Estimating the validity of administrative and survey variables through structural equation modeling

A simulation study on robustness

Sander Scholtus and Bart F.M. Bakker

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201302)

Statistics Netherlands

The Hague/Heerlen, 2013

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2012–2013	2012 to 2013 inclusive
2012/2013	average for 2012 up to and including 2013
2012/'13	crop year, financial year, school year etc. beginning in 2012 and ending in 2013
2010/'11–2012/'13	crop year, financial year, etc. 2010/'11 to 2012/'13 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

60083201302 X-10

Estimating the validity of administrative and survey variables through structural equation modeling: a simulation study on robustness

Sander Scholtus and Bart F.M. Bakker

Summary: Over the past years, the use of administrative data in both official statistics and academic research has grown. This development has made the problem of assessing the quality of administrative sources for statistical use increasingly important. Two of the main aspects of this are validity and reliability of measurement. Although this problem is often mentioned in qualitative terms, so far, not much research has been done on methods that assess the validity or reliability of administrative variables in a quantitative way. The objective of this paper is to describe a quantitative method for estimating validity and to present results obtained with this method in a simulation study.

Following Bakker (2012), the classical test theory can be applied to estimate the validity of administrative variables. This approach requires linked data from at least two sources (surveys and/or registers) that are supposed to measure the same concepts. The validity of the observed variables for measuring these concepts is estimated through a linear structural equation model with a measurement component.

We performed a simulation study to test the robustness of this method to different amounts of measurement error, to misspecification of the measurement model, and to small sample size. The results of these simulations indicate that the method provides reasonable estimates of validity in many situations. Although partial misspecification of the measurement model typically yields biased validity estimates for the variables that are directly involved, it appears that these effects are not propagated to the rest of the model. In addition, a moderate sample size appears to be sufficient to obtain stable parameter estimates unless the model contains many variables with low validity.

Keywords: administrative data; validity; structural equation model; measurement model; simulation study

1 Introduction

Over the past years the use of administrative data has grown, both in official statistics and academic research. As a result, it has become increasingly important to assess and compare the quality of administrative sources for statistical use (Bakker and Daas, 2012). Two important aspects of this are validity (absence of bias) and reliability of measurement. In general, the concept that is measured for administrative purposes may differ to some extent from the concept that is needed for statistical purposes. If these two concepts are substantially different, one should expect the administrative variable to have a low validity for statistical use. Although this problem has been recognised in recent years, so far, not much research has been done on methodology for assessing the validity or reliability of administrative variables in a *quantitative* way.

An approach that is sometimes used in practice proceeds by linking and comparing administrative data to data from a reference source (e.g. a survey) that are – implicitly or explicitly – assumed to be error-free. Obviously, this assumption greatly simplifies matters: it implies that a value in the administrative data set is measured correctly if, and only if, it matches the corresponding value in the reference data set. It seems doubtful, however, that any data set obtained under real-world conditions can be assumed to be completely free of measurement error. Therefore, this approach may be used as a first approximation at best, provided that the validity of the reference data is high.[1] Without the assumption of error-free reference data, the above set-up can be used to assess differences in measurement between sources rather than errors of measurement in one source, which may also be useful (Groen, 2012). In the remainder of this paper, however, we are interested in obtaining direct estimates of the validity of administrative variables when error-free reference data are not available.

More precisely, we consider the situation that a second data source *is* available but that the validity of this data is not high enough to warrant their use as error-free reference data (not even as a first approximation). Bakker (2012) suggested that the classical test theory can be applied in this context to estimate the validity of administrative variables as well as survey variables. This requires linked data from at least two sources (surveys and/or registers) that are supposed to measure the same concepts. The validity of the observed variables for measuring these concepts is estimated through a linear structural equation model with a measurement component (Bollen, 1989; Saris and Andrews, 1991). In this way, the amount of measurement error is estimated for each variable in each linked data source. A more detailed description is given in Section 2 below.

Structural equation modeling is a well-established technique for assessing the validity and reliability of survey data, in particular in the context of questionnaire design (Andrews, 1984; Saris and Gallhofer, 2007). The present application to administrative data is somewhat different, however. For instance, applications in the context of questionnaire design typically include at least three versions of each survey item (being different with respect to the wording of the question, the response scale, etc.), which means that each concept in the model is measured by three or more indicators. In the application considered here, by contrast, each concept is measured at most twice:

[1] The same remark applies to the more common reverse approach of using an administrative source for reference data to assess the quality of survey data.

once in a register and once in a survey. It is therefore important to gain insight into the usefulness and limitations of this method when applied to administrative data.

Two issues that require clarification are the following. Firstly, one would intuitively expect the method to perform better for larger samples and for observed variables with higher validities. Conversely, the method may give inappropriate results if the sample is too small and/or if the validities of some variables in the model are too low. We would like to make this statement more precise: what is a 'too small' sample size and what is a 'too low' validity? Secondly, we would like to know whether the method is robust to certain forms of misspecification of the measurement model. In particular, we are interested in the case that certain measurement errors are correlated and the model does not take this into account. In principle, having a misspecified model may completely invalidate the outcome of the method. However, the method may be robust against partial misspecification of the model and yield reasonable estimates for the validity of variables that are not directly involved in the misspecified part. In other words, it would be convenient if a partial misspecification only has a 'local' effect on the outcome of the method.

In order to investigate the above properties, we performed a simulation study. In this study, the method was applied to various artificial data sets for which the data generating mechanism was known. In this way, we could compare the estimated validities of the observed variables with their theoretical values. The results of this simulation study are described in Sections 3 and 4 below. We also give some theoretical arguments to clarify the observed results. A discussion and conclusion follow in Section 5. Finally, some more technical material is provided in two appendices.

2 Validity and Measurement Models

2.1 Estimating Validity through Structural Equation Modeling

The classical test theory [e.g. Novick (1966)] distinguishes two aspects of measurement quality: *reliability* and *validity*. According to McCall (2001, p. 308), reliability 'refers to whether the measurement procedures assign the same value to a characteristic each time it is measured under essentially the same circumstances'. Reliability is associated with random measurement error. A perfectly reliable measure yields exactly the same value every time it is used (provided the circumstances are 'essentially the same'). Note that a perfectly reliable measure may still be biased. Validity, on the other hand, 'refers to the extent to which the measurement procedures assign values that accurately reflect the conceptual variable being measured' (McCall, 2001, p. 309). Validity is associated with systematic measurement error, i.e. bias. In the present paper we will focus on estimating validity, although the method discussed here can be extended to also estimate reliability (see Section 2.3).

Consider the following simple measurement model for a variable y that is supposed to measure a conceptual variable η:

$$y = \lambda\eta + \varepsilon, \qquad (1)$$

where it is assumed that ε, the measurement error, has mean 0 and is uncorrelated with the concept η. In addition, we assume that both y and η are standardised to have zero mean and unit variance. We take the factor loading λ as a measure of the validity of y for measuring the concept η.[2] In regression terms, λ^2 represents the fraction of the total variance in y that is explained by η. Clearly, a higher fraction of explained variance corresponds to a more valid measurement process. The values $\lambda = 1$ and $\lambda = 0$ occur in the special cases of perfect measurement and absence of correlation between y and η, respectively. Alternatively, the quantity

$$\mathrm{sd}(\varepsilon) = \sqrt{1 - \lambda^2}$$

can be taken as a measure of invalidity.

Obviously, a single instance of model (1) cannot be estimated, as we do not have access to the scores of the latent variable η.[3] Estimation may become possible if we simultaneously consider several variables that measure several concepts, as well as the causal relations between these concepts. It is convenient to formulate this as a linear structural equation model (SEM) with a measurement component (Saris and Andrews, 1991).

To introduce the notation used in the rest of this paper, we briefly discuss the SEM in general before continuing with the application to estimating validity. More technical details are given in Appendix A. The general set-up will be illustrated with an example in Section 2.2.

The first part of an SEM concerns the causal relations between the latent variables. It is customary to distinguish endogenous and exogenous latent variables, represented by $\eta_1, \ldots, \eta_m$ and $\xi_1, \ldots, \xi_n$, respectively. The relations between these variables are expressed in a set of m linear equations (one for each endogenous variable):

$$\eta_i = \sum_{i'=1}^{m} \beta_{ii'} \eta_{i'} + \sum_{j=1}^{n} \gamma_{ij} \xi_j + \zeta_i, \quad i = 1, \ldots, m, \tag{2}$$

with ζ_i a disturbance term that is assumed to be uncorrelated with $\xi_1, \ldots, \xi_n$. The coefficients $\beta_{ii'}$ and γ_{ij} represent the direct effects from $\eta_{i'}$ to η_i and from ξ_j to η_i, respectively. By definition, $\beta_{ii} = 0$ for all i. Further model parameters define the covariance structures of the disturbances and the exogenous factors: $\psi_{ii'} = \mathrm{cov}(\zeta_i, \zeta_{i'})$ and $\varphi_{jj'} = \mathrm{cov}(\xi_j, \xi_{j'})$. Usually when a model is formulated, some of the β, γ, ψ, and φ parameters are fixed to zero.

The second part of an SEM describes the relations between the latent variables and the observed variables that measure them. The observed variables corresponding to the endogenous and exogenous concepts are denoted by $y_1, \ldots, y_p$ and $x_1, \ldots, x_q$, respectively. For our present purpose, we can assume that each observed variable measures

[2] To simplify matters, we assume throughout this paper that all factor loadings are positive. In general, validity would be measured by the absolute value of λ.

[3] For some theoretical concepts – 'intelligence' being the canonical example – direct measurement is intrinsically impossible. In the context of official statistics, we are often dealing with relatively concrete concepts that could, in theory, be measured exactly, given sufficient resources. Examples considered below include 'age', 'gender', and 'hourly wages'. In practice, however, measurement is never without error and we can only obtain imperfect measures of these concepts.

a single latent variable. With this simplifying assumption, the measurement model is as follows:

$$y_k = \lambda_{yk}\eta_{i(k)} + \varepsilon_k, \quad k = 1,\dots,p, \tag{3}$$

$$x_l = \lambda_{xl}\xi_{j(l)} + \delta_l, \quad l = 1,\dots,q, \tag{4}$$

where $i(k)$ represents the index of the endogenous latent variable that is measured by y_k, and similarly for $j(l)$. The covariances of the error terms are denoted by $\theta_{\varepsilon kk'} = \mathrm{cov}(\varepsilon_k, \varepsilon_{k'})$ and $\theta_{\delta ll'} = \mathrm{cov}(\delta_l, \delta_{l'})$.

The above model formulation requires that all variables (latent and observed) are centered to have zero mean and that the measurement errors are such that all pairs (ζ_i, ε_k), (ζ_i, δ_l), (ξ_j, ε_k), (ξ_j, δ_l), and $(\varepsilon_k, \delta_l)$ are uncorrelated (Bollen, 1989). In addition, we will assume here that all η_i, ξ_j, y_k, and x_l are standardised to have unit variance and that the covariance matrices of the disturbances and measurement errors are diagonal.

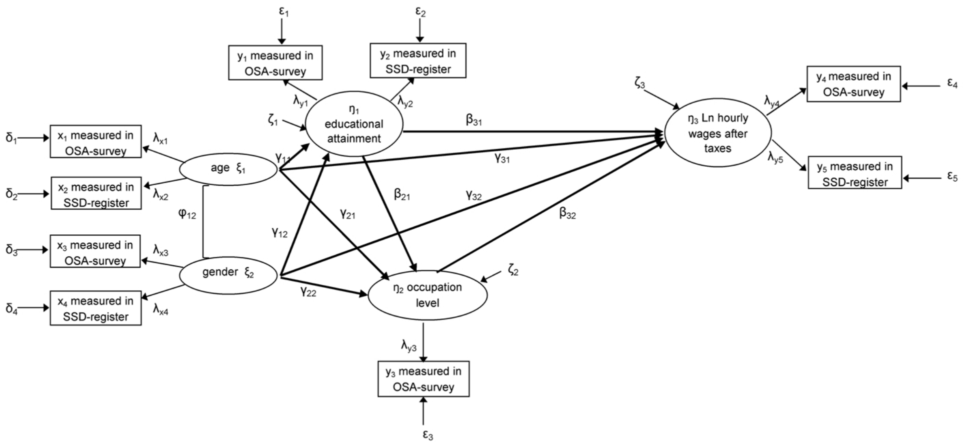
The SEM given by (2), (3), and (4) can be estimated from the observed joint covariance matrix[4] of $y_1,\dots,y_p$ and $x_1,\dots,x_q$ (Bollen, 1989; Jöreskog and Sörbom, 1996); see also Appendix A. Various software packages are available for estimating and analysing SEMs; see Narayanan (2012) for a recent overview.

For the analysis of an SEM, it is important that all parameters of the model are identified. In general, model identification requires that we have at least two observed variables for each latent variable. In the absence of sufficient observed variables, identification can be enforced by fixing certain model parameters to a constant; obviously, this increases the risk of misspecification. Bollen (1989) discusses various criteria for establishing model identification as well as measures of model fit. A common test of overall model fit uses a statistic X^2 that is (under an assumption of multivariate normal data) asymptotically chi-square distributed for correct models; see Appendix A.

We now return to the application at hand. As mentioned in the introduction, we assume that there are two linked data sets available that are supposed to measure the same concepts. The relations between the observed variables and the concepts that they are supposed to measure, as well as the relations between these concepts can be modeled as an SEM by specifying appropriate instances of the equations (2)–(4). We also assume that previous substantive research provides information on plausible values for the structural parameters in the model, i.e. the direct effects between the latent concepts.

Having estimated the model parameters and assuming that the model fits the data sufficiently to not be rejected, we may now assess the validity of the observed variables for measuring the associated concepts as follows (Bakker, 2012). First, the estimated direct effects $\beta_{ii'}$ and γ_{ij} between the latent variables are inspected and compared with their expected values from previous research. If these estimated effects are implausible, we have to conclude that at least some of the latent variables in the model do not correspond with the theoretical concepts that they are supposed to represent. This might be due to undetected misspecification of the model or to very large measurement errors in at least one of the observed variables. On the other hand, if the estimated effects are plausible, then we are confident that the latent variables in the model represent

[4] Given that the observed variables are standardised, this is actually a correlation matrix.



ε₁
ε₂
y₁ measured in OSA-survey
y₂ measured in SSD-register
λy1
λy2
ζ₁
η₁ educational attainment
β31
ζ₃
η₃ Ln hourly wages after taxes
λy4
λy5
y₄ measured in OSA-survey
y₅ measured in SSD-register
ε₄
ε₅
δ₁
δ₂
δ₃
δ₄
x₁ measured in OSA-survey
x₂ measured in SSD-register
x₃ measured in OSA-survey
x₄ measured in SSD-register
λx1
λx2
λx3
λx4
age ξ₁
gender ξ₂
φ12
γ11
γ12
γ21
γ22
γ31
γ32
β21
β32
η₂ occupation level
ζ₂
λy3
y₃ measured in OSA-survey
ε₃

Figure 1: Structural equation model for estimating the validity of register and survey variables; adapted from Bakker (2012).

the posited theoretical concepts. In this case, in analogy with the basic model (1), we take the factor loading λ_{yk} (resp. λ_{xl}) as a measure of the validity of y_k (resp. x_l) for measuring the concept $\eta_{i(k)}$ (resp. $\xi_{j(l)}$).

2.2 Example

Bakker (2012) described an application of the above method, which we summarise here both as an illustration and as a starting point for the simulation study described below. The application concerned four concepts ('age', 'gender', 'educational attainment', and 'ln hourly wages after taxes') that were measured in a register (the Social Statistical Database at Statistics Netherlands; SSD) and a survey (the OSA supply panel 2004). In addition, the model contained one concept ('occupational level') that was only measured in the survey. Linked data were available for a sample of $N = 574$ Dutch persons between 15 and 50 years of age.

Figure 1 displays the SEM used in the form of a path diagram. The structural part of this model – i.e. the part corresponding to (2) – is known as an earnings function model. We refer to Bakker (2012) and the references therein for an overview of previous research on earnings function models, as well as a discussion of the data sources and the procedure used to link the survey data to the register data.

Bakker (2012) estimated the above model using the LISREL software (Jöreskog and Sörbom, 1996). The first panel in Table 1 shows the estimated direct effects in the earnings function model. Based on previous research, these effects were considered sufficiently plausible. In a chi-square test of overall fit, the value of the test statistic was 48 at 18 degrees of freedom, which was also deemed acceptable for the present purpose. The second panel in Table 1 shows the estimated variances of the disturbance terms in the model. In addition, the correlation between the exogenous concepts 'age' and 'gender' was estimated at $\varphi_{12} = -0.07$.

The third panel in Table 1 shows the estimated factor loadings λ_{xl} and λ_{yk} from the model. Since 'occupational level' was only measured in the survey, the factor loading for this variable (y_3) was fixed to 1 in order to have an identified model; nothing could be concluded from this about the validity of y_3. For the remaining variables, the factor loadings were interpreted as measures of validity. Thus, for 'educational attainment' the register provided a more valid measurement than the survey ($\lambda_{y2} = 0.94$ vs. $\lambda_{y1} = 0.82$), while for 'hourly wages' it was the other way around ($\lambda_{y5} = 0.87$ vs. $\lambda_{y4} = 0.95$). In addition, 'age' and 'gender' were measured more or less perfectly in both data sources.

Table 1: Estimated parameters in the model of Figure 1.

from \ to	edu.att. (η_1)	occ.lev. (η_2)	h.wages (η_3)
age (ξ_1)	$\gamma_{11} = -0.22$	$\gamma_{21} = 0.13$	$\gamma_{31} = 0.32$
gender (ξ_2)	$\gamma_{12} = 0.00$	$\gamma_{22} = -0.09$	$\gamma_{32} = -0.18$
educational attainment (η_1)	–	$\beta_{21} = 0.59$	$\beta_{31} = 0.36$
occupational level (η_2)	–	–	$\beta_{32} = 0.32$

	edu.att. (η_1)	occ.lev. (η_2)	h.wages (η_3)
unexplained variance	$\psi_{11} = 0.95$	$\psi_{22} = 0.66$	$\psi_{33} = 0.54$

concept	OSA survey	SSD register
age (ξ_1)	$\lambda_{x1} = 1.00$	$\lambda_{x2} = 1.00$
gender (ξ_2)	$\lambda_{x3} = 1.00$	$\lambda_{x4} = 1.00$
educational attainment (η_1)	$\lambda_{y1} = 0.82$	$\lambda_{y2} = 0.94$
occupational level (η_2)	($\lambda_{y3} = 1.00$)	–
ln hourly wages (η_3)	$\lambda_{y4} = 0.95$	$\lambda_{y5} = 0.87$

2.3 A More Complex Measurement Model

As far as the measurement component is concerned, the SEM of Figure 1 consists of several instances of the basic model (1). This measurement model is very simple, in particular because it does not distinguish between different sources of errors.[5] More complex measurement models have been proposed in the literature on nonsampling errors. While we will not pursue the use of these models in this paper, it is interesting to consider briefly their relation to the basic model (1).

Saris and Andrews (1991) and Scherpenzeel and Saris (1997) proposed the following general measurement model (using the notation of the second reference):

$$y_i = h_{il} T_l + e_i, \tag{5}$$

$$T_l = b_{lj} F_j + g_{lk} M_k + u_l. \tag{6}$$

In this model, the observed variable y_i is decomposed into a stable part T_l and a random part e_i, where 'stable' refers to a component that would be observed again if the measurement procedure were repeated under identical circumstances. The stable part is decomposed further into three components: F_j, being the concept that one wishes to measure [η in our model (1)]; M_k, being a stable component specific to the method

[5] It can be shown that, as a result of this simplification, λ in (1) expresses both the validity *and* the reliability of y as a measure for η; see Biemer and Stokes (1991).

of observation; and u_l, being a stable component that depends on the interaction of the method and the concept being observed. An assumption of this model is that all components are mutually uncorrelated.

Following Heise and Bohrnstedt (1970), Saris and Andrews (1991) proposed to use the coefficient b_{lj} to quantify the validity of y_i as a measure for the concept F_j, and to use h_{il} to quantify the reliability. On the other hand, our λ would correspond with $b_{lj}h_{il}$ under the model (5)–(6). This quantity is also a commonly used measure of validity (Andrews, 1984). To distinguish the two types, Saris and Andrews (1991) suggested the terms *true score validity* for b_{lj} and *indicator validity* for $b_{lj}h_{il}$. A potential drawback of using indicator validity is that it is not a 'pure' measure of validity, since it is also influenced by random measurement errors.

Technically, the model (5)–(6) is an SEM (Saris and Andrews, 1991). Therefore, on paper, estimating validity and reliability under this model is a straightforward extension of the method discussed here. However, identification of the extended model places additional requirements on the data: more observed variables from different sources are needed for each concept. See e.g. Scherpenzeel and Saris (1997) for a discussion of methods to estimate a model of the form (5)–(6) in the context of survey data. It remains an open question how these methods can be adapted to the context of administrative data, where a researcher has much less control over the data collection process.

It should be noted that under the model (5)–(6), the simplified model (1) is still appropriate for the purpose of estimating indicator validity (as opposed to true score validity or reliability). Namely, this model subsumes various components under one new error term that are irrelevant to indicator validity. Moreover, by examining the plausibility of the direct effects between the latent variables in the SEM as described in Section 2.1, we are actually testing whether it can be assumed that $T_l = F_j$. In that case, the true score validity equals 1 by definition.

3 Simulation Study, Part 1: Multinormal Data

3.1 Introduction

To investigate the robustness of the method of Section 2 for estimating validity in various situations, we performed a simulation study. For this study, we generated a large number of data sets with the same variable structure as in the example of Section 2.2, in which synthetic measurement errors were introduced according to various known mechanisms. Subsequently, we estimated the validity of the variables in these synthetic data sets through the SEM of Figure 1. As the distribution of the measurement errors was known in this study, the appropriateness of the estimated validities could be assessed.

For the first part of our simulation study, we worked with synthetic data drawn from a multivariate normal distribution. This represents an idealised situation, since exactly multinormal data are virtually never encountered in practice. For the second part of the study, to be described in Section 4, we worked with more realistic data sets. Throughout, we used the statistical software R to generate data and analyse the results and we used LISREL (version 8.8) to fit the SEMs.

In general, the estimated parameters of an SEM imply an estimated covariance structure for the latent variables in the model. That is, along with the parameter estimates we obtain an estimated covariance matrix for the vectors $\vec{\eta} = (\eta_1, \ldots, \eta_m)'$ and $\vec{\xi} = (\xi_1, \ldots, \xi_n)'$, given by

$$\text{cov}(\vec{\eta}, \vec{\xi}) = \begin{pmatrix} E(\vec{\eta}\vec{\eta}') & E(\vec{\eta}\vec{\xi}') \\ E(\vec{\xi}\vec{\eta}') & E(\vec{\xi}\vec{\xi}') \end{pmatrix} = \begin{pmatrix} \text{C} & \text{G} \\ \text{G}' & \Phi \end{pmatrix},$$

where the entries in the matrices C, G, and Φ are known functions of the parameters in the model; see Appendix A. In all simulations with multivariate normal data considered below, we drew scores for the latent variables $(\eta_1, \eta_2, \eta_3, \xi_1, \xi_2)$ from a multinormal distribution with mean zero and the covariance matrix as implied by the parameter estimates from Section 2.2. Rounded to three decimals, this implied covariance matrix is

$$\begin{array}{c} \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \xi_1 \\ \xi_2 \end{array} \begin{array}{c} \begin{array}{ccccc} \eta_1 & \eta_2 & \eta_3 & \xi_1 & \xi_2 \end{array} \\ \begin{pmatrix} 1.000 & & & & \\ 0.563 & 1.000 & & & \\ 0.462 & 0.535 & 1.000 & & \\ -0.219 & 0.004 & 0.259 & 1.000 & \\ 0.020 & -0.088 & -0.225 & -0.073 & 1.000 \end{pmatrix} \end{array}. \tag{7}$$

In particular, this means that in all simulations the true values of the structural parameters were the same as in Section 2.2.

3.2 Correctly Specified Measurement Models

We first considered various cases for which the measurement model posited in Figure 1 is an adequate description of the mechanism that generated the measurement errors. More precisely, we drew random vectors

$$(\eta_1, \eta_2, \eta_3, \xi_1, \xi_2, \varepsilon_1, \ldots, \varepsilon_5, \delta_1, \ldots, \delta_4)'$$

from a joint multinormal distribution having the zero vector as its mean and the following covariance matrix:

$$\begin{array}{c} \\ \vec{\eta} \\ \vec{\xi} \\ \vec{\varepsilon} \\ \vec{\delta} \end{array} \begin{array}{c} \begin{array}{cccc} \vec{\eta} & \vec{\xi} & \vec{\varepsilon} & \vec{\delta} \end{array} \\ \begin{pmatrix} \text{C}^* & \text{G}^* & \text{O} & \text{O} \\ (\text{G}^*)' & \Phi^* & \text{O} & \text{O} \\ \text{O} & \text{O} & \Theta^*_\varepsilon & \text{O} \\ \text{O} & \text{O} & \text{O} & \Theta^*_\delta \end{pmatrix} \end{array} \tag{8}$$

where the top left part is given by (7), O denotes a block of zeros, and Θ^*_ε and Θ^*_δ are diagonal matrices having pre-specified values $\theta^*_{\varepsilon 11}, \ldots, \theta^*_{\varepsilon 55}$ and $\theta^*_{\delta 11}, \ldots, \theta^*_{\delta 44}$ (all between 0 and 1) on their main diagonal.[6] From each realisation of this random vector

[6] We use an asterisk to indicate that a parameter belongs to the data generating model. The parameters of the SEM used for estimating validity do not have asterisks. This distinction will become important when we consider model misspecification.

we computed the observed variables in our data set as follows:

$$\left.\begin{array}{rcl}
y_1 &=& \lambda^*_{y1}\eta_1+\varepsilon_1,\\
y_2 &=& \lambda^*_{y2}\eta_1+\varepsilon_2,\\
y_3 &=& \lambda^*_{y3}\eta_2+\varepsilon_3,\\
y_4 &=& \lambda^*_{y4}\eta_3+\varepsilon_4,\\
y_5 &=& \lambda^*_{y5}\eta_3+\varepsilon_5,\\
x_1 &=& \lambda^*_{x1}\xi_1+\delta_1,\\
x_2 &=& \lambda^*_{x2}\xi_1+\delta_2,\\
x_3 &=& \lambda^*_{x3}\xi_2+\delta_3,\\
x_4 &=& \lambda^*_{x4}\xi_2+\delta_4,
\end{array}\right\} \tag{9}$$

with $\lambda^*_{yk}=\sqrt{1-\theta^*_{\varepsilon kk}}$ and $\lambda^*_{xl}=\sqrt{1-\theta^*_{\delta ll}}$. By construction, $\mathrm{var}(y_k)=\mathrm{var}(x_l)=1$ for all k and l. Moreover, by comparison with the basic model (1) it is seen that the correct validities of y_k and x_l are given by λ^*_{yk} and λ^*_{xl}, respectively.

Note that different versions of the above model can be specified by varying the choice of $\theta^*_{\varepsilon 11},\dots,\theta^*_{\varepsilon 55}$ and $\theta^*_{\delta 11},\dots,\theta^*_{\delta 44}$. Each of these parameters corresponds to the fraction of the total variance in an observed variable that is due to measurement error. For *any* choice of these parameters, the data generating model (9) is in complete agreement with the SEM of Section 2.2.[7] Thus, one would intuitively expect the factor loadings λ_{yk} and λ_{xl} to be identical to the above validities λ^*_{yk} and λ^*_{xl}. In fact, it is not difficult to show that this identity would hold exactly if the population covariance matrix were analysed; see Appendix A. In practice, however, one has to work with an estimate of the population covariance matrix based on a finite sample and this introduces estimation uncertainty. The resulting estimated factor loadings are still consistent estimators for the validities λ^*_{yk} and λ^*_{xl} under rather general regularity conditions (Bollen, 1989, p. 416ff.). Thus, for large samples, we expect them to agree closely with the true validities.

We also conjecture that, for smaller samples in particular, the behaviour of the estimated validities is likely to become erratic as $\theta^*_{\varepsilon kk}$ and $\theta^*_{\delta ll}$ get close to 1. In this situation, the observed variables are mainly determined by random measurement error. Therefore, a typical realisation of the sample covariance matrix may be relatively far from the population covariance matrix, making it difficult to estimate the SEM.

In our first set of simulations, we generated and analysed data sets according to the following versions of the above model.

1. Increasing amounts of measurement error in all observed variables:
   a. all error variances equal to 0.1;
   b. all error variances equal to 0.2;
   c. all error variances equal to 0.5;
   d. all error variances equal to 0.9.
2. Increasing amounts of measurement error in one observed variable for one exogenous factor:
   a. error variance of x_1 equal to 0.2; all other error variances equal to 0.1;

[7] There is one exception: the model of Section 2.2 cannot handle errors in y_3 because the factor loading of this variable is fixed to 1. This point will be taken up later.

    b. error variance of x_1 equal to 0.5; all other error variances equal to 0.1;
    c. error variance of x_1 equal to 0.9; all other error variances equal to 0.1.

3. Increasing amounts of measurement error in both observed variables for one exogenous factor:
    a. error variances of x_1 and x_2 equal to 0.2; all other error variances equal to 0.1;
    b. error variances of x_1 and x_2 equal to 0.5; all other error variances equal to 0.1;
    c. error variances of x_1 and x_2 equal to 0.9; all other error variances equal to 0.1.
4. Increasing amounts of measurement error in one observed variable for one endogenous factor:
    a. error variance of y_5 equal to 0.2; all other error variances equal to 0.1;
    b. error variance of y_5 equal to 0.5; all other error variances equal to 0.1;
    c. error variance of y_5 equal to 0.9; all other error variances equal to 0.1.
5. Increasing amounts of measurement error in both observed variables for one endogenous factor:
    a. error variances of y_4 and y_5 equal to 0.2; all other error variances equal to 0.1;
    b. error variances of y_4 and y_5 equal to 0.5; all other error variances equal to 0.1;
    c. error variances of y_4 and y_5 equal to 0.9; all other error variances equal to 0.1.

The theoretical validities corresponding to the above error variances are given in Table 2.

Table 2: Theoretical validities for different error variances.

error variance	0.1	0.2	0.5	0.9
validity	0.95	0.89	0.71	0.32

It should be noted that we fixed $\theta^*_{\varepsilon 33}$ to 0 in this set of simulations. Recall that the factor loading of y_3 is fixed to 1 in the SEM of Section 2.2 to ensure identification. As a result, any measurement errors in y_3 would be seen as part of the disturbance term ζ_2 and hence could have an adverse effect on the outcome of the method. This effect was investigated in a later simulation (see Section 3.3), but it was excluded here.

We used a moderate sample size of $N = 600$, which is comparable to the real-world example of Section 2.2. As a benchmark, we repeated the same analyses with a very large sample size ($N = 60000$). For each combination of a model and a sample size we performed $R = 100$ simulations.

Table 3 displays the results for $N = 600$. The second column lists the number of simulations R_c for which LISREL converged[8] to an optimal solution. Severe problems with convergence occurred only for the model with all error variances equal to 0.9 (model 1d). However, even if an optimal solution is found, it may contain negative estimates for certain variance parameters in the model (or, equivalently, factor loadings with $\lambda > 1$). This invalidates the solution to some extent and complicates the interpretation of the estimated parameters. The third column in Table 3 lists the number of simulations R_v for which all estimated variance parameters were higher than -0.05 (i.e. either positive or 'almost' positive). As can be seen, negative estimated variances occurred in all models to some extent, but they were encountered particularly often in models with one or more error variances equal to 0.9.

[8] The maximum number of iterations in LISREL's optimisation routine was set to 1000 throughout.

Table 3: Simulation results for correctly specified models with $N = 600$; apart from R_c and R_v, all reported values are averages over R_c simulations, with standard deviations in brackets.

model	R_c	R_v	X^2	λ_{y1}	λ_{y2}	λ_{y4}	λ_{y5}	λ_{x1}	λ_{x2}	λ_{x3}	λ_{x4}
1a	100	97	19.0	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
			(6.3)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
1b	100	97	18.0	0.89	0.90	0.89	0.90	0.90	0.90	0.90	0.89
			(6.1)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.05)	(0.05)
1c	99	95	17.9	0.71	0.71	0.71	0.70	0.70	0.71	0.71	0.73
			(5.3)	(0.03)	(0.03)	(0.04)	(0.04)	(0.07)	(0.07)	(0.13)	(0.18)
1d	59	24	16.4	0.33	0.33	0.34	0.34	0.96	0.94	1.11	0.96
			(5.7)	(0.07)	(0.08)	(0.15)	(0.20)	(2.32)	(2.10)	(2.36)	(2.16)
2a	100	93	19.1	0.95	0.95	0.95	0.95	0.89	0.95	0.95	0.95
			(6.4)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
2b	100	82	19.3	0.95	0.95	0.95	0.95	0.70	0.96	0.95	0.95
			(6.1)	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)	(0.04)
2c	100	64	17.4	0.95	0.95	0.95	0.95	0.31	0.98	0.95	0.95
			(6.1)	(0.01)	(0.01)	(0.01)	(0.01)	(0.06)	(0.15)	(0.04)	(0.04)
3a	100	90	18.6	0.95	0.95	0.95	0.95	0.90	0.89	0.95	0.95
			(6.3)	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.04)	(0.05)
3b	100	99	17.4	0.95	0.95	0.95	0.95	0.70	0.71	0.95	0.95
			(6.1)	(0.01)	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.03)	(0.03)
3c	98	89	17.3	0.95	0.95	0.95	0.95	0.32	0.32	0.95	0.95
			(5.7)	(0.01)	(0.01)	(0.01)	(0.01)	(0.11)	(0.10)	(0.04)	(0.04)
4a	100	94	17.6	0.95	0.95	0.95	0.89	0.95	0.95	0.95	0.95
			(6.3)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
4b	100	96	18.5	0.95	0.95	0.95	0.71	0.95	0.95	0.94	0.95
			(6.2)	(0.01)	(0.01)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)
4c	100	80	17.5	0.95	0.95	0.95	0.32	0.95	0.95	0.95	0.95
			(5.8)	(0.01)	(0.01)	(0.08)	(0.05)	(0.02)	(0.02)	(0.04)	(0.04)
5a	100	90	18.8	0.95	0.95	0.89	0.90	0.95	0.95	0.95	0.95
			(6.1)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)	(0.04)
5b	100	92	18.5	0.95	0.95	0.70	0.71	0.95	0.95	0.96	0.94
			(6.0)	(0.01)	(0.01)	(0.04)	(0.03)	(0.02)	(0.02)	(0.04)	(0.04)
5c	100	60	17.2	0.95	0.95	0.30	0.30	0.94	0.96	0.95	0.95
			(7.5)	(0.01)	(0.01)	(0.08)	(0.08)	(0.03)	(0.03)	(0.07)	(0.07)

The remaining results in Table 3 are reported as averages over all simulations for which a solution was found by LISREL, with the standard deviation across simulations shown in brackets. The fourth column presents the chi-square test statistic X^2 for assessing overall model fit. Since the posited model was known to hold exactly in these simulations, asymptotically X^2 should follow a chi-square distribution with 18 degrees of freedom (cf. Section 2.2). Thus, its theoretical mean and standard deviation were 18 and 6, respectively. Under the normal approximation, a 99% confidence interval for the mean of X^2 over R_c simulations is given by

$$\left(18 - \frac{2.58 \times 6}{\sqrt{R_c}}, 18 + \frac{2.58 \times 6}{\sqrt{R_c}}\right).$$

In terms of these confidence limits, no significant deviations were found.[9]

The remaining columns present the estimated validities of the observed variables. The fixed parameter λ_{y3} is left out. The average values are mostly in line with the theoretical validities in Table 2. Erratic behaviour occurred only for model 1d, in particular for the x variables. It is interesting to note that the standard deviations across simulations became larger with increasing error variances, and more so for x variables than for y variables. Consider for instance model 1c. Although the estimated validities for this model agree well with their theoretical counterparts *on average*, a single point estimate for λ_{x4} might be off by more than ± 0.1 even in non-exceptional cases. Such a large deviation could lead to the wrong conclusion about which data source contains the most valid measurement of a particular concept.

We also examined the estimates for the structural parameters of the model, i.e. the β, γ, ψ, and φ parameters. For the sake of brevity, these are not tabulated here. For most models, the estimated parameters were in line with their theoretical values; recall that, by construction, the theoretical values of the *structural* parameters are given in Table 1. Substantial deviations occurred only for some of the models with an error variance equal to 0.9, in particular model 1d.

Turning to the results with $N = 60000$, these were clear-cut and we do not report them in a separate table. For all models considered here, LISREL converged to a solution in all simulations and no negative variance estimates occurred. In all cases, the average estimated factor loadings agreed with the theoretical validities up to two decimal places. Moreover, the standard deviations across simulations were substantially lower than for $N = 600$. Thus, it appears that the problems that occurred with $N = 600$ for models with high error variances can be attributed to the relatively small sample size. If the sample is large enough, these problems disappear, at least in this idealised situation.

3.3 Misspecified Measurement Models

For our second set of simulations, we considered several variations on the data generating model of Section 3.2 that cause this model to depart from the SEM of Section 2.2.

[9] If a 95% confidence interval were used instead, then the values for models 1d and 2b would be significant. In fact, the use of a normal confidence interval seems problematic for model 1d because the $R_c = 59$ simulations for which LISREL converged to a solution may form a selective subset.

The aim of this part of the simulation study is to see the effect of model misspecification on the estimated validities.

We considered three basic forms of model misspecification:[10]

- correlated measurement errors in two variables that measure the same concept (models 6 and 7 below);
- correlated measurement errors in two variables that measure different concepts (models 8 and 9 below); and
- measurement errors in a variable for which no alternative measure is available (model 10 below).

These three departures from the data generating model of Section 3.2 could all arise in practice due to limitations of the data collection process. Correlated errors in variables that measure the same concept may occur if these variables have similar measurement procedures. This would happen, for instance, with concepts for which essentially only one type of measurement procedure is available. Correlated errors of the second type mentioned above may occur in practice when related variables are measured using the same instrument – for instance, as related questions in the same survey.

The third type of misspecification may occur in practice when it is impossible to obtain more than one measure for a certain concept in the SEM. This happened for instance with 'occupational level' (η_2 measured only by y_3) in Bakker (2012). To have an identified model, the factor loading of the single measure must be fixed to a constant value. Typically, the factor loading is fixed to 1 and the measure is implicitly assumed to have perfect validity. As we do not expect this assumption to hold exactly in any practical application, it is interesting to know to what extent the inclusion of an imperfect single measure with a fixed factor loading of 1 in the model affects the estimated validities of the other variables.

Technically, the first two types of misspecification were achieved by introducing nonzero elements $\theta^*_{\varepsilon kk'}$ $(k \neq k')$ in off-diagonal positions of the matrix Θ^*_ε in (8). That is, the measurement errors ε_k and $\varepsilon_{k'}$ were given a correlation $\rho^*_{\varepsilon kk'}$ by setting

$$\theta^*_{\varepsilon kk'} = \rho^*_{\varepsilon kk'} \sqrt{\theta^*_{\varepsilon kk} \theta^*_{\varepsilon k'k'}}$$

for the appropriate entry in Θ^*_ε. Measurement errors in y_3 were introduced analogously to errors in the other variables, i.e. by setting $\theta^*_{\varepsilon 33}$ to a non-zero value.

Before presenting the results of the simulation study, we report some theoretical findings. Recall that, in the absence of model misspecification, the factor loadings in the SEM are consistent estimates of the theoretical validities. Now consider the same model (9) with misspecification due to correlated errors in two variables that measure

[10] We only considered misspecifications in the measurement model and not in the structural part of the SEM. As explained in Section 2.1, it is essential to the method that the structural part of the model has been well-studied in previous research. Therefore, we assume that no problems are encountered in specifying the structural part.

the same concept, say y_4 and y_5 with $\theta^*_{\varepsilon 45} \neq 0$. In Appendix A we derive that for large samples – under certain conditions – the estimated factor loadings for y_4 and y_5 satisfy

$$\lambda_{y4} \approx \lambda^*_{y4}\sqrt{1+\frac{\theta^*_{\varepsilon 45}}{\lambda^*_{y4}\lambda^*_{y5}}},$$
$$\lambda_{y5} \approx \lambda^*_{y5}\sqrt{1+\frac{\theta^*_{\varepsilon 45}}{\lambda^*_{y4}\lambda^*_{y5}}},$$

when the model is estimated under the erroneous assumption that $\theta_{\varepsilon 45} = 0$. In this situation, the factor loadings of the other variables are still approximately equal to their theoretical validities λ^*_{yk} and λ^*_{xl}. That is, the validity of y_4 and y_5 is *over*estimated in the presence of positively correlated errors, and *under*estimated when the errors are negatively correlated. In addition, the covariance of η_3 (the common factor measured by y_4 and y_5) with the other latent variables is divided by the same factor $\sqrt{1+\theta^*_{\varepsilon 45}/\lambda^*_{y4}\lambda^*_{y5}}$, while the other latent covariances are estimated consistently. This implies in particular that the estimates for all structural parameters that do not involve η_3 are not affected.

An obvious necessary condition for this property to hold is that the expression under the square root must be positive, i.e.

$$\theta^*_{\varepsilon 45} > -\lambda^*_{y4}\lambda^*_{y5}. \tag{10}$$

This condition is violated for sufficiently large negative correlations, in which case the above result does not hold. In summary, we can say that – for sufficiently large samples – this type of model misspecification has only a 'local effect' in the sense that the estimated validities of the other variables are not affected. We also argue in Appendix A that the overall model fit as measured by X^2 should not be affected by correlated errors of this type. In other words: this type of model misspecification cannot be detected by means of the chi-square test alone. Finally, it is shown in Appendix A that, with only two indicators per latent variable, the correct model with $\theta_{\varepsilon 45}$ as a free parameter is not identified. Thus, it is not obvious how to resolve this type of misspecification (when detected), given the available data.

Turning to the second type of misspecification considered here (correlated errors in variables that measure different concepts), a similar analysis shows that the X^2 statistic *is* sensitive to the presence of correlated errors of this type; see Appendix A. In addition, model identification is less likely to be a problem in this case. An analytical expression for the large-sample values of the factor loadings for this situation could not be obtained.

Finally, consider the third type of model misspecification (measurement errors in y_3 while the factor loading λ_{y3} is fixed to 1). We show in Appendix A that the validities of all variables except y_3 are estimated consistently. The covariance of η_2 (the factor measured by y_3) with all the other latent variables is multiplied by $\lambda^*_{y3} = \sqrt{1-\theta^*_{\varepsilon 33}}$, i.e. attenuated towards 0. Hence, the values of the structural parameters that involve η_2 are incorrect, while the other parameters are estimated consistently. As with the first type, we also argue in Appendix A that the chi-square test is not capable of detecting this type of misspecification, because the distribution of X^2 is not affected by it. As mentioned before, the correct model with λ_{y3} as a free parameter is not identified.

Using the same basic set-up as in Section 3.2, we performed simulations for the following data generating models. (In these model descriptions, all non-specified error variances are set to 0.1 except for $\theta^*_{\varepsilon 33}$, which is 0 unless stated otherwise.)

6. Increasing amounts of lightly correlated measurement errors in both observed variables for one factor:

    a. error variances of y_4 and y_5 equal to 0.2; $\rho^*_{\varepsilon 45} = +0.1$;

    b. error variances of y_4 and y_5 equal to 0.5; $\rho^*_{\varepsilon 45} = +0.1$;

    c. error variances of y_4 and y_5 equal to 0.9; $\rho^*_{\varepsilon 45} = +0.1$;

    d. same as model 6a with $\rho^*_{\varepsilon 45} = -0.1$;

    e. same as model 6b with $\rho^*_{\varepsilon 45} = -0.1$;

    f. same as model 6c with $\rho^*_{\varepsilon 45} = -0.1$.

7. Increasing amounts of heavily correlated measurement errors in both observed variables for one factor:

    a. error variances of y_4 and y_5 equal to 0.2; $\rho^*_{\varepsilon 45} = +0.5$;

    b. error variances of y_4 and y_5 equal to 0.5; $\rho^*_{\varepsilon 45} = +0.5$;

    c. error variances of y_4 and y_5 equal to 0.9; $\rho^*_{\varepsilon 45} = +0.5$;

    d. same as model 7a with $\rho^*_{\varepsilon 45} = -0.5$;

    e. same as model 7b with $\rho^*_{\varepsilon 45} = -0.5$;

    f. same as model 7c with $\rho^*_{\varepsilon 45} = -0.5$.

8. Increasing amounts of lightly correlated measurement errors in two observed variables for different factors:

    a. error variances of y_1 and y_4 equal to 0.2; $\rho^*_{\varepsilon 14} = +0.1$;

    b. error variances of y_1 and y_4 equal to 0.5; $\rho^*_{\varepsilon 14} = +0.1$;

    c. error variances of y_1 and y_4 equal to 0.9; $\rho^*_{\varepsilon 14} = +0.1$;

    d. same as model 8a with $\rho^*_{\varepsilon 14} = -0.1$;

    e. same as model 8b with $\rho^*_{\varepsilon 14} = -0.1$;

    f. same as model 8c with $\rho^*_{\varepsilon 14} = -0.1$.

9. Increasing amounts of heavily correlated measurement errors in two observed variables for different factors:

    a. error variances of y_1 and y_4 equal to 0.2; $\rho^*_{\varepsilon 14} = +0.5$;

    b. error variances of y_1 and y_4 equal to 0.5; $\rho^*_{\varepsilon 14} = +0.5$;

    c. error variances of y_1 and y_4 equal to 0.9; $\rho^*_{\varepsilon 14} = +0.5$;

    d. same as model 9a with $\rho^*_{\varepsilon 14} = -0.5$;

    e. same as model 9b with $\rho^*_{\varepsilon 14} = -0.5$;

    f. same as model 9c with $\rho^*_{\varepsilon 14} = -0.5$.

10. Increasing amounts of measurement error in variable y_3:

    a. error variance of y_3 equal to 0.1;

    b. error variance of y_3 equal to 0.2;

    c. error variance of y_3 equal to 0.5;

    d. error variance of y_3 equal to 0.9.

Table 4: True validities and expected validity estimates for models 6 and 7.

model	6a	6b	6c	6d	6e	6f
true validity	0.89	0.71	0.32	0.89	0.71	0.32
expected estimate	0.91	0.74	0.44	0.88	0.67	0.10
model	7a	7b	7c	7d	7e	7f
true validity	0.89	0.71	0.32	0.89	0.71	0.32
expected estimate	0.95	0.87	0.74	0.84	0.50	–

It should be noted that all instances of models 6 and 7 satisfy the above condition (10), except for model 7f which has $\theta^*_{\varepsilon 45} = -0.45 < -0.1 = -\lambda^*_{y4}\lambda^*_{y5}$. Table 4 lists the theoretical validities and the expected values of the biased validity estimates for y_4 and y_5 under these models.

As before, we performed simulations with $N = 600$ and $N = 60000$. We shall focus on the results with $N = 600$ here, as the large-sample results were very similar. Table 5 displays the results for models 6a through 7f. As expected, the model fit was not affected by the misspecification in these models: the average values of X^2 were all within their 99% confidence intervals under the null hypothesis that the model is correct (which we know to be false in this case). Apart from model 6f, the average estimated validities of y_4 and y_5 were close to their expected (biased) values given above.[11] For model 6f, severe problems occurred with negative variance estimates. For all these models, the average estimated validities of the other variables were close to their true values, which illustrates the 'local effect' of this type of misspecification.

This local effect could also be seen in the estimated values for the structural parameters (second panel in Table 5). We only tabulated the results for the parameters that are directly related to η_3; as expected, the average values for the other structural parameters did not deviate substantially from their true values (i.e. the values from Section 2.2). Table 5 shows some large deviations in the values for the parameters related to η_3. Interestingly though, the estimated direct effects started to deviate strongly from their true values only for large correlations and/or large error variances. For instance, the average values of the β and γ parameters for models 7a and 7b (with $\rho^*_{\varepsilon 45} = 0.5$ and error variances of 0.2 and 0.5, respectively) were still quite close to the true values; in particular, these values might not be considered implausible from a subject-matter point of view. Together with the fact that the model fit measured by X^2 remains the same, this makes this type of model misspecification rather difficult to detect. Moreover, the estimated validities for y_4 and y_5 were already quite biased for model 7a and especially model 7b, which shows the importance of detecting this form of misspecification.

Table 6 displays the results for models 8a through 9f, for which the measurement errors in y_1 and y_4 were correlated. Here, a deterioration was observed in the overall model fit.[12] Using a 99% confidence interval as before, the average values of X^2 were all significant. Note however that models 8a through 8f had average X^2 values between 20 and 24. At 18 degrees of freedom, the SEM would normally not be rejected for

[11] With $N = 60000$, the results for model 6f did agree with the expected values.

[12] This deterioration was even larger for $N = 60000$.

Table 5: Simulation results for misspecified models 6a–7f with $N = 600$ *(using the same format as in Table 3).*

model	R_c	R_v	X^2	λ_{y1}	λ_{y2}	λ_{y4}	λ_{y5}	λ_{x1}	λ_{x2}	λ_{x3}	λ_{x4}
6a	100	91	17.5	0.95	0.95	0.90	0.91	0.95	0.95	0.96	0.94
			(6.7)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
6b	100	89	17.9	0.95	0.95	0.74	0.75	0.95	0.95	0.95	0.95
			(5.6)	(0.01)	(0.01)	(0.03)	(0.03)	(0.02)	(0.02)	(0.05)	(0.05)
6c	100	79	17.1	0.95	0.95	0.44	0.43	0.95	0.95	0.95	0.95
			(5.3)	(0.01)	(0.01)	(0.08)	(0.08)	(0.03)	(0.03)	(0.07)	(0.07)
6d	100	97	18.3	0.95	0.95	0.89	0.88	0.95	0.95	0.95	0.95
			(6.8)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
6e	100	86	17.4	0.95	0.95	0.67	0.67	0.95	0.95	0.96	0.94
			(6.3)	(0.01)	(0.01)	(0.03)	(0.04)	(0.02)	(0.02)	(0.05)	(0.05)
6f	100	12	18.4	0.95	0.95	0.43	0.44	0.95	0.94	0.95	0.96
			(5.6)	(0.01)	(0.01)	(0.39)	(0.44)	(0.03)	(0.03)	(0.08)	(0.08)
7a	100	95	17.9	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
			(5.9)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
7b	100	85	16.8	0.95	0.95	0.87	0.86	0.95	0.95	0.94	0.95
			(5.6)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	(0.05)	(0.05)
7c	100	57	17.2	0.95	0.95	0.75	0.74	0.95	0.95	0.96	0.94
			(5.6)	(0.01)	(0.01)	(0.08)	(0.07)	(0.03)	(0.03)	(0.09)	(0.08)
7d	100	98	18.6	0.95	0.95	0.84	0.84	0.95	0.95	0.95	0.95
			(5.6)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
7e	100	79	18.0	0.95	0.95	0.50	0.50	0.95	0.95	0.95	0.95
			(5.1)	(0.01)	(0.01)	(0.04)	(0.04)	(0.02)	(0.02)	(0.05)	(0.05)
7f	100	0	17.2	0.95	0.95	1.00	1.01	0.95	0.95	0.94	0.96
			(5.1)	(0.01)	(0.01)	(0.00)	(0.35)	(0.03)	(0.03)	(0.07)	(0.08)

model	β_{31}	β_{32}	γ_{31}	γ_{32}	ψ_{33}
6a	0.35	0.31	0.32	−0.18	0.55
	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
6b	0.34	0.30	0.31	−0.17	0.58
	(0.06)	(0.05)	(0.05)	(0.04)	(0.05)
6c	0.26	0.23	0.25	−0.13	0.73
	(0.10)	(0.10)	(0.08)	(0.08)	(0.09)
6d	0.36	0.32	0.33	−0.18	0.53
	(0.05)	(0.05)	(0.04)	(0.03)	(0.03)
6e	0.38	0.33	0.34	−0.19	0.49
	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)
6f	0.63	0.57	0.56	−0.31	−2.15
	(0.62)	(0.59)	(0.60)	(0.38)	(6.24)
7a	0.34	0.30	0.31	−0.17	0.58
	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)
7b	0.29	0.26	0.25	−0.16	0.70
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)
7c	0.16	0.13	0.14	−0.08	0.91
	(0.07)	(0.06)	(0.04)	(0.05)	(0.03)
7d	0.38	0.34	0.35	−0.19	0.47
	(0.05)	(0.04)	(0.04)	(0.03)	(0.03)
7e	0.50	0.45	0.45	−0.26	0.08
	(0.07)	(0.07)	(0.07)	(0.05)	(0.12)
7f	0.11	0.10	0.10	−0.06	−0.44
	(0.03)	(0.03)	(0.03)	(0.03)	(0.16)

Table 6: Simulation results for misspecified models 8a–9f with $N = 600$ *(using the same format as in Table 3).*

model	R_c	R_v	X^2	λ_{y1}	λ_{y2}	λ_{y4}	λ_{y5}	λ_{x1}	λ_{x2}	λ_{x3}	λ_{x4}
8a	99	98	20.6	0.90	0.95	0.90	0.95	0.95	0.95	0.95	0.95
			(6.3)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
8b	100	96	22.3	0.71	0.94	0.71	0.95	0.95	0.95	0.95	0.95
			(7.9)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.04)
8c	100	68	23.0	0.33	0.94	0.31	0.96	0.95	0.95	0.95	0.95
			(7.0)	(0.04)	(0.08)	(0.05)	(0.08)	(0.02)	(0.02)	(0.04)	(0.04)
8d	100	95	20.2	0.89	0.95	0.89	0.95	0.95	0.95	0.95	0.95
			(7.3)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
8e	100	97	22.2	0.71	0.95	0.71	0.95	0.95	0.95	0.95	0.95
			(7.4)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)
8f	100	57	23.9	0.31	0.97	0.32	0.96	0.95	0.95	0.96	0.94
			(8.6)	(0.05)	(0.08)	(0.05)	(0.08)	(0.02)	(0.02)	(0.04)	(0.04)
9a	100	96	102	0.91	0.93	0.91	0.93	0.95	0.95	0.95	0.95
			(20)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
9b	100	94	156	0.73	0.93	0.72	0.92	0.95	0.95	0.95	0.95
			(20)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.04)
9c	100	76	183	0.32	0.95	0.32	0.92	0.94	0.95	0.94	0.95
			(24)	(0.05)	(0.08)	(0.05)	(0.08)	(0.02)	(0.02)	(0.04)	(0.04)
9d	100	94	101	0.88	0.96	0.89	0.96	0.95	0.95	0.95	0.95
			(17)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
9e	100	94	155	0.70	0.96	0.70	0.97	0.95	0.95	0.95	0.95
			(24)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.04)	(0.04)
9f	100	43	186	0.31	0.99	0.31	0.97	0.95	0.95	0.95	0.94
			(26)	(0.06)	(0.13)	(0.05)	(0.12)	(0.02)	(0.02)	(0.05)	(0.04)

these values in a chi-square test. The average X^2 values for models 9a through 9f, on the other hand, would lead to a rejection of the SEM in practice.

Interestingly, the average values of the factor loadings for these models were close to the theoretical validities for all variables, including y_1 and y_4.[13] Thus, this type of correlated measurement error appears to have almost no effect on the correctness of the estimated validities. The same was true of the estimated structural parameters (not shown here). However, these estimated validities and structural parameters are unusable if the model as a whole is rejected by the chi-square test.

Finally, Table 7 displays the results for model 10. As expected, it is seen that in terms of X^2 the SEM fitted the data as well as in the case of correct specification. The validities of all variables beside y_3 were also correctly estimated. The measurement errors in y_3 did have an effect on the structural parameters of the model, as can be seen in the second and third panel of the table. As predicted, some large effects occurred for parameters that are directly related to η_2, the concept measured by y_3, i.e. β_{21}, β_{32}, γ_{21}, γ_{22}, and ψ_{22}.[14]

[13] Recall that the true validity of y_1 and y_4 equals: 0.89 for models 8a, 8d, 9a, and 9d; 0.71 for models 8b, 8e, 9b, and 9e; and 0.32 for models 8c, 8f, 9c, and 9f.

[14] Contrary to expectations, some smaller systematic deviations were also seen for β_{31}, γ_{31}, γ_{32}, and ψ_{33}. These effects persisted with $N = 60000$. A closer inspection revealed that, for model 10 as well as other models, the latent covariance $\text{cov}(\eta_2, \xi_1)$ was consistently overestimated in these simulations. This is probably an artefact due to a rounding problem in the data generating model; recall from (7) that this particular latent covariance is close to 0.

Table 7: Simulation results for misspecified models 10a–10d with $N = 600$ (using the same format as in Table 3).

model	R_c	R_v	X^2	λ_{y1}	λ_{y2}	λ_{y4}	λ_{y5}	λ_{x1}	λ_{x2}	λ_{x3}	λ_{x4}
10a	100	96	17.6	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
			(6.0)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
10b	100	93	17.5	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
			(5.3)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
10c	100	91	18.5	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
			(5.7)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)
10d	100	96	17.4	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
			(5.7)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)

model	β_{21}	β_{31}	β_{32}	γ_{11}	γ_{12}	γ_{21}	γ_{22}	γ_{31}	γ_{32}
10a	0.56	0.39	0.27	−0.21	−0.00	0.12	−0.08	0.33	−0.18
	(0.03)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.03)	(0.03)
10b	0.54	0.41	0.26	−0.22	−0.00	0.11	−0.08	0.34	−0.18
	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)
10c	0.42	0.47	0.18	−0.23	−0.00	0.08	−0.06	0.35	−0.20
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
10d	0.19	0.53	0.07	−0.22	−0.01	0.04	−0.03	0.37	−0.21
	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)	(0.03)

model	ψ_{11}	ψ_{22}	ψ_{33}	φ_{12}
10a	0.95	0.69	0.55	−0.07
	(0.02)	(0.03)	(0.03)	(0.05)
10b	0.95	0.72	0.55	−0.07
	(0.02)	(0.03)	(0.03)	(0.05)
10c	0.95	0.82	0.57	−0.08
	(0.02)	(0.03)	(0.03)	(0.05)
10d	0.95	0.96	0.60	−0.07
	(0.02)	(0.02)	(0.03)	(0.04)

4 Simulation Study, Part 2: Realistic Data

4.1 Introduction

For the second part of the simulation study, we worked with the original data that were analysed by Bakker (2012), as summarised in Section 2.2. We introduced additional errors into this data set by means of a random hot deck imputation method. Hot deck imputation is a commonly used solution for missing data; see e.g. De Waal et al. (2011). It involves the replacement of each missing value by an observed value from a randomly drawn record in the same data set. In a slight deviation from the normal use, we used this technique to randomly replace observed values by (potentially) different observed values, thus introducing artificial errors into the data set.

The resulting artificial data sets are more realistic than the data sets used in Section 3 in two ways. Firstly, they follow a real-world distribution rather than an artificial multinormal one. Secondly, error mechanisms occurring in practice are usually 'intermittent' in the sense that only some of the observed values are incorrect (Di Zio et al., 2008). The hot deck imputation method produces artificial errors according to an intermittent mechanism, while the data generating models of Section 3 produce observed values that are incorrect with probability 1.

We used the same SEM for estimating validity as before, i.e. the model from Figure 1. As in Section 3, we started with a set of simulations for which this SEM is correctly specified (Section 4.2). Subsequently, we examined the effects of different types of model misspecification (Section 4.3).

4.2 Correctly Specified Measurement Models

The cases where the model is correctly specified were obtained by applying random hot deck imputation independently to each observed variable. That is, separately for each variable, we randomly selected a subset of the records in the data set and replaced the observed values in these records with values drawn at random from the original data set.[15] By construction, the errors introduced in this manner are uncorrelated between variables. Note that the data generating method is specified by choosing, for each variable in the data set, the fraction of records to impute, say $0 \leq \pi \leq 1$.

Suppose that the above imputation method is applied to one variable, say z_1. Let π_1 denote the fraction of records in the data set that are imputed. In Appendix B, it is shown that the following properties hold *in expectation*, i.e. on average if the imputation procedure were applied to the same data set an infinite number of times:

- The mean and variance of z_1 are the same before and after imputation.
- The correlation between z_1 and any other variable z_2 is attenuated by a factor $1 - \pi_1$. That is, if the original correlation was $\rho(z_1, z_2)$, then the correlation after imputation equals $(1 - \pi_1)\rho(z_1, z_2)$.

[15] Usually, hot deck imputations are drawn from the non-imputed part of the data set. Using the full data set as a source of imputations is convenient here because it makes the theoretical properties of the method easier to derive.

- More generally, if the imputation method is applied independently also to z_2 (with π_2 the fraction of imputed records), then the correlation between z_1 and z_2 after imputation equals $(1-\pi_1)(1-\pi_2)\rho(z_1,z_2)$.

Note that these properties only hold approximately when the imputation method is applied once.

To put this in perspective, we posit a measurement model of the basic form (1) for both z_1 and z_2 after imputation, i.e.

$$\tilde{z}_1 = \lambda_1 z_1 + \varepsilon_1, \tag{11}$$

$$\tilde{z}_2 = \lambda_2 z_2 + \varepsilon_2, \tag{12}$$

where a tilde indicates that a variable has been imputed. Denoting the variances of the error terms by θ_1 and θ_2, it follows that

$$\begin{aligned}\rho(\tilde{z}_1,z_2) &= \lambda_1\rho(z_1,z_2) = \sqrt{1-\theta_1}\rho(z_1,z_2),\\ \rho(\tilde{z}_1,\tilde{z}_2) &= \lambda_1\lambda_2\rho(z_1,z_2) = \sqrt{(1-\theta_1)(1-\theta_2)}\rho(z_1,z_2).\end{aligned}$$

Comparing this with the expressions derived in Appendix B and solving for π_1 and π_2 yields:

$$\begin{aligned}\pi_1 &= 1-\lambda_1 = 1-\sqrt{1-\theta_1},\\ \pi_2 &= 1-\lambda_2 = 1-\sqrt{1-\theta_2}.\end{aligned}$$

Recall that the SEM is estimated using only the information in the observed correlation matrix. The above result therefore suggests, albeit by a heuristic argument, that applying the hot deck imputation method to a variable z with $\pi = 1-\sqrt{1-\theta}$ as the fraction of imputed records has approximately the same effect on the validity of z as introducing normally distributed errors with an error variance of θ according to the model of Section 3.2. Thus, for instance, an error variance of 0.1 under the normal model reduces the validity by approximately the same factor as taking $\pi = 0.051$.

Since the validity of the observed variables under the normal model is known, we use the above equivalence to derive approximate theoretical validities for the observed variables with different choices of π. For convenience, we take values for π that match the error variances that were used in the previous simulations. Table 8 contains the resulting theoretical validities.

Table 8: Approximate theoretical validities under the hot deck imputation method.

		approximate validity							
π	θ	y_1	y_2	y_4	y_5	x_1	x_2	x_3	x_4
0	0	0.82	0.94	0.95	0.87	1.00	1.00	1.00	1.00
0.051	0.1	0.77	0.89	0.90	0.82	0.95	0.95	0.95	0.95
0.106	0.2	0.73	0.84	0.85	0.78	0.89	0.89	0.89	0.90
0.293	0.5	0.58	0.67	0.67	0.61	0.71	0.71	0.70	0.71
0.684	0.9	0.26	0.30	0.30	0.27	0.32	0.32	0.32	0.32

Note that the validities of Section 2.2 are reproduced when no hot deck imputation is applied ($\pi = 0$). The other validities are obtained by multiplying these original validities by the appropriate factor $\lambda = 1 - \pi = \sqrt{1 - \theta}$.

For the computer simulations, we generated data sets with the random hot deck method according to the following specifications. (In these descriptions, all imputation fractions π are 0 unless stated otherwise.)

11. Increasing amounts of measurement error in all observed variables:
    a. all imputation fractions equal to 0.051;
    b. all imputation fractions equal to 0.106;
    c. all imputation fractions equal to 0.293;
    d. all imputation fractions equal to 0.684.
12. Increasing amounts of measurement error in one observed variable for one exogenous factor:
    a. imputation fraction of x_1 equal to 0.051;
    b. imputation fraction of x_1 equal to 0.106;
    c. imputation fraction of x_1 equal to 0.293;
    d. imputation fraction of x_1 equal to 0.684.
13. Increasing amounts of measurement error in both observed variables for one exogenous factor:
    a. imputation fractions of x_1 and x_2 equal to 0.051;
    b. imputation fractions of x_1 and x_2 equal to 0.106;
    c. imputation fractions of x_1 and x_2 equal to 0.293;
    d. imputation fractions of x_1 and x_2 equal to 0.684.
14. Increasing amounts of measurement error in one observed variable for one endogenous factor:
    a. imputation fraction of y_5 equal to 0.051;
    b. imputation fraction of y_5 equal to 0.106;
    c. imputation fraction of y_5 equal to 0.293;
    d. imputation fraction of y_5 equal to 0.684.
15. Increasing amounts of measurement error in both observed variables for one endogenous factor:
    a. imputation fractions of y_4 and y_5 equal to 0.051;
    b. imputation fractions of y_4 and y_5 equal to 0.106;
    c. imputation fractions of y_4 and y_5 equal to 0.293;
    d. imputation fractions of y_4 and y_5 equal to 0.684.

As before, we performed $R = 100$ simulations for each model. Recall from Section 2.2 that the sample size was $N = 574$.

Table 9 displays the results in the same format as before. These were generally in line with the results found previously using completely artificial data with $N = 600$. In particular, with small to moderate amounts of introduced measurement errors, the estimated factor loadings agreed well on average with the theoretical validities in Table 8. With large amounts of measurement errors, the behaviour of the estimates became erratic. In addition, problems with convergence and large negative variance estimates started to occur. The results for the structural parameters (not tabulated here) were similar. Finally, for all models, the chi-square statistic was much higher than 18, its

Table 9: Simulation results with the hot deck method for correctly specified models (using the same format as in Table 3).

model	R_c	R_v	X^2	λ_{y1}	λ_{y2}	λ_{y4}	λ_{y5}	λ_{x1}	λ_{x2}	λ_{x3}	λ_{x4}
11a	99	58	58	0.77	0.89	0.89	0.83	0.94	0.95	0.94	0.98
			(19)	(0.03)	(0.03)	(0.04)	(0.03)	(0.05)	(0.05)	(0.11)	(0.23)
11b	99	65	57	0.72	0.84	0.85	0.77	0.90	0.89	0.92	0.89
			(19)	(0.05)	(0.06)	(0.05)	(0.04)	(0.08)	(0.08)	(0.16)	(0.13)
11c	90	58	47	0.56	0.66	0.67	0.61	0.72	0.76	0.92	0.71
			(16)	(0.09)	(0.09)	(0.08)	(0.07)	(0.21)	(0.25)	(0.72)	(0.29)
11d	31	6	42	0.67	0.97	0.90	0.34	1.35	1.48	0.55	0.87
			(13)	(1.67)	(3.52)	(2.09)	(0.34)	(2.68)	(3.03)	(1.24)	(2.07)
12a	100	86	53	0.82	0.94	0.95	0.87	0.94	1.00	1.00	1.00
			(7)	(0.00)	(0.00)	(0.00)	(0.00)	(0.03)	(0.03)	(0.00)	(0.00)
12b	100	82	54	0.82	0.94	0.95	0.87	0.89	1.00	1.00	1.00
			(9)	(0.00)	(0.00)	(0.00)	(0.00)	(0.06)	(0.04)	(0.00)	(0.00)
12c	100	63	55	0.82	0.94	0.95	0.87	0.71	1.00	1.00	1.00
			(8)	(0.00)	(0.00)	(0.00)	(0.00)	(0.08)	(0.09)	(0.00)	(0.00)
12d	96	53	55	0.82	0.94	0.95	0.87	0.33	1.09	1.00	1.00
			(9)	(0.00)	(0.00)	(0.00)	(0.00)	(0.11)	(0.35)	(0.00)	(0.00)
13a	100	94	54	0.82	0.94	0.95	0.87	0.95	0.95	1.00	1.00
			(9)	(0.00)	(0.01)	(0.00)	(0.00)	(0.05)	(0.04)	(0.00)	(0.00)
13b	100	92	53	0.82	0.94	0.95	0.86	0.90	0.89	1.00	1.00
			(11)	(0.01)	(0.01)	(0.01)	(0.01)	(0.07)	(0.07)	(0.00)	(0.00)
13c	100	97	49	0.82	0.94	0.96	0.86	0.71	0.72	1.00	1.00
			(11)	(0.01)	(0.01)	(0.01)	(0.01)	(0.12)	(0.12)	(0.00)	(0.00)
13d	71	59	38	0.83	0.92	0.99	0.83	0.39	0.54	1.00	1.00
			(11)	(0.01)	(0.01)	(0.01)	(0.01)	(0.34)	(1.20)	(0.00)	(0.00)
14a	100	100	41	0.82	0.94	0.95	0.82	1.00	1.00	1.00	1.00
			(8)	(0.00)	(0.00)	(0.02)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)
14b	100	100	36	0.82	0.94	0.95	0.78	1.00	1.00	1.00	1.00
			(11)	(0.00)	(0.00)	(0.02)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
14c	100	90	29	0.82	0.94	0.96	0.61	1.00	1.00	1.00	1.00
			(11)	(0.00)	(0.00)	(0.05)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)
14d	93	64	23	0.82	0.94	0.97	0.28	1.00	1.00	1.00	1.00
			(8)	(0.00)	(0.00)	(0.15)	(0.07)	(0.00)	(0.00)	(0.00)	(0.00)
15a	98	98	37	0.81	0.94	0.90	0.83	1.00	1.00	1.00	1.00
			(9)	(0.00)	(0.00)	(0.03)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)
15b	99	99	30	0.81	0.94	0.85	0.79	1.00	1.00	1.00	1.00
			(10)	(0.01)	(0.01)	(0.05)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
15c	91	91	28	0.81	0.95	0.67	0.63	1.00	1.00	1.00	1.00
			(11)	(0.01)	(0.01)	(0.07)	(0.07)	(0.00)	(0.00)	(0.00)	(0.00)
15d	94	66	26	0.80	0.95	0.66	0.36	1.00	1.00	1.00	1.00
			(12)	(0.01)	(0.01)	(1.61)	(0.38)	(0.00)	(0.00)	(0.00)	(0.00)

expected value under the assumption of multinormal data. Note however that this assumption does not hold here and recall that Bakker (2012) found $X^2 = 48$ with the original data. For models 11 through 13, the average values of X^2 were all within one standard deviation of this value. Somewhat surprisingly, increasing the amount of errors under models 14 and 15 actually improved the fit of the SEM.

4.3 Misspecified Measurement Models

For the final part of the simulation study, we used realistic data and examined the effects of the three types of model misspecification from Section 3.3. In order to obtain correlated measurement errors, a modification of the above hot deck imputation method was needed. A positive correlation can be achieved, in principle, by drawing the imputed values for two variables simultaneously from the same donor record. Unfortunately, the size of this correlation is fixed and cannot be prescribed. To have more flexibility in the choice of the size of the correlation, we modified the mechanism that selects the records to impute, as follows.

Suppose that we want to introduce correlated measurement errors in two observed variables z_1 and z_2. We randomly select a subset of the records in the data set to impute, in such a way that the fractions of imputed records for z_1 and z_2 are π_1 and π_2, respectively, and that the fraction of records for which both z_1 *and* z_2 are imputed is π_{12}. Technically, this can be achieved by drawing a random value u from the uniform distribution on $(0,1]$ for each record and

- imputing only z_1 for all records with $0 < u \leq \pi_1 - \pi_{12}$;
- imputing z_1 and z_2 for all records with $\pi_1 - \pi_{12} < u \leq \pi_1$;
- imputing only z_2 for all records with $\pi_1 < u \leq \pi_1 + \pi_2 - \pi_{12}$;
- imputing neither of z_1 and z_2 for all records with $\pi_1 + \pi_2 - \pi_{12} < u \leq 1$.

Moreover, for each record only one random donor record is drawn to impute z_1 and/or z_2. In order for this procedure to be well-defined, we require that

$$0 \leq \pi_{12} \leq \min\{\pi_1, \pi_2\}, \text{ and} \tag{13}$$

$$\pi_1 + \pi_2 - \pi_{12} \leq 1. \tag{14}$$

In Appendix B, it is shown that the following properties hold in expectation:

- The mean and variance of z_1 and z_2 are the same before and after imputation.
- If the original correlation between z_1 and z_2 was $\rho(z_1, z_2)$, then the correlation after imputation equals $(1 - \pi_1 - \pi_2 + 2\pi_{12})\rho(z_1, z_2)$.
- If another variable, say z_3, is imputed by the hot deck method independently of z_1 and z_2 (with π_3 the fraction of imputed records), then the correlation between z_1 and z_3 after imputation equals $(1-\pi_1)(1-\pi_3)\rho(z_1, z_3)$, and the correlation between z_2 and z_3 after imputation equals $(1-\pi_2)(1-\pi_3)\rho(z_2, z_3)$.

Again, these properties are only approximately true when the data set is imputed once.

As in Section 4.2, we can assess the theoretical validity of the imputed variables by comparing the above expressions for the correlations with the expressions that would follow from a measurement model of the form (1). To simplify matters, we only discuss the case $\pi_1 = \pi_2$ here, because this is the only case that occurred in the simulations to be discussed below. Let π, then, denote the common imputation fraction of z_1 and z_2, and let λ and θ be their common factor loading and error variance under a model of the form (11)–(12) (this time with correlated error terms). Analogous to the case that the variables were imputed independently, we obtain:

$$\begin{aligned}
\rho(\tilde{z}_1, \tilde{z}_3) &= \sqrt{(1-\theta)(1-\theta_3)}\rho(z_1, z_3), \\
\rho(\tilde{z}_2, \tilde{z}_3) &= \sqrt{(1-\theta)(1-\theta_3)}\rho(z_2, z_3), \\
\rho(\tilde{z}_1, \tilde{z}_2) &= (1-\theta)\rho(z_1, z_2) + \rho_{12}\theta,
\end{aligned}$$

where in the last line ρ_{12} denotes the correlation between ε_1 and ε_2 under the normal measurement model.

Comparing the expressions for $\rho(\tilde{z}_1, \tilde{z}_3)$ and $\rho(\tilde{z}_2, \tilde{z}_3)$ yields $\pi = 1 - \sqrt{1-\theta}$ and $\pi_3 = 1 - \sqrt{1-\theta_3}$ as before. Equating the two expressions for $\rho(\tilde{z}_1, \tilde{z}_2)$ then yields:

$$\left\{1 - 2(1 - \sqrt{1-\theta}) + 2\pi_{12}\right\}\rho(z_1, z_2) = (1-\theta)\rho(z_1, z_2) + \rho_{12}\theta.$$

After some re-arranging of terms, we find the following expression for ρ_{12}:

$$\rho_{12} = \frac{-2 + 2\sqrt{1-\theta} + \theta + 2\pi_{12}}{\theta}\rho(z_1, z_2). \tag{15}$$

Thus, by varying the choice of π_{12} in the hot deck procedure, we can obtain different values for the correlation between the measurement errors. Note however that the range of correlations that can be obtained with this procedure is more limited than with the data generating model of Section 3. The range of possible values for ρ_{12} is limited by the (fixed) value of $\rho(z_1, z_2)$ in the original data set and the choice of error variance θ, or equivalently the imputation fraction π. The latter choice places restrictions on π_{12} through assumptions (13) and (14). In particular, for $\rho(z_1, z_2) > 0$ it can be shown from (15) that ρ_{12} cannot exceed $\rho(z_1, z_2)$.

Using the above extended hot deck imputation method, we created $R = 100$ data sets corresponding to each of the following models. (As before, all imputation fractions π are 0 unless stated otherwise.)

16. Increasing amounts of lightly correlated measurement errors in both observed variables for one factor:
    a. imputation fractions of y_4 and y_5 equal to 0.051 with $\pi_{45} = 0.007$ ($\rho_{45} = 0.1$);
    b. imputation fractions of y_4 and y_5 equal to 0.106 with $\pi_{45} = 0.018$ ($\rho_{45} = 0.1$);
    c. imputation fractions of y_4 and y_5 equal to 0.293 with $\pi_{45} = 0.073$ ($\rho_{45} = 0.1$);
    d. imputation fractions of y_4 and y_5 equal to 0.684 with $\pi_{45} = 0.368$ ($\rho_{45} = 0.24$).
17. Increasing amounts of heavily correlated measurement errors in both observed variables for one factor:
    a. imputation fractions of y_4 and y_5 equal to 0.051 with $\pi_{45} = 0.032$ ($\rho_{45} = 0.5$);

b. imputation fractions of y_4 and y_5 equal to 0.106 with $\pi_{45} = 0.066$ ($\rho_{45} = 0.5$);

c. imputation fractions of y_4 and y_5 equal to 0.293 with $\pi_{45} = 0.195$ ($\rho_{45} = 0.5$);

d. imputation fractions of y_4 and y_5 equal to 0.684 with $\pi_{45} = 0.507$ ($\rho_{45} = 0.5$).

18. Increasing amounts of lightly correlated measurement errors in two observed variables for different factors:

    a. imputation fractions of y_1 and y_4 equal to 0.051 with $\pi_{14} = 0.014$ ($\rho_{14} = 0.1$);

    b. imputation fractions of y_1 and y_4 equal to 0.106 with $\pi_{14} = 0.030$ ($\rho_{14} = 0.1$);

    c. imputation fractions of y_1 and y_4 equal to 0.293 with $\pi_{14} = 0.104$ ($\rho_{14} = 0.1$);

    d. imputation fractions of y_1 and y_4 equal to 0.684 with $\pi_{14} = 0.368$ ($\rho_{14} = 0.12$).

19. Increasing amounts of heavily correlated measurement errors in two observed variables for different factors:

    a. imputation fractions of y_1 and y_4 equal to 0.051 with $\pi_{14} = 0.051$ ($\rho_{14} = 0.41$);

    b. imputation fractions of y_1 and y_4 equal to 0.106 with $\pi_{14} = 0.106$ ($\rho_{14} = 0.41$);

    c. imputation fractions of y_1 and y_4 equal to 0.293 with $\pi_{14} = 0.293$ ($\rho_{14} = 0.41$);

    d. imputation fractions of y_1 and y_4 equal to 0.684 with $\pi_{14} = 0.684$ ($\rho_{14} = 0.41$).

20. Increasing amounts of measurement error in variable y_3:

    a. imputation fraction of y_3 equal to 0.051; all other imputation fractions equal to 0.051;

    b. imputation fraction of y_3 equal to 0.106; all other imputation fractions equal to 0.051;

    c. imputation fraction of y_3 equal to 0.293; all other imputation fractions equal to 0.051;

    d. imputation fraction of y_3 equal to 0.684; all other imputation fractions equal to 0.051.

For the models with correlated measurement errors, the implied correlation coefficient is mentioned in brackets. Ideally, we wanted to have models with $\rho = 0.1$ and $\rho = 0.5$, as in Section 3.3. In the cases where this was not possible due to the above-mentioned limitations of the hot deck method, we chose the feasible correlation nearest to 0.1 or 0.5.

The results of these simulations are shown in Table 10. Broadly speaking, these results confirm the previous findings with multinormal data. In particular:

- For models 16 and 17 (corresponding to the first type of misspecification from Section 3.3), the estimated validities of variables other than y_4 and y_5 agreed with their theoretical values in Table 8. The validities of y_4 and y_5 were overestimated.

- For models 18 and 19 (corresponding to the second type of misspecification from Section 3.3), the estimated validities agreed reasonably well with their theoretical values for all variables, including y_1 and y_4. For model 19 – but not for model 18 –, we observed a sharp increase in the average value of X^2 as more misspecified measurement errors were introduced.

- For model 20 (corresponding to the third type of misspecification from Section 3.3), the estimated validities again agreed with their theoretical values.

In addition, for most models at least some of the estimated parameters became erratic when large amounts of measurement errors were introduced.

Table 10: Simulation results with the hot deck method for misspecified models (using the same format as in Table 3).

model	R_c	R_v	X^2	λ_{y1}	λ_{y2}	λ_{y4}	λ_{y5}	λ_{x1}	λ_{x2}	λ_{x3}	λ_{x4}
16a	100	100	38	0.81	0.94	0.90	0.83	1.00	1.00	1.00	1.00
			(9)	(0.00)	(0.01)	(0.03)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)
16b	98	98	33	0.81	0.95	0.86	0.79	1.00	1.00	1.00	1.00
			(10)	(0.01)	(0.01)	(0.04)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
16c	85	85	28	0.81	0.95	0.71	0.65	1.00	1.00	1.00	1.00
			(11)	(0.01)	(0.01)	(0.07)	(0.07)	(0.00)	(0.00)	(0.00)	(0.00)
16d	93	89	25	0.81	0.95	0.59	0.53	1.00	1.00	1.00	1.00
			(9)	(0.01)	(0.01)	(0.21)	(0.19)	(0.00)	(0.00)	(0.00)	(0.00)
17a	99	99	41	0.81	0.94	0.93	0.86	1.00	1.00	1.00	1.00
			(9)	(0.00)	(0.01)	(0.02)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
17b	99	99	37	0.81	0.95	0.91	0.84	1.00	1.00	1.00	1.00
			(10)	(0.01)	(0.01)	(0.03)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)
17c	95	95	31	0.81	0.95	0.85	0.78	1.00	1.00	1.00	1.00
			(10)	(0.01)	(0.01)	(0.06)	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)
17d	87	74	30	0.81	0.94	1.17	0.67	1.00	1.00	1.00	1.00
			(25)	(0.02)	(0.10)	(2.57)	(0.21)	(0.00)	(0.00)	(0.00)	(0.00)
18a	100	100	43	0.77	0.94	0.89	0.87	1.00	1.00	1.00	1.00
			(11)	(0.03)	(0.02)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
18b	100	98	43	0.72	0.94	0.84	0.88	1.00	1.00	1.00	1.00
			(11)	(0.05)	(0.03)	(0.04)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
18c	100	88	35	0.56	0.95	0.65	0.88	1.00	1.00	1.00	1.00
			(13)	(0.08)	(0.07)	(0.06)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
18d	97	60	29	0.25	1.05	0.29	0.89	1.00	1.00	1.00	1.00
			(12)	(0.10)	(0.40)	(0.09)	(0.13)	(0.00)	(0.00)	(0.00)	(0.00)
19a	100	100	53	0.78	0.93	0.90	0.87	1.00	1.00	1.00	1.00
			(9)	(0.02)	(0.01)	(0.03)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
19b	100	100	67	0.74	0.92	0.85	0.87	1.00	1.00	1.00	1.00
			(16)	(0.04)	(0.03)	(0.04)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)
19c	99	92	96	0.57	0.92	0.67	0.87	1.00	1.00	1.00	1.00
			(30)	(0.08)	(0.07)	(0.06)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)
19d	94	53	116	0.24	1.00	0.30	0.86	1.00	1.00	1.00	1.00
			(32)	(0.08)	(0.28)	(0.08)	(0.13)	(0.00)	(0.00)	(0.00)	(0.00)
20a	100	64	60	0.77	0.88	0.90	0.83	0.95	0.95	0.95	0.96
			(19)	(0.05)	(0.04)	(0.03)	(0.03)	(0.06)	(0.06)	(0.09)	(0.09)
20b	100	66	59	0.77	0.89	0.89	0.83	0.94	0.96	0.96	0.95
			(19)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)	(0.11)	(0.09)
20c	100	57	61	0.77	0.89	0.90	0.82	0.94	0.96	0.95	0.96
			(20)	(0.03)	(0.05)	(0.04)	(0.03)	(0.06)	(0.06)	(0.09)	(0.09)
20d	99	68	63	0.78	0.88	0.89	0.84	0.95	0.95	0.95	0.96
			(19)	(0.05)	(0.05)	(0.04)	(0.03)	(0.04)	(0.05)	(0.08)	(0.08)

5 Discussion and Conclusion

As outlined in Section 2 and illustrated by Bakker (2012), estimates of the validity of administrative and survey variables can be obtained through an SEM. An important prerequisite for applying this method is that data measuring the same set of concepts are available from two different sources (surveys, registers), and that these sources can be linked at the record level. In addition, prior knowledge should be available about the expected direct effects between the concepts. Assuming that these prerequisites are met, the method can be applied in principle. In the present paper, we have investigated the suitability of the resulting validity estimates in various situations, by means of simulations and theoretical arguments. It should be noted that, while we used a specific example of an SEM in our simulation study, the theoretical arguments apply to all SEMs that satisfy the six numbered assumptions in Appendix A.2.

Firstly, we have seen that if the SEM is correctly specified (in the sense that no parameters in the model are fixed to values that are inconsistent with the underlying structure of the data), then the method provides consistent estimates of validity. The simulation results indicated that the minimal sample size that is required to obtain well-behaved (i.e. stable and approximately unbiased) validity estimates actually depends on the validities of the variables themselves. A moderate sample size of $N = 600$ should be sufficient if all variables contain small to moderate amounts of measurement error (say $\lambda \geq 0.7$). On the other hand, if the data contain variables with (very) low validities, then for $N = 600$ the behaviour of the estimated factor loadings and structural parameters is likely to be erratic and problems with convergence may be encountered during estimation. Thus, if one suspects that a variable in the model may have low validity, it is advisable to try to obtain a large sample. If this is impractical, one could try estimating the SEM both with and without that variable, and compare the results.

Secondly, we looked at cases where the measurement model was incorrectly specified. In particular, we considered cases where the SEM assumed uncorrelated measurement errors while in fact there were non-zero correlations, either between errors in variables that measure the same concept, or between errors in variables that measure different concepts. In principle, such a violation of a model assumption could invalidate all outcomes of the method. Nevertheless, in our study we found the effects of correlated measurement errors to be rather limited and 'local', in the sense that estimates of parameters not directly related to the offending variables remained consistent. A similar result was found for another type of misspecification, where a variable was assumed to be a perfect measure when in fact it contained measurement error. Again, the validities of the other variables were still estimated consistently. It should be noted that this 'local' property does *not* hold for all possible misspecifications. In particular, it is known that the bias due to an omitted path in a factor model or due to a misspecification in the structural part of an SEM can spread to parameter estimates throughout the model (Bollen et al., 2007). The latter types of misspecification are less likely to occur in the context considered here (cf. footnote 10).

Detecting and correcting the above forms of model misspecification can be problematic. We found that the usual chi-square test of overall model fit is not well-suited to finding the above types of misspecification. In fact, two of the three types investigated here were shown to be intrinsically impossible to detect using the X^2 statistic. The remaining type (correlated errors across concepts) can be detected by the chi-square test in principle, but the simulation results indicate that the power of this test is rather low, unless the errors are heavily correlated. For the two types of misspecification that are not detected by the X^2 statistic, an alternative means of detection could be provided by examining the plausibility of the estimated direct effects between the latent variables in the model. In fact, it was shown that these misspecifications do lead to biased values for some of these effects. But again, unless the errors are heavily correlated, the bias would often be too small in practice to make the direct effects implausible according to prior knowledge.

Thus, when the model fit is assessed only by the X^2 statistic and the plausibility of the structural effects, then model misspecifications of the above types are difficult or even impossible to recognise. Other overall fit measures are available, such as the (adjusted) goodness-of-fit index and the standardized root mean squared residual, as well as measures of fit for individual parameters, such as the modification index (Bollen, 1989). We did not test the effects of model misspecification on these other measures here. Previous empirical studies such as Saris et al. (2009) suggest that the standard fit measures are often equally (un)successful in detecting model misspecification.

A second problem can occur when trying to correct the model after detecting that it is misspecified. This requires that we turn a fixed parameter of the model into a free parameter. For the first and third type of misspecification considered here, the resulting correct model is not identified. In particular, correlated measurement errors in two indicators for the same concept cannot be distinguished from uncorrelated errors, unless a third indicator is added to the model. Alternatively, if a reasonable estimate of the offending parameter (i.e. either an error correlation or the validity of a single measure) is available from previous research, the parameter could be fixed to this value instead of 0, resp. 1.

From a theoretical point of view, extending the model to have three indicators per concept is a good idea, since this is also expected to alleviate the above-mentioned problem in the detection of misspecification. [On the other hand, Saris et al. (1987) found empirically that with three indicators measuring the same concept, the chi-square test still has low power for the detection of the first type of misspecification considered here.] From a practical point of view, adding a third indicator may be a feasible option if one has access to a survey with repeated measurements of the same concept. Otherwise, adding a third indicator is difficult, because it requires a third data source that can be linked to the two available data sets while retaining a sufficiently large sample.

In summary, we can say that the method discussed here provides approximately unbiased and stable estimates of the validity of administrative and survey data for a sample size as small as $N = 600$, provided that the SEM is correctly specified and contains variables with moderate to high validities. A partial misspecification of the model produces biased validity estimates for the involved variables, but (for the types of misspecification considered here) this bias is not propagated to the rest of the model. Technically, the method provides consistent estimates of indicator validity rather than true score validity (cf. Section 2.3). Extending the method to also estimate the true score validity and the reliability of administrative data is a topic for further research.

Finally, it should be noted that by standardising the observed variables as we have done here, information is lost about differences in the observed means of administrative and survey variables. This is in fact important information in the context of official statistics, where population totals are often the main parameters of interest. Observed means can be taken into account in an extended form of the SEM (Bollen, 1989). More research is needed on how to use this information for inference about validity.

References

Andrews, F. M. (1984), 'Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach', *Public Opinion Quarterly* **48**, pp. 409–442.

