, 337–357. Dordrecht: Kluwer Academic Publishers. DOI: https://doi.org/10.1007/978-1-4020-1958-6_16

Tucker, L. R., & **Lewis, C.** (1973). Reliability coefficient for maximum likelihood factor-analysis. *Psychometrika, 38*(1), 1–10. DOI: https://doi.org/10.1007/BF02291170

Yourman, L. C., Lee, S. J., Schonberg, M. A., Widera, E. W., & **Smith, A. K.** (2012). Prognostic indices for older adults: A systematic review. *JAMA, 307*(2), 182–192. DOI: https://doi.org/10.1001/jama.2011.1966

Yuan, K. H., & **Bentler, P. M.** (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical & Statistical Psychology, 51*(1), 63–88. DOI: https://doi.org/10.1111/j.2044-8317.1998.tb00667.x

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.143.pr

How to cite this article: Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-Based Manifest and Latent Composite Scores in Structural Equation Models. *Collabra: Psychology,* 5(1): 9. DOI: https://doi.org/10.1525/collabra.143

Submitted: 08 February 2018 **Accepted:** 11 January 2019 **Published:** 22 February 2019

Senior Editor: Victoria Savalei

Editor: Victoria Savalei

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS