Discussion Paper

Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables

2015 | 17

Sander Scholtus
Bart F. M. Bakker
Arnout van Delden

Summary: Administrative data are used more and more in official statistics and academic research. Since administrative variables are usually collected for a different purpose, it is important to assess the suitability of these variables for statistical use. Due to measurement errors, an administrative variable might have low validity for the concept of interest and it might also suffer from bias. In this paper, we describe a method for assessing the measurement validity of administrative and survey variables by means of a structural equation model. An advantage of this approach is that it is not necessary to assume that either the administrative or the survey data are error-free. In addition, we consider the possibility of estimating the bias in these variables, in the presence of a so-called audit sample for which 'gold standard' data have been obtained. An application of the method is discussed to assess the measurement quality of value-added tax Turnover for the Dutch short-term statistics.

1 Introduction

In recent years, the use of administrative data has grown in official statistics as well as in academic research (Bethlehem, 2008; Bakker and Daas, 2012). Governmental organisations, such as tax authorities, social security offices, and municipalities, routinely collect data on a large number of social and economic phenomena as part of their regular activities. In many countries, national statistical institutes (NSIs) and other producers of official statistics have access to these administrative data sources. Many NSIs are looking at ways to use administrative data to reduce – and ideally replace – their own data collection by means of questionnaires. Reasons for this include tighter budgets and a decreasing willingness of persons and businesses to participate in surveys. Administrative data also offer possibilities for more detailed statistical analyses than surveys based on relatively small samples.

When examining the suitability of a given administrative source for statistical purposes, several questions need to be addressed (Bakker, 2011b; Zhang, 2012). In this paper, we will focus on issues related to the quality of measurement. In general, all data sources may contain errors. In the case of administrative data, a particular source of error arises from potential differences between the variable that is measured for administrative purposes and the variable that is needed for statistical purposes. To give an example, Statistics Netherlands is supposed to publish short-term statistics on Turnover, which is a variable with a specific definition according to the short-term statistics regulation (European Commission, 2006). The Dutch tax authorities also collect information on Turnover from businesses to levy value-added tax (VAT). Conceptually, these two Turnover variables are not the same for all businesses; for instance, some economic activities are included in the first type of Turnover but excluded from the second because they are exempt from taxes. (We will return to this example in the application below.) It is, therefore, important to assess the measurement quality of administrative variables for statistical use (Bakker and Daas, 2012; Groen, 2012).

In the context of questionnaire design, there is a well-established tradition of using linear structural equation models (SEMs) to assess the measurement quality of survey variables; key references include Andrews (1984), Saris and Andrews (1991), Scherpenzeel and Saris (1997), Saris and Gallhofer (2007), and Alwin (2007). The models used in this approach can be seen as an extension of the classical test theory from psychology as set out by Lord and Novick (1968) and Jöreskog (1971). Each observed variable is modelled as an imperfect measure of an underlying latent (unobserved) variable. To quantify the measurement quality of an observed variable, one can estimate its *validity* which, under the simplest model, is defined as its standardised factor loading on the underlying latent variable (see Section 2). These models are usually identified by taking repeated measurements on each target variable, which requires a carefully-planned research design. It should also be noted that SEMs require variables that are measured on an interval scale or higher. At the very least, the latent variable should have such a measurement level. For nominal and ordinal variables, latent class models are more appropriate (Biemer, 2011).

Applying the same modelling approach to administrative data is not straightforward. As administrative data are collected by an external party, it is usually not possible to conduct methodological experiments. Bakker (2012) suggested that repeated measurements may be obtained by linking an administrative data set to data from an independent sample survey. This is useful in particular for examining whether questions in an existing survey can be replaced by corresponding administrative variables, at least in terms of validity. Similarly, Pavlopoulos and

Vermunt (2015) and Oberski (2015) have used latent class models to compare the amount of classification error in categorical administrative and survey variables. An important advantage of approaches that use latent variables is that they do not assume that either the administrative or the survey data are error-free. In fact, it is not necessary to know in advance which source provides the measurement with the highest validity: this is estimated from the data. If the fitted model indicates that the administrative data have a validity that is at least as high as that of the survey data, that will be an argument for using the cheaper administrative data.

While validity captures the correlation of an observed variable to the underlying concept, producers of official statistics are often interested in population means or totals. Therefore, in addition to the validity, it may be important to know whether any substantial measurement bias occurs in the levels of individual variables (so-called *intercept bias*). The main objective of the present paper is to discuss how the approach using SEMs may be extended to also assess the bias of an administrative variable. To illustrate, we describe an application at Statistics Netherlands to assess the validity and intercept bias of VAT Turnover for the Dutch short-term statistics.

The remainder of this paper is organised as follows. Section 2 describes the proposed methodology for estimating the validity and intercept bias of observed variables. Some more technical points are confined to an appendix. The above-mentioned application to VAT Turnover is discussed in Section 3. Section 4 closes the paper with a discussion of the possibilities and limitations of the proposed method. While the main focus of this paper is on evaluating the measurement quality of administrative data, some potential other applications of the method are also outlined in Section 4.

2 Methodology

2.1 Assessing validity and intercept bias using SEMs

Let $y_1, \dots, y_p$ denote a set of observed variables that may be affected by measurement errors, and let $\eta_1, \dots, \eta_m$ denote the underlying variables of interest that are error-free and not observed directly. The relationship between each observed and unobserved variable, as well as the relations that exist among the unobserved variables, may be described by an SEM.

For our purposes here, an SEM may be defined as a system of linear regression equations:

$$\eta_i = \alpha_i + \sum_{j \neq i} \beta_{ij} \eta_j + \zeta_i, \quad (i = 1, \dots, m), \tag{1}$$

$$y_k = \tau_k + \lambda_k \eta_{i_k} + \epsilon_k, \quad (k = 1, \dots, p). \tag{2}$$

Equations of the form (1) are *structural equations* relating the unobserved variables to each other: the coefficient β_{ij} represents a direct effect of η_j on η_i, ζ_i represents a zero-mean disturbance term, and α_i represents a structural intercept. Equations of the form (2) are *measurement equations* relating an observed y_k to an unobserved η_{i_k} in terms of a factor loading λ_k, a measurement intercept τ_k, and a zero-mean measurement error ϵ_k that is uncorrelated with η_{i_k}. Note that the same latent variable can be measured by different observed variables. By contrast, we restrict attention in this paper to SEMs in which each observed variable loads on exactly one latent variable. More general SEMs that do not have this restriction are discussed, e.g., by Bollen (1989).

The SEM given by (1)–(2) contains the following parameters: α_i, β_{ij}, τ_k, λ_k, $\psi_{ij} = \mathrm{cov}(\zeta_i, \zeta_j)$, and $\theta_{kl} = \mathrm{cov}(\epsilon_k, \epsilon_l)$. It is standard practice to restrict some of these to zero a priori, based on substantive considerations. Provided that the model is identified, the unknown parameters can be estimated from the observed variance-covariance matrix and the observed vector of means of $y_1, \dots, y_p$; see Section 2.2. The absolute value of the standardised factor loading

$$|\lambda_k^s| \equiv |\lambda_k| \frac{\mathrm{sd}(\eta_{i_k})}{\mathrm{sd}(y_k)} = \sqrt{1 - \frac{\mathrm{var}(\epsilon_k)}{\mathrm{var}(y_k)}} \tag{3}$$

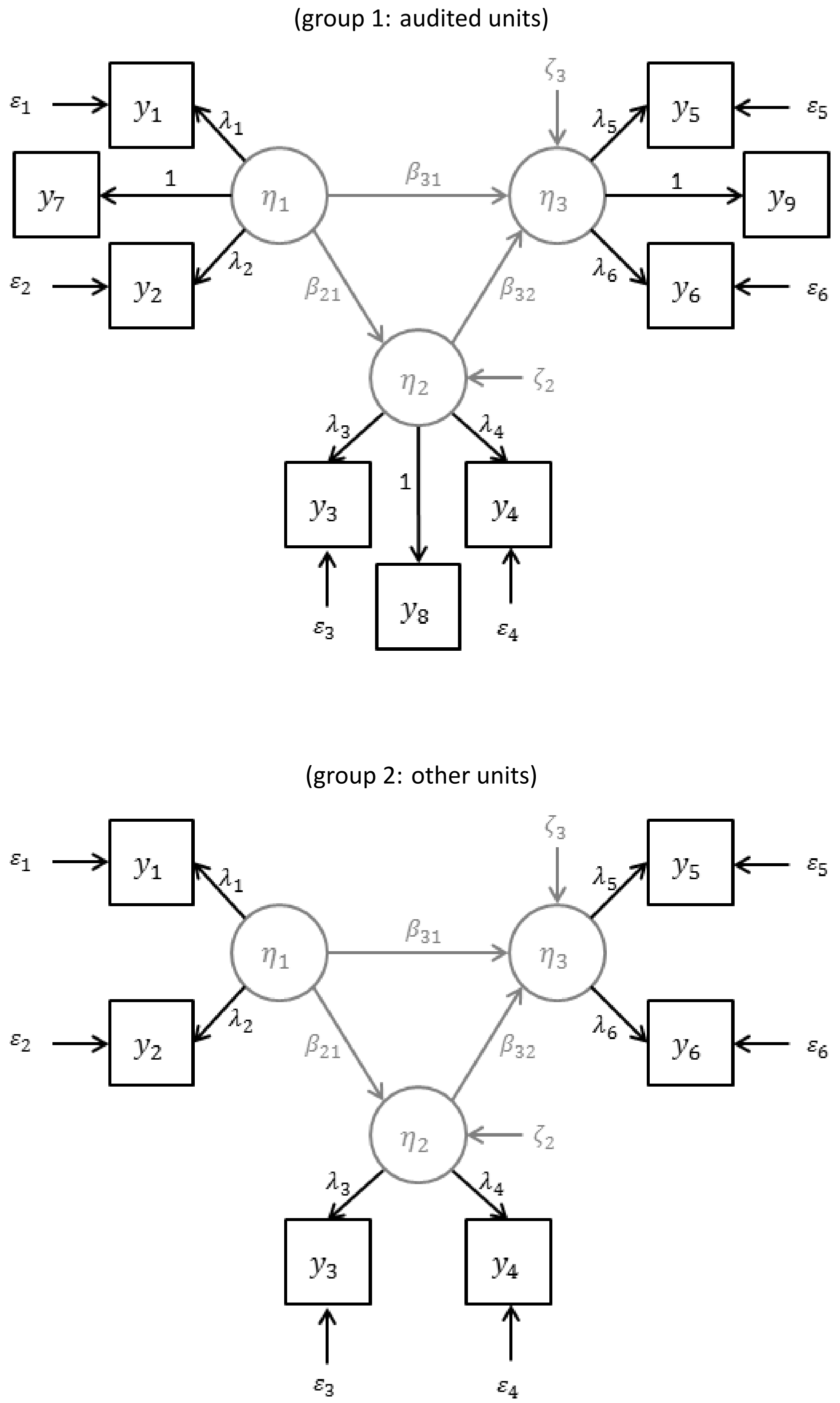
may be used as a measure of the validity[1] of y_k (Bakker, 2012). The intercept bias of y_k may be evaluated by comparing its observed mean to the estimated error-free mean of η_{i_k}. Having estimated the model, we can derive formulae to correct each observed variable to the scale of the corresponding error-free variable; see Section 2.4 for more details.

By linking administrative data to survey data, one will usually obtain at most two indicators per latent variable (Scholtus and Bakker, 2013). The smallest SEM that is then identified has $m = 2$ correlated latent variables with two indicators each. If covariates are available that are considered to be measured (essentially) without error, these can also be included in the model to obtain identification. In addition, identification of any SEM with latent variables requires that each latent variable be given a scale and, if the model contains intercept terms, that the origins of these scales be fixed as well. When one is interested only in the validity, identification may be achieved by standardising each latent variable to have mean 0 and variance 1. However, this is not an option if the intercept bias is to be evaluated. In fact, none of the standard SEM identification procedures [see Little et al. (2006) for an overview] is then suitable because, as argued by Bielby (1986), these procedures define an 'arbitrary' metric for the latent variables.

A procedure for achieving model identification in a 'non-arbitrary' way was suggested by Sobel and Arminger (1986) and discussed in the present context by Scholtus (2014). The basic idea is to collect additional 'gold standard' data on each latent variable for a random subsample of the original data set. Many of the variables encountered in official statistics are factual (e.g., Age, Voting behaviour, Turnover, Number of employees), so that it is theoretically possible to obtain the true score for each unit. In practice, it is usually prohibitively expensive or otherwise inconvenient to do so for the entire population or even for a sizeable sample. But it may often be feasible to obtain 'gold standard' data for a small subsample of units. Provided that this *audit sample* is obtained by randomised selection from the original data set, we can use it to assign a 'non-arbitrary' metric to the latent variables, thereby identifying the SEM. We can still use the entire data set to estimate the model parameters in terms of this metric.

Figure 2.1 shows an example of a path diagram of an SEM that is identified in this way, having $m = 3$ latent variables with two indicators each (outside the audit sample). The task of estimating this model can be cast as a missing-data problem that may be solved by fitting a two-group SEM; see Section 2.3. Results on simulated data in Scholtus (2014) suggested that a relatively small audit sample of 50 units may often be sufficient.

[1] In the terminology of Saris and Andrews (1991), $|\lambda_k^s|$ measures the *indicator validity* of y_k. This is actually the product of its 'pure' validity and its reliability as defined by Saris and Andrews (1991). Biemer (2011) uses the terms *empirical* and *theoretical validity* instead of indicator validity and 'pure' validity, respectively. This point is taken up in Section 4.


(group 1: audited units)

(group 2: other units)

Figure 2.1 Example of a two-group SEM identified by means of an audit sample. The model for the first group contains additional error-free variables that are observed only in the audit sample. The structural part of the model is shown in grey.

In practice, the ‘gold standard’ data could be obtained by some form of re-editing by subject-matter experts, as was done in a different context by Nordbotten (1955). In Figure 2.1 and throughout this paper, it is assumed that the audit data do not contain any measurement errors: in (2) for these variables, $\tau = 0$, $\lambda = 1$, and $\mathrm{var}(\epsilon) = 0$. In fact, the model can be identified by the audit sample also when $\mathrm{var}(\epsilon) \neq 0$ but the other two assumptions do hold. In that case, the ‘gold standard’ data are supposed to contain only measurement errors that do not affect the scale of measurement. While this assumption is theoretically weaker than the assumption of no errors, it is not necessarily more plausible in practice. When the ‘gold standard’ data are obtained by re-editing, it actually seems less plausible from a practical point of view.

2.2 Estimating an SEM

Let $\mu = (\mu_1, \dots, \mu_p)'$ and $\Sigma = (\sigma_{kl})$ denote, respectively, the population mean vector and variance-covariance matrix of the observed variables in the SEM. That is, $\mu_k = E(y_k)$ and $\sigma_{kl} = \mathrm{cov}(y_k, y_l)$. Under the model given by (1)–(2), these moments are expressible in terms of the unknown model parameters: $\mu = \mu(\vartheta)$ and $\Sigma = \Sigma(\vartheta)$, with ϑ a vector containing all distinct parameters. Explicit expressions for $\mu(\vartheta)$ and $\Sigma(\vartheta)$ can be found, e.g., in Bollen (1989).

For a given sample of size n, let $\bar{y}$ and S denote the empirical means and covariances of $y_1, \dots, y_p$. A conventional way to estimate ϑ is by minimising a certain distance function F_{ML} between $(\bar{y}, S)$ and $(\mu(\vartheta), \Sigma(\vartheta))$, which leads to maximum likelihood (ML) estimation if the sample consists of independent, identically distributed (i.i.d.) observations from a multivariate normal distribution. This method of estimation also produces asymptotic standard errors for the estimated parameters, as well as a test statistic that can be used as a measure of overall fit: under the above assumption of normality, $X^2_{\mathrm{ML}} = (n-1)F_{\mathrm{ML}}$ should follow a chi-square distribution with known degrees of freedom if the model holds. More details are given in Appendix I.1 and in Bollen (1989).

In many practical applications, including the VAT application to be discussed in Section 3, the assumption of having i.i.d. observations from a normal distribution is not satisfied. Firstly, the data may come from a different (unknown) distribution. In this situation, it can be shown that minimising F_{ML} still produces a consistent point estimator for ϑ under mild conditions, but the estimated standard errors are typically incorrect and the above test statistic need not follow a chi-square distribution. It is known how to obtain asymptotically correct standard errors (Satorra, 1992; Muthén and Satorra, 1995); see also Appendix I.1. A correction to the chi-square test statistic was proposed by Satorra and Bentler (1986, 1994). The resulting corrected statistic is denoted by X^2_{SB} here. The terms Robust Maximum Likelihood and Pseudo Maximum Likelihood (PML) are used to refer to this estimation strategy when the data are not normally distributed.

Secondly, the above assumption is violated when the sample is obtained through some complex survey design, possibly involving without-replacement sampling from a finite population, stratification, clustering and/or multi-stage selection. In this case, to obtain a consistent point estimator, one should use design-consistent estimates of μ and Σ in place of $\bar{y}$ and S. After this adjustment, essentially the same results apply as in the i.i.d. case with non-normal data (Muthén and Satorra, 1995). Thus, the same PML approach may be used to obtain corrected standard errors and test statistics. For single-group models, this approach has been shown to give good results in simulation studies (Stapleton, 2006). Some more details are given in Appendix I.1. In the application to be discussed below, we used PML estimation to account for both phenomena: non-normal data and finite-population sampling.

Many software packages are available for estimating SEMs, including LISREL, EQS, and Mplus. For the analyses in this paper, we made use of two packages from the R environment for statistical computing (R Development Core Team, 2015): the package `lavaan` (Rosseel, 2012) which contains the basic functionality for estimating a variety of latent variable models and the package `lavaan.survey` (Oberski, 2014) which implements the PML approach for SEM estimation with complex samples. The latter package in turn relies on the R package `survey` (Lumley, 2004) which provides general functionality for complex survey analysis.

2.3 Incorporating the audit sample

In theory, the estimation of an SEM that is identified by means of an audit sample is straightforward. Consider the example in Figure 2.1. We set up a two-group SEM, where the first group contains the n_1 observations from the audit sample and its model is defined in terms of the observed variables $y_1, \dots, y_9$, while the second group contains the $n - n_1$ remaining observations and its model is defined in terms of $y_1, \dots, y_6$ alone. In the first group, the model is identified by assuming that $y_7 = \eta_1$, $y_8 = \eta_2$, and $y_9 = \eta_3$. The model for the second group is identified by restricting all parameters of the overlapping part of the model to be equal in both groups; this makes sense if the audit sample is a random subsample of the original data.[2)]

In practice, some complications arise. The main practical problem is that all standard software packages that can estimate multi-group SEMs require that the same set of observed variables be used in each group. Thus, in the example of Figure 2.1 we need to account for the missing data for y_7, y_8, y_9 in the second group.

Allison (1987) proposed a general-purpose method for estimating SEMs with missing data, which provides ML estimates provided that the data are i.i.d. multivariate normal and the missing values are Missing At Random (MAR) in the terminology of Little and Rubin (2002). In the context considered here, the data are missing by design (i.e., the design of the audit sample) and we can ensure that the MAR condition is satisfied. Baumgartner and Steenkamp (1998) described an extension of Allison's method that is usable in combination with the PML approach of Section 2.2, so that the condition of multivariate normality can be dropped. This method involves imputing random values from a normal distribution for the missing variables in the second group, in such a way that the observed means of these variables are identically zero, the observed variances are identically one, and the observed covariances with all other variables are zero. In the model for the second group, the measurement equations $y_7 = \epsilon_7$, $y_8 = \epsilon_8$, and $y_9 = \epsilon_9$ are included. In addition, we fix $\theta_{77} = \theta_{88} = \theta_{99} = 1$ for this group. Basically, this ensures that the observed means and covariances for y_7, y_8, y_9 in the second group are exactly reproduced while the estimation of the rest of the model is not affected by the imputed values. Because some of the observed moments are now fixed by design, some care must be taken in defining the correct degrees of freedom for the model. See Appendix I.2 for more details.

A more generally applicable way to deal with missing values in an SEM is by multiple imputation (Oberski, 2014). We will not explore this option here. In addition to the missing-data problem,

[2)] To identify the model, it is in fact sufficient to restrict only the parameters of the *measurement* model to be invariant across groups; the *structural* parameters could be estimated freely. In practice, the model fit will improve if these parameters are left free across groups. However, this may be seen as overfitting if the audit sample is truly a random subsample of the original data, as all differences between the groups should then be due to sampling fluctuations. Having a separate structural model in each group would also complicate the derivation of a correction formula for the observed variables (see Section 2.4). We therefore restrict all overlapping parameters to be invariant.

some more subtle issues arise if we want to take a complex survey design into account in an SEM with more than one group. This is discussed in Appendices I.1 and I.4.

2.4 Deriving a correction formula

Having estimated the SEM given by (1)–(2), we obtain for each observed variable y_k an estimate of the validity $|\hat{\lambda}_k^s|$ from (3) and an estimated regression line for y_k given η_{i_k}:

$$\hat{y}_k = \hat{\tau}_k + \hat{\lambda}_k \eta_{i_k}. \tag{4}$$

In broad terms, three possible cases may arise here:

(a) the validity of y_k is high and $(\hat{\tau}_k, \hat{\lambda}_k) \approx (0, 1)$;
(b) the validity of y_k is high but $(\hat{\tau}_k, \hat{\lambda}_k)$ differs significantly from $(0, 1)$;
(c) the validity of y_k is low.

With case (a), the observed values are strongly correlated to the true values and there is no indication of measurement bias. Observed variables that fall under case (c) apparently contain too much measurement error to be of use. In the remainder of this section, we will focus on case (b), where there is a strong correlation between the observed and true values but the observed values are systematically too high or too low. Suppose we would like to correct this measurement bias. Formula (4), which predicts the value of y_k for a given value of η_{i_k}, cannot be used directly for this purpose. Rather, we need a formula that predicts η_{i_k}, given the observed values.

From the literature, it is known how to predict the true scores of the latent variables in an SEM from the observed ones; see, e.g., formula (6) in Meijer et al. (2012). This predictor is unbiased but it involves a linear combination of (in general) all the observed variables from the original model. In the present context, these variables have most likely been obtained specifically for a methodological evaluation study, e.g., by linking data from different sources, and they will typically not all be available during regular statistical production. To correct y_k for measurement error in this situation, we will now derive a formula for predicting $\hat{\eta}_{i_k}$ from y_k alone.

For notational simplicity, we drop the indices k and i_k in the remainder of this section. To find the best predictor of η for an arbitrary given value $y = y_0$, we need to evaluate $E(\eta \mid y = y_0)$. Consider the linear regression model $\eta = a + by + \omega$, with $E(y\omega) = E(\omega) = 0$. Using (2) and the fact that $\sigma_{\epsilon\eta} = 0$, we find that

$$b = \frac{\sigma_{y\eta}}{\sigma_y^2} = \frac{\lambda\sigma_\eta^2 + \sigma_{\epsilon\eta}}{\sigma_y^2} = \lambda\frac{\sigma_\eta^2}{\sigma_y^2}$$

and $a = \mu_\eta - b\mu_y$. Thus, we obtain:

$$\begin{aligned} E(\eta \mid y = y_0) &= a + by_0 \\ &= \mu_\eta + \lambda\frac{\sigma_\eta^2}{\sigma_y^2}(y_0 - \mu_y). \end{aligned} \tag{5}$$

The unknown parameters in expression (5) can all be expressed as simple functions of ϑ (Bollen, 1989). Thus, having estimated the original SEM, we can use the following formula for predicting η given $y = y_0$:

$$\hat{\eta}_0 = \hat{\mu}_\eta + \hat{\lambda}\frac{\hat{\sigma}_\eta^2}{\hat{\sigma}_y^2}(y_0 - \hat{\mu}_y), \tag{6}$$

with $\hat{\mu}_\eta = \mu_\eta(\hat{\vartheta})$, $\hat{\sigma}_\eta^2 = \sigma_\eta^2(\hat{\vartheta})$, etc. Furthermore, since a and b are differentiable functions of ϑ, approximate standard errors for $\hat{a} = \hat{\mu}_\eta - \hat{\lambda}(\hat{\sigma}_\eta^2/\hat{\sigma}_y^2)\hat{\mu}_y$ and $\hat{b} = \hat{\lambda}(\hat{\sigma}_\eta^2/\hat{\sigma}_y^2)$ can be obtained by linearisation; `lavaan` and most other modern SEM software packages provide this option.

Some remarks are in order. Firstly, it should be noted that solving (4) for η directly yields $\tilde{\eta}_0 = (y_0 - \hat{\tau})/\hat{\lambda}$, which is *not* equivalent to (6). This approach is invalid in general because it ignores the fact that ϵ and y are correlated. However, $\hat{\eta}_0$ does converge to $\tilde{\eta}_0$ as the validity of y approaches 1 (see below).

Secondly, a similar formula to (6) can be derived for predicting η from any given subset of the observed variables in the original SEM, by considering the multiple regression of η on those variables. Such a formula may be useful in practice if several (but not necessarily all) observed variables from the SEM are available during regular production, for instance because they come from the same data source. Similarly, if covariates without error are available they can also be incorporated in the prediction of η to improve the correction formula.

Finally, in the context of a repeated survey (where the same set of statistics is produced on a regular basis), it is desirable to use the same instance of formula (6) to correct observations on y for measurement error in multiple survey rounds, without the need to re-estimate the correction every time. Clearly, this requires that the measurement model remains stable over time. In fact, the parameters a and b also depend on μ_η and σ_η^2 and could therefore change as the structural part of the model evolves over time, even when the measurement model remains stable. However, it can be shown that this effect is negligible in practice provided that the validity of y is high enough. Note that, using (2), we can write

$$b = \lambda \frac{\sigma_\eta^2}{\lambda^2\sigma_\eta^2 + \sigma_\epsilon^2} = \frac{1}{\lambda}\frac{1}{1 + (\sigma_\epsilon/\lambda\sigma_\eta)^2}.$$

By (3), $(\sigma_\epsilon/\lambda\sigma_\eta)^2 = (\lambda^s)^{-2} - 1$ is a non-negative quantity that is close to 0 when the validity of y is high. Hence, we can use a first-order Taylor expansion to obtain, as a good approximation:

$$b \approx \frac{1}{\lambda}\left[1 - \left(\frac{\sigma_\epsilon}{\lambda\sigma_\eta}\right)^2\right].$$

Similarly, it can be derived that, to the same order of approximation,

$$a \approx -\frac{\tau}{\lambda} + \left(\frac{\sigma_\epsilon}{\lambda\sigma_\eta}\right)^2\left(\mu_\eta + \frac{\tau}{\lambda}\right).$$

Now let $\sigma_\eta^2(T)$ and $\sigma_\eta^2(T+1)$ denote the variance of η at two time points (not too far apart) and suppose that the parameters of the measurement model remain invariant between T and $T+1$. Then, using the above first-order approximation, we find the following expression for the change in slope parameter b between T and $T+1$:

$$b(T+1) - b(T) \approx \frac{1}{\lambda}\left(\frac{\sigma_\epsilon}{\lambda\sigma_\eta(T+1)}\right)^2 \Delta\sigma_\eta^2(T, T+1),$$

with $\Delta\sigma_\eta^2(T,T+1) = \left[\sigma_\eta^2(T+1) - \sigma_\eta^2(T)\right]/\sigma_\eta^2(T)$ the relative change in σ_η^2 between T and $T+1$. For instance, suppose that $\lambda = 1.1$, the validity of y at time $T+1$ equals 0.95 and $\Delta\sigma_\eta^2(T,T+1) = 5\%$. According to the above expression, the absolute change in b will be about

$$\frac{1}{1.1} \times \left(\frac{1}{(0.95)^2} - 1\right) \times 0.05 \approx 0.005,$$

which is likely to be negligible compared to the standard error of $\hat{b}$. More generally, the change in b will be close to 0 provided that the validity of y is high at both time points and the relative change in σ_η^2 is small. A similar result can be obtained for a with the additional requirement that the relative change in μ_η between T and $T+1$ has to be small.

In summary, provided that the measurement model remains stable over time and the validity of y is high enough, the effect of changes in the structural parameters on a and b will be negligible in practice when the structural parameters evolve gradually over time. Of course, the measurement model cannot be expected to remain stable indefinitely. Therefore, it will be necessary to update formule (6) by conducting a new audit sample at regular intervals and/or whenever changes are made to the data collection process that may affect the measurement parameters of y. In the case of administrative data, an NSI should monitor actively whether such changes are being made by the administrative authority.

3 Application: Using VAT Turnover for the Dutch quarterly short-term statistics

3.1 Introduction

From the second half of 2011 onwards, Statistics Netherlands has been publishing quarterly short-term statistics (STS) on Turnover that are based on a combination of VAT data for small to medium-sized businesses and a census survey for the largest and/or most complex units (Van Delden and De Wolf, 2013). The VAT data are obtained from tax declarations submitted to the Dutch tax authorities. The primary output of STS consists of estimated growth rates of Turnover for different sectors of the Dutch economy (classified by NACE code). Levels of total Turnover by sector are also estimated and used to calibrate the Dutch structural business statistics (SBS) and to weight the contribution of each sector to the Dutch national accounts. Given this secondary use of the STS estimates, it is vital that they do not suffer from intercept bias. The relation between VAT Turnover and STS Turnover is known to vary by type of economic activity, depending on the specific tax regulations that apply (Van Delden et al., 2015). Hence, direct use of the VAT data may give a distorted view of the contribution of each sector to the Dutch economy.

Van Delden et al. (2015) previously analysed the measurement quality of VAT data by a direct linear regression of Turnover as measured in the SBS survey[3)] on VAT Turnover. This analysis was done separately for each NACE group[4)]. The results of these analyses were used, in combination with qualitative knowledge on tax regulations, to decide for each NACE group whether:

(a) VAT data could be used as a direct replacement of survey data;
(b) VAT data could be used after applying a linear correction to VAT Turnover; or
(c) VAT data could not be used.

[3)] The definitions of SBS and STS Turnover are identical in nearly all cases.

[4)] We use the term *NACE group* to refer to a stratification of units by NACE code. In the study by Van Delden et al. (2015), NACE groups were defined at the most detailed level used during data collection and processing at Statistics Netherlands. In our application, we defined NACE groups at the level where separate STS estimates are published.

The correction formulae for NACE groups in class (b) followed directly from the linear regression model [similar to formula (6) in this paper, but with SBS Turnover taking the role of true Turnover]. Class (c) also included NACE groups for which the analysis was inconclusive, e.g. because the results did not agree with expectations based on the tax regulations. In fact, a drawback of linear regression is that measurement errors in the SBS and VAT data are not explicitly taken into account. It is well known that estimates of regression parameters may be biased in the presence of measurement errors (Bound et al., 2001). Van Delden et al. (2015) did use a robust regression to avoid bias due to incidental (large) errors, but this does not provide protection against the effects of structural measurement errors. Therefore, we decided to do an alternative analysis using an SEM to account for potential measurement errors. The results of this analysis will be compared to those obtained by the method of Van Delden et al. (2015) below.

As mentioned above, the analyses in this application were done in R. Interested readers can request a copy of the R code by sending an email to `sshs@cbs.nl`.

3.2 Data

The present application focused on eight NACE groups, listed in Table 3.1. The first four groups are part of the sector "Trade", while the last four groups are part of the sector "Transportation and storage". For all of these NACE groups, the VAT data are currently not used to produce STS. In addition to Turnover, we included the following concepts in the SEM: Number of employees, Costs of purchases, and Total operating costs. All data referred to the year 2012.

Table 3.1 Overview of NACE groups considered in this application

NACE code	description
45112	Sale and repair of passenger cars and light motor vehicles
45190	Sale and repair of trucks, trailers, and caravans
45200	Specialised repair of motor vehicles
45400	Sale and repair of motorcycles and related parts
50100	Sea and coastal passenger water transport and ferry-services
50300	Inland passenger water transport and ferry-services
52100	Warehousing and storage
52290	Forwarding agencies, ship brokers and charterers; weighing and measuring

For all concepts, one indicator is available from the SBS sample survey data. As a second indicator for the Number of employees, we used the value listed in the General Business Register (GBR) which is the population frame of business units maintained by Statistics Netherlands. Additional indicators for the other three variables were obtained from the Profit Declaration Register (PDR). This is an administrative data set provided to Statistics Netherlands by the tax authorities. Businesses are obliged to provide information to the PDR annually, but delayed reporting is accepted by the tax authorities up to several years after the reference period. In this study, we used the PDR data that were available by October 2014. Finally, VAT data on Total turnover were included. Businesses usually declare VAT on a monthly or quarterly basis. In this study, we used the derived annual VAT Turnover. Table 3.2 summarises the available data sources for each concept.

Table 3.3 lists the population size in each NACE group as well as the number of units for which data were available. Businesses from the group of very large and/or complex units were excluded from this analysis, as Statistics Netherlands is not planning to use administrative data to replace the STS survey for this group at the moment. The SBS data set has survey weights to account for

Table 3.2 Overview of variables used in this application, with available data sources

variable name	available sources
Number of employees	GBR, SBS (+ audit)
Costs of purchases	PDR, SBS (+ audit)
Total operating costs	PDR, SBS (+ audit)
Total turnover	VAT, PDR, SBS (+ audit)

Table 3.3 Number of units in each NACE group. All figures refer to 2012 and, apart from the first line, to the population with large and/or complex units excluded.

NACE group	45112	45190	45200	45400
population (total)	18 680	1 790	6 054	1 763
population (w/o complex units)	18 556	1 739	6 018	1 759
SBS net sample	934	180	281	76
SBS net sample linked to admin. data	819	170	238	60
net audit sample	44	47	44	43
NACE group	50100	50300	52100	52290
population (total)	943	4 500	764	2 968
population (w/o complex units)	889	4 475	695	2 835
SBS net sample	197	383	194	450
SBS net sample linked to admin. data	98	323	145	375
net audit sample	33	44	38	40

the sampling design and non-response. The SBS uses simple random sampling stratified by NACE group and size class, based on the number of employees according to the GBR. Correction for non-response is based on a weighting model that involves NACE group, size class and legal form.

We could not link all units from the SBS data set to the two administrative data sets used here (PDR and VAT), mainly because not all fiscal units from the administrative data could be linked unambiguously to an SBS unit. In addition, some units had missing data in the PDR or VAT data sets (unit non-response). This explains the loss of units between the third and fourth line in Table 3.3. To account for potential selectivity introduced in this step, we recalibrated the survey weights within each NACE group, using a simplified version of the standard SBS weighting model.

Since SBS Turnover was available for all units in the third row of Table 3.3, we could check whether the loss of records that were not linked to administrative data yielded selection bias that was not corrected by reweighting, at least for our target variable Turnover. Let n' and n denote the number of records in the SBS sample before and after linkage, respectively. First, we computed the difference between the reweighted estimate of the population total of SBS Turnover after linkage and the original weighted estimate before linkage; this is reported as a percentage of the estimate before linkage in the row "difference" of Table 3.4. Next, we simulated, for each NACE group, $10,000$ simple random samples without replacement of size n from the SBS sample of size n', applied our reweighting model to each subsample, and computed the resulting estimate of total SBS Turnover. This produced an approximate reference distribution of outcomes for this estimate under the assumption that the linkage loss was completely at random and thus not selective. The row "p value" in Table 3.4 reports, for each NACE group, the fraction of simulated values that differed from the original SBS estimate by at least as much (in absolute value) as our original reweighted estimate after linkage. It is seen that, for all NACE groups except 45112, the difference was not significant; thus, there was no indication of selection bias with respect to SBS Turnover. For NACE group 45112, the significant effect could be explained by two units with unusually large values of Turnover (relative to their size classes) that were not linked to administrative data. When these two outliers were removed from the population, the difference for NACE group 45112 was not significant anymore (p value: 0.082).[5)]

Table 3.4 Effect of unlinked records on population estimate of SBS Turnover

NACE group	45112	45190	45200	45400	50100	50300	52100	52290
units linked	88%	94%	85%	79%	50%	84%	75%	83%
difference	−5.8%	−2.8%	2.4%	−8.5%	−2.7%	1.8%	−4.9%	1.7%
p value	0.003	0.150	0.423	0.221	0.883	0.646	0.392	0.656

The necessary ‘gold standard’ data for evaluating intercept bias were obtained by having two senior subject-matter experts re-edit the SBS data for an audit sample of 50 units in each NACE group. The audit sample was stratified by a coarsened version of size class, reduced to just two strata, with 25 units taken in each stratum. The net audit sample was slightly smaller (see Table 3.3), mainly because we had selected the audit sample before linking the SBS data to administrative data. In addition, a few audited units turned out to be inactive or misclassified by type of economic activity, which means that they were not part of the target population.

All variables considered here have skew distributions. For instance, most of the Turnover in each NACE group is concentrated among a few largest units. In theory, the PML estimation method should account for the fact that the data are not normally distributed. We also considered possible transformations to obtain data that were closer to being normal, or to account for heteroskedastic measurement errors. In some cases, this led to a slightly improved model fit (not shown here). On the other hand, these transformations made the interpretation of the measurement model in terms of the original variables less intuitive. We therefore decided to work with the untransformed data in this application, since we could find a model that fitted these data reasonably well (see below). In what follows, all financial variables are measured in millions of Euros.

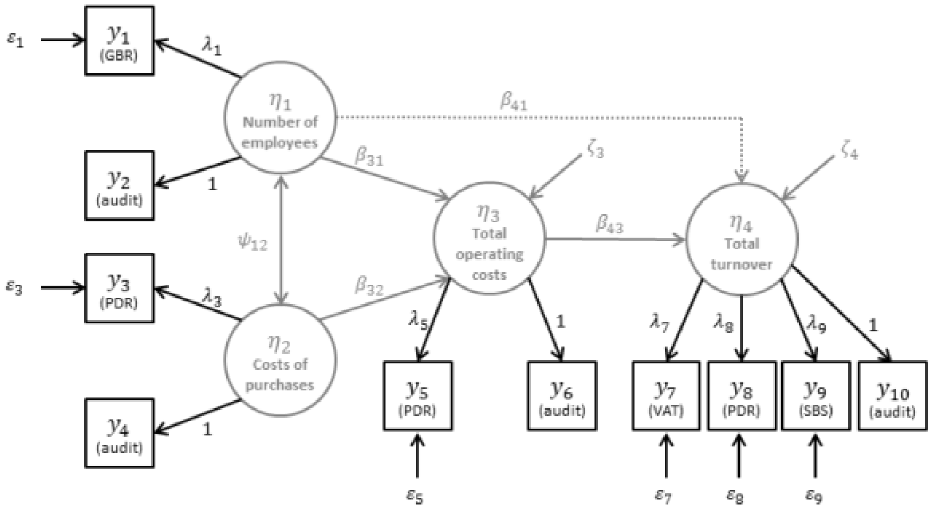


Figure 3.1 Path diagram of the basic model used in this application (intercepts not shown). For the group of non-audited units, remove variables y_2, y_4, y_6, and y_{10}.

Preliminary analyses revealed that some correlations between the original SBS data and the audited data were extremely high. This could be explained by the fact that relatively few values were changed during the audit, combined with the skewness of the data. These correlations

[5)] In the remainder of this paper, we report results for NACE group 45112 with these two outliers removed prior to reweighting. The results with the two outliers included were virtually identical.

close to 1 led to some computational problems, with covariance matrices close to being singular, so that lavaan could not estimate the parameters of the SEM. To avoid these problems, we decided to only include SBS Turnover in the model and exclude the other SBS variables.

The path diagram of the basic SEM used here is shown in Figure 3.1. For the structural part of the model, we used a nearly-saturated recursive model. The direction of the arrows was prompted by accounting rules that underlie these conceptual variables: Costs of purchases (η_2) is a component of Total operating costs (η_3), which in turn contributes to the Total turnover (η_4); in addition, Number of employees (η_1) is closely related to Staff costs which is another component of η_3. The structural model is not fully saturated: we excluded the direct effect of η_2 on η_4 because there is no substantive reason why Costs of purchases would have an additional effect on Total turnover besides its contribution to Total operating costs. The direct effect of Number of employees on Total turnover was included in the initial model but turned out to be insignificant for some NACE groups, in which cases it was set to zero.

3.3 Results

Table 3.5 shows selected fit measures for the final chosen model in each NACE group: the robust chi-square test statistic X^{2*}_{SB} and robust versions of the CFI, TLI, and RMSEA; see Appendices I.2 and I.3 for precise definitions of the fit measures in this table. For NACE groups 45112, 45200, and 50300, all measures indicate an excellent fit. For NACE groups 45190, 52100, and 52290, the robust chi-square statistic is somewhat high compared to the degrees of freedom and the other fit measures mostly indicate a reasonable fit. For the remaining two groups (45400 and 50100), the overall fit is rather poor. Note that the sample sizes in these last two groups are small, both compared to the other NACE groups and compared to the minimal sample sizes that are recommended in the SEM literature [see, e.g., Boomsma (1982)].

Table 3.5 Fit measures for the final model

NACE group	45112	45190	45200	45400	50100	50300	52100	52290
X^{2*}_{SB}	57.5	91.0	41.0	172.9	98.7	61.5	82.8	94.5
df^*	62	61	61	62	62	61	62	62
p value	0.637	0.008	0.977	0.000	0.002	0.458	0.040	0.005
$\hat{c}^*_{\text{SB}}$	36.7	11.7	43.9	2.4	6.1	22.7	11.9	16.2
CFI^*_{SB}	1.000	0.966	1.000	0.925	0.887	0.998	0.898	0.974
TLI^*_{SB}	1.001	0.966	1.034	0.927	0.890	0.998	0.902	0.975
$RMSEA^*_{\text{SB}}$	0.000	0.077	0.000	0.255	0.116	0.007	0.070	0.053

We checked the residuals of the fitted models. In cases where the overall model fit was poor, some large residuals did occur for the exogenous variables Number of employees and/or Costs of purchases, but never for Turnover. Thus, to the extent that the model may be misspecified, we assumed that these misspecifications were related only to other variables than Turnover. Moreover, results on simulated data in Scholtus and Bakker (2013) suggest that, for the type of SEM considered here, the effects of local model misspecifications are not propagated to other parts of the model. Hence, for the purpose of making valid inferences about the measurement quality of Turnover, we considered the fitted models to be adequate.

Table 3.6 displays the estimated factor loadings, measurement intercepts, and validities of Turnover as observed in VAT, PDR, and SBS. Recall that the validity λ^s is given by (3). It is seen that the validity of VAT Turnover was high in nearly all NACE groups, the one exception being

Table 3.6 Parameter estimates for Turnover (with standard errors)

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
λ (VAT)	0.79	0.01	0.89	0.02	1.29	0.19	0.80	0.04
τ (VAT)	−0.04	0.04	−0.00	0.05	−0.04	0.08	0.01	0.03
λ^s (VAT)	0.98		0.97		0.99		0.97	
λ (PDR)	1.02	0.01	0.95	0.02	1.23	0.20	0.99	0.03
τ (PDR)	0.00	0.05	0.06	0.03	−0.02	0.08	0.00	0.01
λ^s (PDR)	1.00		0.98		0.99		1.00	
λ (SBS)	1.01	0.01	1.01	0.00	1.21	0.19	0.98	0.02
τ (SBS)	−0.01	0.05	−0.00	0.00	−0.04	0.08	0.00	0.01
λ^s (SBS)	0.99		1.00		0.98		1.00	

	50100		50300		52100		52290	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
λ (VAT)	0.92	0.05	0.75	0.03	1.24	0.15	0.69	0.08
τ (VAT)	0.15	0.16	0.18	0.03	0.03	0.55	1.14	0.44
λ^s (VAT)	0.96		0.95		0.93		0.72	
λ (PDR)	1.12	0.06	0.89	0.02	1.09	0.08	1.01	0.06
τ (PDR)	−0.69	0.21	0.07	0.03	0.32	0.41	0.24	0.41
λ^s (PDR)	0.97		1.00		0.94		0.99	
λ (SBS)	1.00	0.00	0.83	0.05	0.93	0.12	0.73	0.09
τ (SBS)	0.04	0.04	0.11	0.04	0.21	0.42	0.84	0.47
λ^s (SBS)	1.00		0.94		0.94		0.83	

NACE group 52290 ($\lambda^s = 0.72$). On the other hand, the unstandardised measurement parameters indicate that the observed VAT Turnover was systematically too high or too low compared to the true Turnover. In most groups, the hypothesis that $\lambda = 1$ was rejected. For the intercept τ, significant deviations from 0 were found only in NACE groups 50300 and 52290.

It is interesting to note that, overall, the measurement quality of PDR Turnover was better than that of VAT Turnover: the validity was high in all NACE groups (often very close to 1) and the intercept and unstandardised factor loading were usually closer to the reference values of 0 and 1. Thus, from the point of view of measurement quality, it would often be preferable to use the PDR data as a source for Turnover. Unfortunately, these data cannot be used directly for STS, because they are available only on an annual basis and because they suffer from administrative delay as mentioned above.

For all NACE groups, a correction formula for VAT Turnover was derived as described in Section 2.4. The results are shown in Table 3.7. Thus, for instance, to correct VAT Turnover to the scale of true Turnover in NACE group 45112, the following formula was obtained:

$$\widehat{\text{Turnover}} = 0.11 + 1.22 \times \text{VAT Turnover}.$$

Analogous correction formulae could be derived, if necessary, for the other observed Turnover variables (SBS and PDR).

For completeness, the full set of parameter estimates is given for all NACE groups in Appendix II. Regarding the other variables in the model, an interesting result was that the variable Number of employees as measured in the GBR had the worst measurement quality. For all NACE groups considered in this study, the GBR underestimated the true Number of employees according to the model (λ significantly lower than 1).

Regarding the structural parameters of the model, it is interesting to note that the standardised

Table 3.7 Intercept and slope of a correction formula for VAT Turnover (with standard errors)

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
a (VAT)	0.11	0.05	0.07	0.06	0.04	0.06	0.01	0.03
b (VAT)	1.22	0.02	1.07	0.04	0.76	0.12	1.18	0.06
	50100		50300		52100		52290	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
a (VAT)	0.01	0.19	-0.16	0.07	0.35	0.41	0.94	0.51
b (VAT)	1.01	0.07	1.21	0.10	0.70	0.08	0.75	0.07

versions of the direct effects β_{31} and β_{32} (not shown here) suggest that, for most NACE groups in the Trade sector, the Total operating costs are dominated by Costs of purchases rather than the Number of employees. Official figures from the Dutch SBS confirm that, in these NACE groups, the Costs of purchases are usually much higher than the Staff costs. On the other hand, for most NACE groups in the Transportation sector, the difference between β_{31} and β_{32} is smaller after standardisation and this is again confirmed by the official figures. Thus, in this respect the estimated parameters are plausible.

Finally, it is interesting to compare the results of the SEM method to those that would be obtained by the robust regression method of Van Delden et al. (2015) mentioned in Section 3.1. Since the results in Van Delden et al. (2015) are not directly comparable to ours (they are based on a different stratification into NACE groups and different data), we applied their method to our data to obtain comparable results; see Table 3.8. Recall that this method involves a robust linear regression of SBS Turnover on VAT Turnover. The fitted regression line is then used to correct VAT Turnover if necessary. Hence, the estimates in Table 3.8 should be compared to Table 3.7.

For NACE groups 45190, 45400, and 50100, the two methods yielded similar correction formulae for VAT Turnover. In the other NACE groups, some large and significant differences occurred, in particular for the slope parameter. In all cases, the direction of the difference is as expected from the estimated measurement parameters of SBS Turnover under the SEM model (Table 3.6). For instance, in NACE group 50300, the robust regression suggested that VAT Turnover and true Turnover are well-aligned ($\hat{b}_{RR} = 1.00$, $\hat{a}_{RR} = 0.00$), whereas the SEM yielded a correction formula for VAT Turnover with slope 1.21 and intercept -0.16. The difference can be explained by noting that, according to Table 3.6, both SBS Turnover and VAT Turnover suffer from systematic under-reporting in this NACE group; hence, the assumption of the robust regression method that SBS Turnover is a good proxy for true Turnover is not satisfied here.

Table 3.8 Intercept and slope of a correction formula for VAT Turnover estimated by robust regression (with standard errors)

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
a_{RR} (VAT)	0.01	0.00	-0.00	0.00	0.00	0.00	0.00	0.00
b_{RR} (VAT)	1.27	0.01	1.06	0.01	0.97	0.01	1.17	0.04
R^2	0.98		0.97		0.99		0.96	
	50100		50300		52100		52290	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
a_{RR} (VAT)	0.00	0.00	0.00	0.00	0.01	0.01	0.04	0.01
b_{RR} (VAT)	0.99	0.01	1.00	0.00	0.83	0.02	0.96	0.01
R^2	0.96		0.98		0.90		0.71	

4 Conclusions and discussion

4.1 Discussion of results

In this paper, we explored the possibility of using structural equation modelling to assess the measurement quality of administrative variables for statistical use. We specifically looked at validity and intercept bias. Estimating the intercept bias of an observed variable in a meaningful way requires the collection of additional 'gold standard' data for a random subsample of the original data. To illustrate the method, we applied it to assess the suitability of VAT data on Turnover for the Dutch quarterly STS.

As the method is relatively expensive and complex, it might be useful in practice to apply a staged approach. Begin by making some preliminary comparisons between the administrative data and data from other sources (e.g., survey data), for instance by visual inspection of scatter plots or by robust linear regression, as was done by Van Delden et al. (2015). This preliminary analysis may already be conclusive in two possible ways: either by revealing that the administrative data are only weakly correlated to the other data (in which case the data are probably not useful), or by revealing that the two data sources contain nearly identical values (in which case the data may be considered to have high validity, provided that the errors in the two sources are independent). In all other cases, it seems dangerous to draw conclusions about the validity of the administrative data at this stage. Moreover, nothing can be concluded at this stage about the presence of systematic bias in either data source; the results for NACE group 50300 discussed at the end of Section 3 may serve as a cautionary example.

For the second stage, if the preliminary analysis is inconclusive, one may proceed with the estimation of an SEM to evaluate the validity. For this, the collection of additional audit data is not required. If the validity turns out to be low, the administrative data should probably not be used.

If the validity is high and one is also interested in the bias, then one may proceed to the final stage. For this stage, an audit sample is conducted. An extended SEM can then be used to evaluate the intercept bias and, if necessary, estimate a formula for correcting the bias. Results on simulated data in Scholtus (2014) suggest that the precision of the estimated SEM parameters increases slowly with the size of the audit sample, so from a cost-benefit point of view it may be reasonable to keep the audit sample small in practice. On the other hand, the audit samples in these simulations were selected by simple random sampling and it may in fact be possible to obtain significant improvements in precision by optimising the design of the audit sample. This could be an interesting topic for future research. Within the method as discussed here, any form of probability sampling can be used to select the audit sample so long as the design is known.

Identification of an SEM typically requires that each variable be measured at least two times, and if only two indicators are available their measurement errors should be uncorrelated. In traditional applications of SEMs to survey data, identification is often achieved by asking multiple variants of the same question, either within the same interview or in a follow-up interview (Saris and Gallhofer, 2007). For panel surveys, an alternative is the so-called simplex design which involves asking the same question to the same respondents at (≥ 3) different time points (Alwin, 2007). With administrative data sources, asking follow-up questions is almost never possible. In addition, while many longitudinal administrative data sources are available, the recorded values often remain unchanged until an event occurs that triggers an alteration (Bakker, 2011a). This implies that measurement errors in a single administrative source at different time points are often correlated. A more generally applicable way to obtain multiple measurements with administrative data may be to link them to survey data, as we did in this paper. This approach does require that the data sources can be perfectly linked (no linkage errors). In practice, there may be records that cannot be linked. In that case, one should check whether the linked data are sufficiently representative of the population, and possibly weigh the data to improve this.

In applications where multiple data sources are available during regular statistical production, an SEM may be used to estimate the underlying true scores from the combined information in all observed variables (Meijer et al., 2012). In this paper, we considered the situation that combined data sources are available only in the context of a methodological study. We showed how an SEM can be used to estimate a correction formula for the intercept bias in a single observed variable, which can subsequently be applied during regular production. This type of formula is relevant for variables with high validity but significant intercept bias, in applications where the absolute levels of individual variables are of concern.

4.2 Assumptions and limitations

A strong assumption of the method is that it is possible to obtain ‘gold standard’ versions of the target variables, at least for a small subsample of units. In practice, applications where absolute levels are of concern are likely to arise only for ‘factual’ variables. For such variables, an objective true value can be determined in principle, although the measurement procedure that is required to obtain this value may be difficult, expensive, or otherwise inconvenient to implement in practice. Clearly, the outcome of the method relies on the quality of the audit data. In our application, the audit data were obtained through re-editing by subject-matter experts. An important, albeit difficult, question is whether it is realistic to consider these data as a ‘gold standard’. As a topic for future research, it may be interesting to investigate the re-editing process in more detail and to find out how confident the experts are about their decisions. It is conceivable that the quality of the audit data actually differs by sub-population, e.g., because less information is available on smaller units or in specific NACE groups. It may also be interesting to conduct a simulation study to investigate to what extent the estimate of validity and the correction formula for intercept bias are robust to minor violations of the assumption that the audit data do not contain measurement error.

A limitation of the application in Section 3 is that the model was fitted to data of only one year, so we could not test whether the estimated measurement parameters change over time. It would be good to repeat the analysis on data from a different year. Note that this would also require a new audit sample. As noted in Section 2.4, gradual longitudinal changes in the structural part of the model should have a negligible effect on the correction formula in practice, provided that the validity of the observed variable is high enough.

A precondition for applying a correction formula in practice is that the estimated parameters should not be affected too much by individual observations with unusual values (outliers or points of high leverage). Standard estimation techniques for SEMs are based on ordinary sample covariances and means and therefore sensitive to outliers. In the application of Section 3, we identified a few outliers on an ad hoc basis and removed them from the data before estimating the model, because we wanted to find the validity and intercept bias for the bulk of the data.

Clearly, this could be improved. In the literature, some advanced methods have been proposed for robust estimation of SEMs in the presence of outliers; see, e.g., Yuan and Bentler (1998). However, no software is currently available that can apply these methods in combination with a finite population sampling design. More work could be done in this area.

In the method as described here, we did not introduce any prior assumptions about the relative measurement quality of each observed variable (apart from the audit data). Thus, in our application we did not presuppose that some of the sources in Table 3.2 were more or less prone to measurement errors than other ones. In this respect, the comparison between the validities of the variables in this application was completely data-driven. If a researcher does have prior knowledge about the relative merits of each data source, these could be incorporated in the model by means of equality or inequality constraints on parameters (Rosseel, 2012). Alternatively, it may then be natural to use a Bayesian SEM (Palomo et al., 2007).

In the type of model that was used here, measurement errors are considered to follow a continuous distribution. In practice, measurements on the same theoretical variable in a survey and an administrative source are sometimes found to be exactly equal for a substantial subset of the units. This is often explained by assuming that measurement errors in survey and administrative data are 'intermittent', i.e., there is a non-zero probability of observing the true value. Guarnera and Varriale (2015) consider a latent class model for measurement errors in numerical variables which explicitly takes this property into account. An alternative interpretation of the above phenomenon is that measurement errors in different sources are correlated because the measurement procedures cannot be considered independent.[6)] For instance, it might happen that some units simply report the same value of Turnover in the survey that they provided previously to the tax authorities, without going back to their original records. Correlated measurement errors can be taken into account in the SEM framework, provided that sufficient other indicators of the latent variables are available.

As remarked in footnote 1, the SEM in this paper yields estimates of the so-called indicator validity or empirical validity of the observed variables. Estimating the theoretical validity by factoring out the reliability component requires a more complex SEM, the so-called *multitrait-multimethod model*. This approach has been applied successfully in survey questionnaire design (Scherpenzeel and Saris, 1997), but it is not readily applicable to administrative data. Oberski et al. (2015) have recently proposed a generalisation of multitrait-multimethod models that may be useful for estimating the theoretical validity of administrative data. Their framework also provides alternative ways to model dependencies between measurement errors in different sources.

4.3 Potential applications

In the context of the application in Section 3, estimating the validity and intercept bias was useful to help deciding whether a specific administrative source could replace an existing sample survey, possibly after a model-based correction. Another, similar type of application might involve comparing several potential (administrative) sources for the same target variable and choosing the best one. This could be relevant for instance for NSIs that are moving towards a population census based on register data (Berka et al., 2012). Of course, the decision to use or

[6)] Thanks to Daniel Oberski for pointing this out.

not to use an administrative data source for statistics should be based on other criteria as well, besides the measurement quality. See, e.g., Daas et al. (2011) for a comprehensive overview of relevant quality indicators for administrative data. In addition, the outcome of a model-based analysis should always be compared with expectations based on other, qualitative knowledge about an administrative data source. For statistics that are already based on administrative data or mixed sources, the method described in this paper could be useful to quantify the influence of measurement errors on published statistical results.

The multi-group SEM with an audit sample as used in this paper can be applied to answer other research questions too. One interesting application in official statistics might be to compare the effects of automatic editing and manual editing on administrative or survey data. Many NSIs are now applying selective editing, which means that the most influential errors are edited manually by subject-matter specialists and the rest of the data are edited automatically (or not edited at all) for reasons of efficiency (De Waal et al., 2011). In evaluations of the quality of editing, the manually-edited data are usually considered to be the 'gold standard'. The quality of an automatic editing method is then assessed by comparing its outcome to that of manual editing. One drawback of this evaluation approach is that it requires the same data to be edited both automatically and manually, which is sometimes done in evaluation studies but not during regular production. Ilves and Laitila (2009) proposed to estimate the residual bias after selective editing as part of the regular production process by applying probability sampling (rather than cut-off sampling) to select records for manual editing. They retained the assumption that the manually-edited data are error-free. It may be interesting to combine the probability editing approach with our SEM approach, by selecting a small, random subset of the records as an audit sample which is submitted to a more intense form of manual editing. The outcome of this intense form of editing is then taken as the 'gold standard', to which both automatic and regular (i.e., less intensive) manual editing can be compared. A model similar to the one used here could then be applied to obtain separate estimates of the residual bias after automatic and regular manual editing. In some applications, a model-based bias correction might replace (part of) the regular manual editing to yield a more efficient data editing process. This alternative seems interesting in particular for large administrative data sets, where even traditional selective editing methods may be too resource-demanding.

The use of an audit sample to identify an SEM may also be relevant in some applications outside official statistics. SEMs are frequently used as an analysis technique in sociology, political science, and other social sciences, as well as in econometrics. Here, the role of a latent variable differs from the application in this paper: it represents a theoretical construct that is defined by the model, rather than a true value that exists independently of the model. As such, researchers in these areas are seldom interested in the true metrics of latent variables, and intercept bias is not usually a direct concern. However, this type of study often involves a comparison between groups (e.g., across countries, across subpopulations, or across time) and in that case different amounts of intercept bias or unequal factor loadings between groups can invalidate the outcome. Bielby (1986) described a hypothetical example where the use of reference indicators to achieve model identification leads to the wrong substantive conclusions in a two-group comparison, because the variables that are used as reference indicators have different values of τ and λ in each group. If the measurement parameters of these indicators are restricted to be invariant across groups for identification purposes, the implied metric of the underlying latent variable is different in each group, and across-group comparisons of the structural parameters are therefore misleading. One could, for instance, conclude that the effect of Education on Income is the same for men and women in a country, when in reality the effect is different (or vice versa), simply because men and women reacted differently (on average) to some of the questions by which Education and Income were measured. This is particularly problematic as the invariance of potential reference indicators cannot be tested or otherwise checked with the data themselves (Bielby, 1986).

For variables where the collection of 'gold standard' data is theoretically possible – if only for a small subset of the units by subjecting them to a non-standard, expensive form of observation – the use of an audit sample may provide a solution to this problem (Sobel and Arminger, 1986). If an audit sample is conducted in each group (country, subpopulation, time point), this yields a set of partially observed 'gold standard' variables that can be used as secure reference indicators to identify the multi-group model. Of course, the assumption that these indicators are truly invariant across groups still cannot be tested. But the data collection process for the audit sample could be designed in such a way that across-group differences are ruled out as much as possible, which may be easier to control for a small sample than during the large-scale field work for the actual survey. As noted in Section 2.1, it is in fact possible to allow for measurement errors in the audit data, provided that these do not affect the scale of measurement. Depending on the type of variable and the way the audit data are collected, the latter assumption may sometimes be more realistic than the assumption of no errors.

Acknowledgements

We would like to thank Arjen de Boer, Guus van de Burgt, Danny van Elswijk, Joost van der Leeden, Roy Manglie, and Mathieu van der Vijgh for providing the data for the application in Section 3. We also want to thank Cees Elzinga, Harry Ganzeboom, and Barry Schouten for their helpful comments on previous versions of the paper.

References

Allison, P. D. (1987). Estimation of Linear Models with Incomplete Data. *Sociological Methodology 17*, 71–103.

Alwin, D. F. (2007). *Margins of Errors*. New York: John Wiley & Sons.

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

Bakker, B. F. M. (2011a). Micro-Integration. Method Series, Statistics Netherlands, The Hague.

Bakker, B. F. M. (2011b). Micro-Integration: State of the art. In *ESSnet on Data Integration, Report on WP1*, pp. 77–107.

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

Bakker, B. F. M. and P. J. H. Daas (2012). Methodological Challenges of Register-Based Research. *Statistica Neerlandica 66*, 2–7.

Baumgartner, H. and J.-B. E. M. Steenkamp (1998). Multi-Group Latent Variable Models for Varying Numbers of Items and Factors with Cross-National and Longitudinal Applications. *Marketing Letters 9*, 21–35.

Berka, C., S. Humer, M. Moser, M. Lenk, H. Rechta, and E. Schwerer (2012). Combination of Evidence from Multiple Administrative Data Sources: Quality Assessment of the Austrian Register-Based Census 2011. *Statistica Neerlandica 66*, 18–33.

Bethlehem, J. (2008). Surveys without Questions. In De Leeuw, Hox, and Dillman (Eds.), *International Handbook of Survey Methodology*, pp. 500–511. New York: Psychology Press.

Bielby, W. T. (1986). Arbitrary Metrics in Multiple-Indicator Models of Latent Variables. *Sociological Methods and Research 15*, 3–23.

Biemer, P. P. (2011). *Latent Class Analysis of Survey Error*. Hoboken, New Jersey: John Wiley & Sons.

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

Boomsma, A. (1982). The Robustness of LISREL against Small Sample Sizes in Factor Analysis Models. In Jöreskog and Wold (Eds.), *Systems under Indirect Observation*, Volume I, pp. 149–173. Amsterdam, The Netherlands: North-Holland Publishing Company.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement Error in Survey Data. In Heckman and Leamer (Eds.), *Handbook of Econometrics*, Volume 5, pp. 3705–3843. Amsterdam, The Netherlands: Elsevier.

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, A. Bernardi, F. Cerroni, T. Laitila, A. Wallgren, and B. Wallgren (2011). List of Quality Groups and Indicators Identified for Administrative Data Sources. BLUE-ETS Project, Deliverable 4.1. Available at: http://www.blue-ets.istat.it/.

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

European Commission (2006). Commission Regulation (EC) No 1503/2006 of 28 September 2006 implementing and amending Council Regulation (EC) No 1165/98 concerning short-term statistics as regards definitions of variables, list of variables and frequency of data compilation. Published in the Official Journal of the European Union L281, 12 October 2006, pp. 15–30.