Bakker, B. F. M. (2012), 'Estimating the Validity of Administrative Variables', *Statistica Neerlandica* **66**, pp. 8–17.

Bakker, B. F. M. and Daas, P. (2012), 'Some Methodological Issues of Register Based Research', *Statistica Neerlandica* **66**, pp. 2–7.

Biemer, P. and Stokes, S. L. (1991), Approaches to the Modeling of Measurement Error, *in* Biemer, Groves, Lyberg, Mathiowetz and Sudman, eds, 'Measurement Errors in Surveys', John Wiley & Sons, New York, pp. 487–516.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, John Wiley & Sons, New York.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M. and Chen, F. (2007), 'Latent Variable Models under Misspecification: Two-Stage Least Squares (2SLS) and Maximum Likelihood (ML) Estimators', *Sociological Methods & Research* **36**, pp. 48–86.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*, John Wiley & Sons, New Jersey.

Di Zio, M., Guarnera, U. and Luzi, O. (2008), 'Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data'. Working Paper No. 22, UN/ECE Work Session on Statistical Data Editing, Vienna.

Groen, J. A. (2012), 'Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures', *Journal of Official Statistics* **28**, pp. 173–198.

Heise, D. R. and Bohrnstedt, G. W. (1970), Validity, Invalidity, and Reliability, *in* E. F. Borgatta and G. W. Bohrnstedt, eds, 'Sociological Methodology', Jossey Bass, San Francisco, pp. 104–129.

Jöreskog, K. and Sörbom, D. (1996), *LISREL 8: User's Reference Guide*, Scientific Software International, Chicago.

McCall, R. B. (2001), *Fundamental Statistics for Behavioral Sciences*, 8th edn, Wadsworth, Belmont.

Narayanan, A. (2012), ‘A Review of Eight Software Packages for Structural Equation Modeling’, *The American Statistician* **66**, pp. 129–138.

Novick, M. R. (1966), ‘The Axioms and Principal Results of Classical Test Theory’, *Journal of Mathematical Psychology* **3**, pp. 1–18.

Saris, W. E. and Andrews, F. M. (1991), Evaluation of Measurement Instruments Using a Structural Modeling Approach, *in* Biemer, Groves, Lyberg, Mathiowetz and Sudman, eds, ‘Measurement Errors in Surveys’, John Wiley & Sons, New York, pp. 575–597.

Saris, W. E. and Gallhofer, I. N. (2007), *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, John Wiley & Sons, New York.

Saris, W. E., Satorra, A. and Sörbom, D. (1987), The Detection and Correction of Specification Errors in Structural Equation Models, *in* ‘Sociological Methodology’, pp. 105–129.

Saris, W. E., Satorra, A. and Van der Veld, W. M. (2009), ‘Testing Structural Equation Models or Detection of Misspecifications?’, *Structural Equation Modeling* **16**, pp. 561–582.

Scherpenzeel, A. C. and Saris, W. E. (1997), ‘The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies’, *Sociological Methods & Research* **25**, pp. 341–383.

Appendix A Derivation of Results on Measurement Error Models

In this appendix, we derive some theoretical results on the measurement error models from Section 3. The first subsection summarises, without proof, some basic properties of the SEM that are needed later. The reader is referred to Bollen (1989) for more details and proofs. The second subsection discusses the four basic data generating models used in Section 3.

A.1 Some Properties of the SEM

A definition of the SEM was given in Section 2.1. Using matrix algebra, a more concise formulation of the same model can be given as follows. Let $\vec{\eta}$, $\vec{\xi}$, $\vec{\zeta}$, $\vec{y}$, $\vec{x}$, $\vec{\varepsilon}$, and $\vec{\delta}$ be column vectors containing the variables η_i, ξ_j, ζ_i, y_k, x_l, ε_k, and δ_l, respectively. Expressions (2), (3), and (4) are equivalent to:

$$\left.\begin{array}{rcl} \vec{\eta} & = & \mathrm{B}\vec{\eta}+\Gamma\vec{\xi}+\vec{\zeta}, \\ \vec{y} & = & \Lambda_y\vec{\eta}+\vec{\varepsilon}, \\ \vec{x} & = & \Lambda_x\vec{\xi}+\vec{\delta}, \end{array}\right\}$$

where B, Γ, Λ_y, and Λ_x are matrices containing the parameters $\beta_{ii'}$, γ_{ij}, λ_{yk}, and λ_{xl}, respectively.[16] In addition, the model contains four matrices of covariance parameters: $\Psi = (\psi_{ii'})$, $\Phi = (\varphi_{jj'})$, $\Theta_\varepsilon = (\theta_{\varepsilon kk'})$, and $\Theta_\delta = (\theta_{\delta ll'})$. Together, the choice of these eight matrices completely describes an SEM.

Let I denote the $m \times m$ identity matrix. The joint covariance matrix of the observed vectors $\vec{y}$ and $\vec{x}$ can be expressed in terms of the eight parameter matrices of the SEM, as follows:

$$\Sigma = \begin{pmatrix} \Lambda_y \mathrm{C} \Lambda_y' + \Theta_\varepsilon & \Lambda_y \mathrm{G} \Lambda_x' \\ \Lambda_x \mathrm{G}' \Lambda_y' & \Lambda_x \Phi \Lambda_x' + \Theta_\delta \end{pmatrix}, \tag{16}$$

with

$$\mathrm{C} = \mathrm{cov}(\vec{\eta},\vec{\eta}) = (\mathrm{I}-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)[(\mathrm{I}-\mathrm{B})^{-1}]', \tag{17}$$

and

$$\mathrm{G} = \mathrm{cov}(\vec{\eta},\vec{\xi}) = (\mathrm{I}-\mathrm{B})^{-1}\Gamma\Phi. \tag{18}$$

The free parameters of an SEM are estimated in practice by minimising the distance between the sample covariance matrix of $\vec{y}$ and $\vec{x}$ (say S) and expression (16). A standard choice of distance function to minimise is the maximum likelihood function:

$$F_{ML}(\mathrm{S},\Sigma) = \log|\Sigma| + \mathrm{tr}(\mathrm{S}\Sigma^{-1}) - \log|\mathrm{S}| - (p+q).$$

Minimising this function provides consistent estimates of the model parameters under light regularity conditions. Note that $F_{ML} = 0$ for $\Sigma = \mathrm{S}$.

A SEM is called identified if it is not possible to find two different sets of values for the free parameters in the model that produce the same Σ. An interesting fact that emerges from (16) is that the parameter matrices B, Γ, and Ψ only appear in Σ in an

[16] Recall that it is assumed here that each observed variable loads on one latent variable. Under this assumption, each row of Λ_y and Λ_x contains exactly one non-zero entry. We (ab)use this fact to identify the parameters in Λ_y and Λ_x by a single rather than a double index.

indirect way, in the form of the latent covariance matrices C and G. This means that a necessary condition for the model to be identified is that the system of equations (17) and (18) does not admit multiple solutions for the free parameters in B, Γ, and Ψ, given the other matrices in this system. Namely, if this system of equations had more than one solution, then each of these solutions would produce the same Σ and the model would not be identified.

For identified models, the quantity $X^2 = (N-1)F_{ML}$ is often used to test the fit of the model, with N the sample size. Implicitly, it is assumed that the observed variables follow a multivariate normal distribution. Then, under the null hypothesis that the model fits the data (more formally: that Σ equals the population covariance matrix of the model from which the sample that produced S was drawn), X^2 is asymptotically distributed as a chi-square variate with $(p+q)(p+q+1)/2-t$ degrees of freedom, t being the number of free parameters in the model. In particular, model identification requires that $t \leq (p+q)(p+q+1)/2$.

A.2 Some Results on Measurement Error Models

In the remainder of this appendix, we consider the set-up of the simulation study in Section 3: a multivariate normal data set is generated from one model and subsequently analysed using a second model that may or may not match the first model. Both models have the form of an SEM. In what follows, the notation from Section A.1 is reserved for the second model. The parameter matrices of the data generating model are denoted by B^*, Γ^*, Λ_y^*, Λ_x^*, Ψ^*, Φ^*, Θ_ε^*, and Θ_δ^*. Analogous to (16), the population covariance matrix of $\vec{y}$ and $\vec{x}$ under this model is given by

$$\Sigma^* = \begin{pmatrix} \Lambda_y^* \mathrm{C}^* (\Lambda_y^*)' + \Theta_\varepsilon^* & \Lambda_y^* \mathrm{G}^* (\Lambda_x^*)' \\ \Lambda_x^* (\mathrm{G}^*)' (\Lambda_y^*)' & \Lambda_x^* \Phi^* (\Lambda_x^*)' + \Theta_\delta^* \end{pmatrix}. \tag{19}$$

with C^* and G^* defined analogous to (17) and (18).

Below, we examine each data generating model while assuming that the population covariance matrix (19) is known. In this simplified situation, we can fit the second SEM by minimising $F_{ML}(\Sigma^*, \Sigma)$ directly. Of course, in practice one only has a sample-based estimate of Σ^* to work with. Typically, the resulting parameter values are consistent estimates of the parameters that would be obtained if the population covariance matrix were analysed. Thus, the theoretical results given below indicate the behaviour of the estimated parameters that is to be expected for large samples.

We will now discuss each of the four basic data generating models used in Section 3. The SEM used for data analysis is correctly specified in the first case and misspecified in the three remaining cases. Throughout this discussion, we shall make the following simplifying assumptions:

1. The model used for analysis is identified.

2. All variables are centered to have mean 0 and the observed variables in $\vec{y}$ and $\vec{x}$ are also standardised to have variance 1. (Thus Σ and Σ^* are correlation matrices rather than covariance matrices.)

For the cases with misspecification, some additional assumptions will be made below.

A.2.1 Correctly Specified Models

We start with the situation from Section 3.2, where the SEM used in the analysis is correctly specified with respect to the data generating model. That is, all *fixed* parameters have the same values in both models and all *free* parameters appear in exactly the same positions of the parameter matrices that describe these models. In this case, it seems reasonable to expect that the validity of each variable is correctly estimated.

Actually, it is easy to see from (16) and (19) that the system $\Sigma = \Sigma^*$ has an exact solution in this case: we simply set $\mathrm{B} = \mathrm{B}^*$, $\Gamma = \Gamma^*$, etc. Moreover, this solution is unique because the model is identified. Thus, if we would fit the second SEM to the population correlation matrix Σ^*, we would obtain a solution with $F_{ML} = 0$ and parameter values identical to those of the first SEM. In particular, the estimated validities λ_{yk} and λ_{xl} would be identical to their theoretical counterparts λ^*_{yk} and λ^*_{xl}.

A.2.2 Model Misspecification, Type 1

Next, we consider the three different situations from Section 3.3, where the SEM used in the analysis is misspecified with respect to the data generating model. In the remainder of this section, we make the following additional assumptions:

3. Each observed variable loads on one latent variable. ('The factor complexity of the observed variables is 1.')
4. A scale is provided for the latent variables in $\vec{\eta}$ and $\vec{\xi}$ by fixing $\mathrm{var}(\eta_i) = c_{ii} = c^*_{ii} = 1$ and $\mathrm{var}(\xi_j) = \varphi_{jj} = \varphi^*_{jj} = 1$.
5. In the model used for analysis, the matrices Θ_ε and Θ_δ are diagonal.
6. Each latent variable has at most two indicators.[17]

These assumptions are satisfied for all models that were used in the simulation study.

The first type of misspecification in Section 3.3 concerns correlated measurement errors in two observed variables that measure the same latent variable. By assumption 6, this latent variable is not measured by any other variable. Without essential loss of generality, we use the SEM from Section 2.2 as a working example. We take y_4 and y_5 as the variables with correlated errors and η_3 as the associated latent variable. In other words, the parameter matrices of the two SEMs are exactly matched with one exception: $\theta^*_{\varepsilon 45} \neq 0$ whereas $\theta_{\varepsilon 45}$ is fixed to 0.

[17] Actually, a less restrictive assumption suffices. The results given below remain valid if some latent variables have three or more indicators, provided that these latent variables are not directly related to any indicators involved in the misspecification.

The system $\Sigma = \Sigma^*$ consists of $(p+q)(p+q+1)/2$ non-redundant equations. Using (16), (19), and assumptions 1–5, these equations can be written as follows:

$$\left.\begin{array}{ll}
\lambda_{yk}^2 + \theta_{\varepsilon kk} = 1, & \text{(for all } k\text{)}, \\
\lambda_{xl}^2 + \theta_{\delta ll} = 1, & \text{(for all } l\text{)}, \\
\lambda_{y4}\lambda_{y5} = \lambda_{y4}^*\lambda_{y5}^* + \theta_{\varepsilon 45}^*, & \\
c_{i(k),i(k')}\lambda_{yk}\lambda_{yk'} = c_{i(k),i(k')}^*\lambda_{yk}^*\lambda_{yk'}^*, & \text{(for all } k \neq k' \text{ with } \{k,k'\} \neq \{1,2\}\text{)}, \\
g_{i(k),j(l)}\lambda_{yk}\lambda_{xl} = g_{i(k),j(l)}^*\lambda_{yk}^*\lambda_{xl}^*, & \text{(for all } k,l\text{)}, \\
\varphi_{j(l),j(l')}\lambda_{xl}\lambda_{xl'} = \varphi_{j(l),j(l')}^*\lambda_{xl}^*\lambda_{xl'}^*, & \text{(for all } l \neq l'\text{)}.
\end{array}\right\} \qquad (20)$$

In the third equation, we used that $c_{i(4),i(5)} = c_{33} = c_{i(4),i(5)}^* = c_{33}^* = 1$. An alternative way to write this equation is:

$$\lambda_{y4}\lambda_{y5} = \lambda_{y4}^*\lambda_{y5}^*\left(1 + \frac{\theta_{\varepsilon 45}^*}{\lambda_{y4}^*\lambda_{y5}^*}\right).$$

Note that the left-hand-sides of the equations in (20) contain the parameters that need to be estimated. The *values* of the expressions on the right-hand-sides are known at the estimation stage, even when the individual parameter values that occur in these expressions are not.

Suppose that condition (10) from Section 3.3 is satisfied. By inspection, it can be seen that the following choice of parameter values is an exact solution to (20):

$$\begin{aligned}
\lambda_{y4} &= \lambda_{y4}^*\sqrt{1 + \theta_{\varepsilon 45}^*/\lambda_{y4}^*\lambda_{y5}^*}, \\
\lambda_{y5} &= \lambda_{y5}^*\sqrt{1 + \theta_{\varepsilon 45}^*/\lambda_{y4}^*\lambda_{y5}^*}, \\
c_{3i'} &= c_{3i'}^*/\sqrt{1 + \theta_{\varepsilon 45}^*/\lambda_{y4}^*\lambda_{y5}^*}, \quad \text{(for all } i' \neq 3\text{)}, \\
g_{3j} &= g_{3j}^*/\sqrt{1 + \theta_{\varepsilon 45}^*/\lambda_{y4}^*\lambda_{y5}^*}, \quad \text{(for all } j\text{)},
\end{aligned}$$

and, for the remaining parameters,

$$\begin{aligned}
\lambda_{yk} &= \lambda_{yk}^*, && \text{(for all other } k\text{)}, \\
\lambda_{xl} &= \lambda_{xl}^*, && \text{(for all } l\text{)}, \\
c_{ii'} &= c_{ii'}^*, && \text{(for all other combinations } i,i'\text{)}, \\
g_{ij} &= g_{ij}^*, && \text{(for all other combinations } i,j\text{)}, \\
\varphi_{jj'} &= \varphi_{jj'}^*, && \text{(for all } i,j\text{)}.
\end{aligned}$$

Moreover, this solution is unique because of assumption 1.

The above result shows that, if the population correlation matrix Σ^* were analysed in this case, it would be possible to obtain a solution that fits exactly, as in the case without misspecification. In this solution, most parameters are correctly estimated. The only exceptions are the factor loadings of y_4 and y_5 on η_3, which are multiplied by $\sqrt{1 + \theta_{\varepsilon 45}^*/\lambda_{y4}^*\lambda_{y5}^*}$, and the non-zero correlations between η_3 and the other latent variables, which are divided by the same amount. In particular, the validities of all variables except y_4 and y_5 would be estimated correctly.

Note that the incorrectly estimated validities of y_4 and y_5 satisfy

$$\lambda_{y4} - \lambda_{y5} = (\lambda_{y4}^* - \lambda_{y5}^*)\sqrt{1 + \theta_{\varepsilon 45}^*/\lambda_{y4}^*\lambda_{y5}^*} = \kappa \times (\lambda_{y4}^* - \lambda_{y5}^*),$$

for some $\kappa > 0$. Hence, the difference $\lambda_{y4} - \lambda_{y5}$ has the same sign as $\lambda^*_{y4} - \lambda^*_{y5}$. In other words, although the actual values of the estimated validities of y_4 and y_5 are incorrect, these values do indicate correctly which of the measures has the highest validity. Moreover, the difference between the highest and lowest validity is inflated by the factor κ. For positively correlated errors, $\kappa > 1$ and the difference is too large. For negatively correlated errors, $\kappa < 1$ and the difference is too small.

Note also that, since we have again $\Sigma = \Sigma^*$ in this case, the null hypothesis of the chi-square test is satisfied. Thus, somewhat counterintuitively, although the model is misspecified, X^2 has the same distribution as if it were correctly specified. It follows that this type of misspecification cannot be detected using the chi-square test alone.

Finally, note that the above result implies that the data generating model (i.e. the original SEM with $\theta_{\varepsilon 45}$ as an additional free parameter), is not identified: there exist two distinct sets of parameter values that produce the same covariance matrix. In other words, for a latent variable with two indicators, it is not possible to distinguish uncorrelated measurement errors from correlated ones. This requires at least three indicators.

A.2.3 Model Misspecification, Type 2

The second type of misspecification in Section 3.3 concerns correlated measurement errors in two observed variables that measure different latent variables. Using assumption 6, we restrict the discussion to the case that both of the involved latent variables are also measured by one other variable. To clarify the issue, we take y_1 and y_4 as the variables with correlated errors and η_1 and η_3 as the associated latent variables. These latent variables are also measured by y_2 and y_5, respectively. This time, the parameter matrices of the two SEMs are exactly matched with the exception that $\theta^*_{\varepsilon 14} \neq 0$ while $\theta_{\varepsilon 14} = 0$.

Upon examining the system $\Sigma = \Sigma^*$ in this case, we find that it contains in particular the following equations:

$$c_{13}\lambda_{y1}\lambda_{y4} = c^*_{13}\lambda^*_{y1}\lambda^*_{y4} + \theta^*_{\varepsilon 14}, \tag{21}$$

$$c_{13}\lambda_{y2}\lambda_{y5} = c^*_{13}\lambda^*_{y2}\lambda^*_{y5}, \tag{22}$$

$$c_{13}\lambda_{y1}\lambda_{y5} = c^*_{13}\lambda^*_{y1}\lambda^*_{y5}, \tag{23}$$

$$c_{13}\lambda_{y2}\lambda_{y4} = c^*_{13}\lambda^*_{y2}\lambda^*_{y4}. \tag{24}$$

Let us assume that $c^*_{13} \neq 0$ (i.e. that the two involved latent variables are correlated). Multiplying (21) and (22) on both sides yields

$$c^2_{13}\lambda_{y1}\lambda_{y2}\lambda_{y4}\lambda_{y5} = (c^*_{13})^2\lambda^*_{y1}\lambda^*_{y2}\lambda^*_{y4}\lambda^*_{y5} + c^*_{13}\lambda^*_{y2}\lambda^*_{y5}\theta^*_{\varepsilon 14}, \tag{25}$$

while multiplying (23) and (24) yields

$$c^2_{13}\lambda_{y1}\lambda_{y2}\lambda_{y4}\lambda_{y5} = (c^*_{13})^2\lambda^*_{y1}\lambda^*_{y2}\lambda^*_{y4}\lambda^*_{y5}. \tag{26}$$

Since the second term on the right-hand-side of (25) is non-zero, it is clearly impossible to find values for $(c_{13}, \lambda_{y1}, \lambda_{y2}, \lambda_{y4}, \lambda_{y5})$ that satisfy (25) and (26) simultaneously.

Consequently, it is also impossible to satisfy (21)–(24) simultaneously, and by extension the system $\Sigma = \Sigma^*$ does not have an exact solution.

Thus, if the population correlation matrix were analysed with this type of model misspecification, a solution with $F_{ML}(\Sigma^*, \Sigma) > 0$ would be obtained. (We did not attempt to find an analytical expression for this solution.) The above result implies that the null hypothesis of the chi-square test is violated in this case ($\Sigma \neq \Sigma^*$). Hence, in principle, the chi-square test statistic can be used to detect this type of misspecification. In addition, the above result suggests – but does not prove – that the correct model with $\theta_{\varepsilon 14}$ as a free parameter is identified.[18]

A.2.4 Model Misspecification, Type 3

The third and final type of misspecification in Section 3.3 concerns errors in a single measure for a latent variable. Working in the context of the model from Section 2.2 as before, y_3 is the only observed variable that measures η_2. To ensure identification of the SEM that is used to analyse the data, the parameters $\lambda_{y3} = 1$ and $\theta_{\varepsilon 33} = 0$ are fixed. Misspecification occurs when the data generating model has $\lambda^*_{y3} \neq 1$ and $\theta^*_{\varepsilon 33} \neq 0$.

Analogous to the above discussion of the first type of misspecification, it can be shown that the following choice of parameter values yields an exact solution to the system $\Sigma = \Sigma^*$:

$$\begin{aligned}
\lambda_{y3} &= 1, & & \\
\lambda_{yk} &= \lambda^*_{yk}, & &\text{(for all other } k), \\
\lambda_{xl} &= \lambda^*_{xl}, & &\text{(for all } l), \\
c_{2i'} &= c^*_{2i'}\lambda^*_{y3}, & &\text{(for all } i' \neq 2), \\
c_{ii'} &= c^*_{ii'}, & &\text{(for all other combinations } i, i'), \\
g_{2j} &= g^*_{2j}\lambda^*_{y3}, & &\text{(for all } j), \\
g_{ij} &= g^*_{ij}, & &\text{(for all other combinations } i, j), \\
\varphi_{jj'} &= \varphi^*_{jj'}, & &\text{(for all } i, j).
\end{aligned}$$

As before, assumption 1 guarantees that this solution is unique. Note that all validities are correctly estimated in this case. (Obviously, $\lambda_{y3} \neq \lambda^*_{y3}$. Recall however that we do not use the method to make inference about the validity of y_3.) In fact, incorrect values occur only for the latent covariances that involve η_2: these are attenuated towards 0.

Similarly to the first case with correlated errors, the null hypothesis $\Sigma = \Sigma^*$ remains satisfied in spite of the presence of model misspecification. Again, we conclude that the chi-square test is not capable of detecting that a model is misspecified in this way.

[18] Technically, the model could still be underidentified due to the existence of two distinct sets of parameter values that produce the same covariance matrix, with $\theta_{\varepsilon 14}$ taking non-zero values in both sets. We did not attempt to prove or disprove this.

Appendix B Derivation of Results on Hot Deck Imputation

B.1 Appendix to Section 4.2

Suppose that the hot deck imputation method is only applied to a variable z_1, with π_1 the fraction of imputed records in the data set. We denote the original distribution of z_1 (before imputation) by $F(z_1)$. Let $\tilde{z}_1$ denote the imputed version of z_1. Then, formally, we have:

$$\tilde{z}_1 = (1-\tau_1)z_1 + \tau_1\breve{z}_1,$$

with τ_1 a dichotomous random variable taking the value 1 with probability π_1 and 0 with probability $1-\pi_1$, and with $\breve{z}_1$ a variable drawn at random from $F(z_1)$. Note that τ_1 and $\breve{z}_1$ are drawn independently of each other.

Technically, $\tilde{z}_1$ is a mixture of two random variables, z_1 and $\breve{z}_1$, having the *same* distribution $F(z_1)$. It follows that $\tilde{z}_1$ is also distributed according to $F(z_1)$, so that in particular $E(\tilde{z}_1) = E(z_1)$ and $\text{var}(\tilde{z}_1) = \text{var}(z_1)$. Thus, in expectation, the mean and variance of z_1 are the same before and after imputation.[19]

We now consider the covariance between $\tilde{z}_1$ and an unimputed variable z_2. An application of the conditional covariance formula yields:

$$\begin{aligned}\text{cov}(\tilde{z}_1,z_2) &= E_{\tau_1}\{\text{cov}(\tilde{z}_1,z_2|\tau_1)\} + \text{cov}_{\tau_1}\{E(\tilde{z}_1|\tau_1),E(z_2|\tau_1)\}\\ &= (1-\pi_1)\text{cov}(z_1,z_2) + \pi_1\text{cov}(\breve{z}_1,z_2) + 0\\ &= (1-\pi_1)\text{cov}(z_1,z_2).\end{aligned}$$

In the last line, we used that $\text{cov}(\breve{z}_1,z_2) = 0$ because $\breve{z}_1$ is drawn from the univariate distribution of z_1, independently of z_2. Combining this with the above result on the variance of $\tilde{z}_1$ yields $\rho(\tilde{z}_1,z_2) = (1-\pi_1)\rho(z_1,z_2)$.

Finally, suppose that, independently of the construction of $\tilde{z}_1$, we construct an imputed version of z_2 using the same technique:

$$\tilde{z}_2 = (1-\tau_2)z_2 + \tau_2\breve{z}_2,$$

where $\tau_2 = 1$ with probability π_2 and $\breve{z}_2$ is drawn from $F(z_2)$. Obviously, $E(\tilde{z}_2) = E(z_2)$ and $\text{var}(\tilde{z}_2) = \text{var}(z_2)$ as before. For the covariance between $\tilde{z}_1$ and $\tilde{z}_2$, we find:

$$\begin{aligned}\text{cov}(\tilde{z}_1,\tilde{z}_2) &= E_{\tau_1,\tau_2}\{\text{cov}(\tilde{z}_1,\tilde{z}_2|\tau_1,\tau_2)\} + \text{cov}_{\tau_1,\tau_2}\{E(\tilde{z}_1|\tau_1,\tau_2),E(\tilde{z}_2|\tau_1,\tau_2)\}\\ &= P(\tau_1=0,\tau_2=0)\text{cov}(z_1,z_2) + 0\\ &= (1-\pi_1)(1-\pi_2)\text{cov}(z_1,z_2).\end{aligned}$$

In the second line, all other terms vanish due to independence. In particular, it holds that $\text{cov}(\breve{z}_1,\breve{z}_2) = 0$. We conclude that $\rho(\tilde{z}_1,\tilde{z}_2) = (1-\pi_1)(1-\pi_2)\rho(z_1,z_2)$.

[19] These properties can also be proved directly by conditioning on τ_1 and applying the standard formulas for conditional expectation and variance.

B.2 Appendix to Section 4.3

The alternative version of the hot deck imputation method described in Section 4.3 can be put into more mathematical terms as follows. As in the main text, we take z_1 and z_2 as the variables to impute. Let (τ_1, τ_2) be two dichotomous random variables having a joint distribution such that

$$\begin{aligned} P(\tau_1 = 1) &= \pi_1, \\ P(\tau_2 = 1) &= \pi_2, \\ P(\tau_1 = 1, \tau_2 = 1) &= \pi_{12}. \end{aligned}$$

Note that the joint distribution of (τ_1, τ_2) is completely determined by the choice of π_1, π_2, and π_{12}. In addition, let $(\breve{z}_1, \breve{z}_2)$ be a random draw from the joint distribution $F(z_1, z_2)$, independently of (τ_1, τ_2). Now we define the imputed versions of z_1 and z_2 by

$$\begin{aligned} \tilde{z}_1 &= (1-\tau_1)z_1 + \tau_1\breve{z}_1, \\ \tilde{z}_2 &= (1-\tau_2)z_2 + \tau_2\breve{z}_2. \end{aligned}$$

Note that, as in Section B.1, $\tilde{z}_1$ is a mixture of two random variables having the same univariate distribution $F(z_1)$. Therefore it still holds that $E(\tilde{z}_1) = E(z_1)$ and $\text{var}(\tilde{z}_1) = \text{var}(z_1)$. Similarly, $E(\tilde{z}_2) = E(z_2)$ and $\text{var}(\tilde{z}_2) = \text{var}(z_2)$.

For the covariance between the imputed versions of z_1 and z_2, we find:

$$\begin{aligned} \text{cov}(\tilde{z}_1, \tilde{z}_2) &= E_{\tau_1,\tau_2}\{\text{cov}(\tilde{z}_1, \tilde{z}_2 | \tau_1, \tau_2)\} + \text{cov}_{\tau_1,\tau_2}\{E(\tilde{z}_1|\tau_1,\tau_2), E(\tilde{z}_2|\tau_1,\tau_2)\} \\ &= \pi_{12}\text{cov}(\breve{z}_1, \breve{z}_2) + (1-\pi_1-\pi_2+\pi_{12})\text{cov}(z_1, z_2) + 0 \\ &= (1-\pi_1-\pi_2+2\pi_{12})\text{cov}(z_1, z_2). \end{aligned}$$

In the second line, we used that

$$P(\tau_1 = 0, \tau_2 = 0) = 1 - P(\tau_1 = 1) - P(\tau_2 = 1) + P(\tau_1 = 1, \tau_2 = 1);$$

note that this probability is well-defined because of assumption (14). Furthermore, we used in the third line that $\text{cov}(\breve{z}_1, \breve{z}_2) = \text{cov}(z_1, z_2)$.

Thus, we find that in this case $\rho(\tilde{z}_1, \tilde{z}_2) = (1-\pi_1-\pi_2+2\pi_{12})\rho(z_1, z_2)$. If a third variable, say z_3, is imputed by the hot deck method independently of z_1 and z_2, then it can be shown as before that

$$\begin{aligned} \rho(\tilde{z}_1, \tilde{z}_3) &= (1-\pi_1)(1-\pi_3)\rho(z_1, z_3), \\ \rho(\tilde{z}_2, \tilde{z}_3) &= (1-\pi_2)(1-\pi_3)\rho(z_2, z_3), \end{aligned}$$

with π_3 the fraction of imputed records for z_3.