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

Guarnera, U. and R. Varriale (2015). Estimation and Editing for Data from Different Sources. An Approach Based on Latent Class Model. Working Paper No. 32, UN/ECE Work Session on Statistical Data Editing, Budapest.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.

Ilves, M. and T. Laitila (2009). Probability-Sampling Approach to Editing. *Austrian Journal of Statistics 38*, 171–182.

Jöreskog, K. G. (1971). Statistical Analysis of Sets of Congeneric Tests. *Psychometrika 36*, 109–133.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (second ed.). New York: John Wiley & Sons.

Little, T. D., D. W. Slegers, and N. A. Card (2006). A Non-Arbitrary Method of Identifying and Scaling Latent Variables in SEM and MACS Models. *Structural Equation Modeling 13*, 59–72.

Lord, F. M. and M. R. Novick (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software 9*, 1–19.

Meijer, E., S. Rohwedder, and T. Wansbeek (2012). Measurement Error in Earnings Data: Using a Mixture Model Approach to Combine Survey and Register Data. *Journal of Business & Economic Statistics 30*, 191–201.

Muthén, B. O. and A. Satorra (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology 25*, 267–316.

Nordbotten, S. (1955). Measuring the Error of Editing the Questionnaires in a Census. *Journal of the American Statistical Association 50*, 364–369.

Oberski, D. L. (2014). lavaan.survey: An R Package for Complex Survey Analysis of Structural Equation Models. *Journal of Statistical Software 57*, 1–27.

Oberski, D. L. (2015). Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model. Preprint.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter (2015). Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models. Manuscript, submitted for publication.

Palomo, J., D. B. Dunson, and K. Bollen (2007). Bayesian Structural Equation Modeling. In Lee (Ed.), *Handbook of Latent Variable and Related Models*, pp. 163–188. Amsterdam: Elsevier.

Papadopoulos, S. and Y. Amemiya (2005). Correlated Samples with Fixed and Nonnormal Latent Variables. *The Annals of Statistics 33*, 2732–2757.

Pavlopoulos, D. and J. K. Vermunt (2015). Measuring Temporary Employment. Do Survey or Register Data Tell the Truth? *Survey Methodology 41*, 197–214.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software 48*, 1–36.

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

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

Särndal, C.-E., B. Swensson, and J. H. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Satorra, A. (1992). Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures. *Sociological Methodology 22*, 249–278.

Satorra, A. (2002). Asymptotic Robustness in Multiple Group Linear-Latent Variable Models. *Econometric Theory 18*, 297–312.

Satorra, A. and P. M. Bentler (1986). Some Robustness Issues of Goodness of Fit Statistics in Covariance Structure Analysis. In *ASA 1986 Proceedings of the Business and Economic Statistics Section*, pp. 549–554.

Satorra, A. and P. M. Bentler (1994). Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis. In Von Eye and Clogg (Eds.), *Latent Variables Analysis: Applications to Developmental Research*, pp. 399–419. Thousand Oaks: SAGE Publications.

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

Scholtus, S. (2014). Explicit and Implicit Calibration of Covariance and Mean Structures. Discussion Paper 2014-09, Statistics Netherlands, The Hague.

Scholtus, S. and B. F. M. Bakker (2013). Estimating the Validity of Administrative and Survey Variables by means of Structural Equation Models. Paper presented at the conference New Techniques and Technologies for Statistics 2013, Brussels.

Sobel, M. E. and G. Arminger (1986). Platonic and Operational True Scores in Covariance Structure Analysis. *Sociological Methods and Research 15*, 44–58.

Stapleton, L. M. (2006). An Assessment of Practical Solutions for Structural Equation Modeling with Complex Survey Data. *Structural Equation Modeling 13*, 28–58.

Van Delden, A., R. Banning, A. De Boer, and J. Pannekoek (2015). Analysing whether Sample Survey Data can be Replaced by Administrative Data. Paper presented at the conference New Techniques and Technologies for Statistics 2015, Brussels.

Van Delden, A. and P.-P. De Wolf (2013). A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data. Paper presented at the conference New Techniques and Technologies for Statistics 2013, Brussels.

Yuan, K.-H. and P. M. Bentler (1998). Structural Equation Modeling with Robust Covariances. *Sociological Methodology 28*, 363–396.

Zhang, L.-C. (2012). Topics of Statistical Theory for Register-Based Statistics and Data Integration. *Statistica Neerlandica 66*, 41–63.

Appendix

I Additional methodology

In this appendix, a more detailed and technical description is given of the methodology from Section 2. Section I.1 reviews some general results on SEM estimation. Section I.2 considers adjustments regarding missing data that are needed for the application in this paper. Section I.3 provides formulae for additional fit measures used in this application. Finally, an additional adjustment to account for interdependencies between groups is discussed in Section I.4.

I.1 PML estimation for SEMs

We begin by reviewing some of the theory behind SEM estimation, starting with ML estimation for i.i.d. normal data and moving on to PML estimation for non-normal data and complex survey data. A more comprehensive discussion of these topics, as well as other estimation methods, can be found in Muthén and Satorra (1995) or Oberski (2014) (for a single group) and Satorra (2002) (for multiple groups).

We consider a multiple-group SEM, from which the single-group model follows as a special case. Suppose there are G groups with n_g sampled units in group g, and the samples are independent between groups. The total sample size is $n = \sum_{g=1}^{G} n_g$. Let $y_{gi} = (y_{gi1}, \dots, y_{gip})'$ denote the vector of observed variables for unit i in group g. In contrast with the notation of Section 2, we use a matrix of uncentered cross-product moments to summarise the observed data in each group: $S_g^0 = (1/n_g)\sum_{i=1}^{n_g} y_{gi} y_{gi}'$, where it is assumed that a constant 1 is included as one of the observed variables. Note: this last assumption implies that we can obtain the observed means and covariances $\bar{y}_g$ and S_g from S_g^0, and vice versa. In addition, let $s_g^0 = \mathrm{vech}(S_g^0)$, where $\mathrm{vech}(.)$ denotes the operator that vectorises a symmetric matrix by stacking the non-redundant elements column-wise (Harville, 1997). The population equivalents of S_g^0 and s_g^0 (i.e., the matrix and vector to which these quantities converge as $n_g \to \infty$) are denoted by Σ_g^0 and σ_g^0, respectively. We also define S^0 as the block-diagonal matrix with $S_1^0, \dots, S_G^0$ as blocks along the main diagonal, and $s^0 = ((s_1^0)', \dots, (s_G^0)')'$; analogously, we define Σ^0 and σ^0.

For G groups, the distance function F_{ML} that was mentioned in Section 2 is given by:

$$F_{\mathrm{ML}}(\vartheta) = \sum_{g=1}^{G} \frac{n_g}{n} \left\{ \log|\Sigma_g^0(\vartheta)| + \mathrm{tr}(S_g^0 \Sigma_g^0(\vartheta)^{-1}) - \log|S_g^0| - p \right\}, \tag{7}$$

where $\mathrm{tr}(.)$ denotes the trace of a matrix. It can be shown that minimising F_{ML} is asymptotically equivalent to minimising the following 'weighted least squares' function:

$$\begin{aligned} F_{\mathrm{WLS}}(\vartheta) &= \sum_{g=1}^{G} \frac{n_g}{n} \left\{ s_g^0 - \sigma_g^0(\vartheta) \right\}' \hat{V}_{g,\mathrm{ML}} \left\{ s_g^0 - \sigma_g^0(\vartheta) \right\} \\ &= \left\{ s^0 - \sigma^0(\vartheta) \right\}' \hat{V}_{\mathrm{ML}} \left\{ s^0 - \sigma^0(\vartheta) \right\}, \end{aligned} \tag{8}$$

with

$$\hat{V}_{g,\mathrm{ML}} = \frac{1}{2} D' \left\{ (S_g^0)^{-1} \otimes (S_g^0)^{-1} \right\} D,$$

and D the so-called duplication matrix (Harville, 1997). Also, $\hat{V}_{\mathrm{ML}}$ is a block-diagonal matrix with $(n_g/n)\hat{V}_{g,\mathrm{ML}}$ as blocks along the main diagonal. Let V_{ML} denote the population equivalent of $\hat{V}_{\mathrm{ML}}$. If the data are i.i.d. multivariate normal, then it can be shown that V_{ML} is identical to Γ^{-1}, where Γ denotes the asymptotic variance-covariance matrix of $\sqrt{n}s^0$.

Let $\hat{\vartheta}$ be the estimator that is obtained by minimising (7) or (8). If the assumption of i.i.d. multivariate normal data holds, then the asymptotic variance-covariance matrix of $\hat{\vartheta}$ is given by

$$\mathrm{avar}(\hat{\vartheta}) = \frac{1}{n}(\Delta' V_{\mathrm{ML}}\Delta)^{-1}, \tag{9}$$

with $\Delta = \partial\sigma^0(\vartheta)/\partial\vartheta'$. Furthermore, under the hypothesis that the model holds, the test statistic $X^2_{\mathrm{ML}} = (n-1)F_{\mathrm{ML}}$ is asymptotically distributed as a chi-square variate with degrees of freedom equal to $df = \mathrm{rank}(\Delta'_\perp \Gamma \Delta_\perp)$. Here, $\Delta_\perp$ denotes an orthogonal complement to the matrix Δ (Harville, 1997; Satorra, 2002). Typically, though not always, df equals the number of distinct observed moments (means and covariances) used to estimate the model minus the number of free parameters to be estimated (Bollen, 1989).

When the data are not normally distributed (but the i.i.d. assumption does hold), minimising (7) or (8) still provides consistent point estimates under rather general conditions (Bollen, 1989). Asymptotic standard errors based on (9) may be too small in this case. The correct expression for the asymptotic variance-covariance matrix of $\hat{\vartheta}$ is now:

$$\mathrm{avar}(\hat{\vartheta}) = \frac{1}{n}(\Delta' V_{\mathrm{ML}}\Delta)^{-1}\Delta' V_{\mathrm{ML}}\Gamma V_{\mathrm{ML}}\Delta(\Delta' V_{\mathrm{ML}}\Delta)^{-1}, \tag{10}$$

which reduces to (9) when $V_{\mathrm{ML}} = \Gamma^{-1}$. The asymptotic distribution of X^2_{ML} also need not be chi-square in this case. Satorra and Bentler (1994) proposed a relatively simple adjustment to X^2_{ML}. Define

$$\hat{U} = \hat{V}_{\mathrm{ML}} - \hat{V}_{\mathrm{ML}}\hat{\Delta}(\hat{\Delta}'\hat{V}_{\mathrm{ML}}\hat{\Delta})^{-1}\hat{\Delta}'\hat{V}_{\mathrm{ML}}$$

and

$$\hat{c}_{\mathrm{SB}} = \mathrm{tr}(\hat{U}\hat{\Gamma})/df. \tag{11}$$

In (11), $\hat{\Delta}$ is obtained by evaluating Δ at $\vartheta = \hat{\vartheta}$ and $\hat{\Gamma}$ is an appropriate estimate of Γ (see below). The Satorra-Bentler-corrected test statistic is $X^2_{\mathrm{SB}} = X^2_{\mathrm{ML}}/\hat{c}_{\mathrm{SB}}$, with the chi-square distribution with df degrees of freedom as its reference distribution (if the model holds).

To estimate Γ, the following technique is useful. Define $d^0_{gi} = \mathrm{vech}(y_{gi}y'_{gi})$, so that $s^0_g = (1/n_g)\sum_{i=1}^{n_g} d^0_{gi}$. Since this re-defines s^0_g as a sample mean, it can be shown that an appropriate estimator for $\mathrm{avar}(\sqrt{n_g}s^0_g)$ is given by

$$\hat{\Gamma}_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(d^0_{gi} - s^0_g)(d_{gi} - s^0_g)'.$$

Hence, $\Gamma = \mathrm{avar}(\sqrt{n}s^0)$ may be estimated by

$$\hat{\Gamma} = \begin{bmatrix} \frac{n}{n_1}\hat{\Gamma}_1 & & & \\ & \frac{n}{n_2}\hat{\Gamma}_2 & & \\ & & \ddots & \\ & & & \frac{n}{n_G}\hat{\Gamma}_G \end{bmatrix}. \tag{12}$$

For complex survey designs, one should first of all replace S^0 by a design-consistent estimator of Σ^0. Muthén and Satorra (1995) and Oberski (2014) consider the general case of a survey design that involves stratification, multi-stage selection and clustering. Essentially, in this case we can write $s_g^0 = (1/N_g) \sum_{i=1}^{n_g} w_{gi} d_{gi}^0$ for some weights w_{gi} that depend on the survey design, with $N_g = \sum_{i=1}^{n_g} w_{gi}$. To apply the PML approach, we can still use (10), (11), and (12), provided that each $\hat{\Gamma}_g$ is replaced by a variance estimator that takes the sample design for group g into account. The R package `lavaan.survey` implements this by referring to the variance estimation functionality of the `survey` package.

It should be noted that expression (12) is based on the assumption that the samples are independent between groups. For survey designs that involve without-replacement sampling, this will be false in general unless the survey happens to be stratified by the variable that defines the groups. In particular, this assumption is violated in the application of Section 3. In Section I.4 below, an approximate correction is proposed. As this correction is rather involved and turns out to have only a minor effect in our application, it is ignored in the rest of this paper.

I.2 Missing data

The use of an audit sample leads naturally to a two-group SEM with some of the variables missing by design in the second group. As described in Section 2.3, Baumgartner and Steenkamp (1998) suggested that these missing values can be accounted for by imputing random, normally-distributed values with mean zero and variance one, such that the imputed variables are uncorrelated to all other variables in the second group. (That is, they are both uncorrelated amongst themselves and uncorrelated to the observed variables.) A practical algorithm for obtaining such imputations works as follows:

1. Start by imputing random values from $N(0,1)$ for each missing variable separately.
2. For each missing variable y_k in turn, estimate a linear regression model with all other variables as predictors (in the second group). Replace y_k by its residual $\hat{e}_k = y_k - \hat{y}_k$ from the estimated model. By construction, this residual is uncorrelated to the other variables in the second group. Note: When treating the r^{th} missing variable, the updated values are used for missing variables $1, \dots, r-1$.
3. Finally, rescale each missing variable in the second group to have mean 0 and variance 1.

In case a complex survey design is used, this should be taken into account when estimating the regression models in the second step. Also, the rescaling in the last step should then be done so that the design-consistent estimates of the mean and variance equal 0 and 1, respectively.

As mentioned in Section 2.3, the measurement equations for the missing variables in the second group are chosen in such a way that the means, variances and covariances involving these variables are reproduced exactly by the SEM, while the estimation of the rest of the model is not affected by these variables. The sample moments involving the missing variables have thus been fixed so that they do not contribute to F_{ML} (or any other fitting function). The degrees of freedom of the model should be corrected to take this into account. Let q denote the number of missing variables in the second group and let df the denote the uncorrected degrees of freedom of the model, computed as if the imputed values were ordinary observed values. Since we have fixed q means and

$$p + (p-1) + \cdots + (p-q+1) = pq - \frac{q(q-1)}{2}$$

distinct covariances, the correct degrees of freedom should be:

$$df^* = df - q\left(p - \frac{q-3}{2}\right). \tag{13}$$

Baumgartner and Steenkamp (1998) applied the above approach only in the context of standard ML estimation. For PML estimation, we have to make an additional adjustment to $\hat{\Gamma}_2 = \widehat{\text{avar}}(\sqrt{n_2}s_2^0)$ (or, more generally, to $\hat{\Gamma}_g$ for each group g in which missing variables have been imputed in this way). Since the observed means and covariances involving the imputed variables are fixed, all elements of the corresponding rows and columns of $\hat{\Gamma}_2$ should be set to zero. With this adjustment, PML estimation yields the same results as if the variables that are missing by design were left out of the model.

Note that, in particular, the Satorra-Bentler correction factor of formula (11) is replaced in this context by

$$\hat{c}_{\text{SB}}^* = \text{tr}(\hat{U}\hat{\Gamma}^*)/df^*, \tag{14}$$

where df^* is given by (13) and $\hat{\Gamma}^*$ is obtained by making the above-mentioned adjustment to $\hat{\Gamma}$ from (12). The overall fit of the model can now be tested by comparing $X_{\text{SB}}^{2*} = X_{\text{ML}}^2/\hat{c}_{\text{SB}}^*$ to a chi-square distribution with df^* degrees of freedom.

I.3 Other fit measures

In the application from Section 3, several other measures were used in addition to X_{SB}^{2*} to evaluate the model fit. For the sake of completeness, we provide expressions for the robust (PML) versions of these fit measures, with adjustments to account for the imputed values in the second group (see Section I.2). Note: The definitions of these fit measures are not fully standardised. The following formulae are based on the default implementation in lavaan.

- Comparative Fit Index (CFI):

$$CFI_{\text{SB}}^* = 1 - \frac{\max\left\{X_{\text{SB}}^{2*} - df^*, 0\right\}}{\max\left\{X_{\text{SB}}^{2*} - df^*, X_{\text{SB},0}^{2*} - df_0^*, 0\right\}}. \tag{15}$$

- Tucker-Lewis Index (TLI):

$$TLI_{\text{SB}}^* = \frac{(X_{\text{SB},0}^{2*}/df_0^*) - (X_{\text{SB}}^{2*}/df^*)}{(X_{\text{SB},0}^{2*}/df_0^*) - 1}. \tag{16}$$

- Root Mean Square Error of Approximation (RMSEA):

$$RMSEA_{\text{SB}}^* = \sqrt{G\max\left\{N^{-1}(X_{\text{SB}}^{2*} - df^*), 0\right\}/df^*}. \tag{17}$$

Note: The CFI and TLI compare the fit of the model to that of a so-called baseline model. In the application of Section 3, we used the default baseline model selected by lavaan: this is the independence model with no restrictions across groups and with each observed variable modelled as $y_k = \tau_k + \epsilon_k$, with $\text{cov}(\epsilon_k, \epsilon_l) = 0$ for all $k \neq l$. In expressions (15) and (16), $X_{\text{SB},0}^{2*}$ and df_0^* refer to this baseline model. These adjusted quantities can be obtained from their unadjusted versions $X_{\text{SB},0}^2$ and df_0 analogously to Section I.2, with one subtle difference in the definition of df_0^*. Under the baseline model, the intercepts and error variances of the q imputed variables in the second group are not fixed (as in our original model) but estimated. This means that our adjustment to df_0 needs to account for $2q$ degrees of freedom less than before. Hence, the correction formula for the degrees of freedom of the baseline model becomes:

$$df_0^* = df_0 - q\left(p - \frac{q-3}{2}\right) + 2q = df_0 - q\left(p - \frac{q+1}{2}\right). \tag{18}$$

I.4 A two-group model with correlated samples

As was suggested at the end of Section I.1, for applications where the samples are not independent across groups, it would make sense from a theoretical point of view to include the off-diagonal blocks $\frac{n}{\sqrt{n_g n_h}}\hat{\Gamma}_{gh}$ in (12), where $\hat{\Gamma}_{gh}$ denotes an estimate of $\text{acov}(\sqrt{n_g}s_g^0, \sqrt{n_h}s_h^0)$ $(g \neq h)$. As far as we are aware, this problem has not been treated in the SEM literature. Papadopoulos and Amemiya (2005) considered correlation between groups in the case where the groups represent waves of a longitudinal study and the same respondents are observed multiple times, but they did not take other aspects of finite-population sampling into account. Here, we do not attempt to tackle the general case, but only consider an approximate solution that works for applications like the one in Section 3.

Consider a two-group SEM, where the first group of n_1 units represents a subsample (the audit sample) taken without replacement from a larger original sample of n units and the second group consists of the remaining $n - n_1$ units. The sample for the first group is seen to be a two-phase sample (Särndal et al., 1992), with the first phase given by the design of the original sample and the second phase given by the design used for subsampling. Similarly, the sample for the second group is also a two-phase sample: the first phase is the same as for the first group, while the second phase amounts to taking the complement of the first-group subsample. It is obvious that the samples for the two groups are dependent.

For clarity we consider the means and covariances separately again. We use $\mathcal{S}$ to denote the first-phase sample and $\mathcal{S}_g$ to denote the subsample for group g. Let w_i denote the weight of unit i in the first-phase sample. The final weight for two-phase sampling is denoted as $w_{gi} = w_i v_{gi}$ for units in group g $(g = 1, 2)$. Design-consistent estimates of the means μ and (vectorised) covariances σ are obtained from group g as follows:

$$\bar{y}_g = \frac{1}{N_g}\sum_{i\in\mathcal{S}_g} w_{gi}y_i,$$

$$s_g = \frac{1}{N_g}\sum_{i\in\mathcal{S}_g} w_{gi}d_{gi},$$

with $d_{gi} = \text{vech}((y_i - \bar{y}_g)(y_i - \bar{y}_g)')$ and $N_g = \sum_{i\in\mathcal{S}_g} w_{gi}$ [cf. Muthén and Satorra (1995)]. Similarly, design-consistent estimates based on the entire first-phase sample are given by:

$$\bar{y} = \frac{1}{N}\sum_{i\in\mathcal{S}} w_i y_i,$$

$$s = \frac{1}{N}\sum_{i\in\mathcal{S}} w_i d_i,$$

with $d_i = \text{vech}((y_i - \bar{y})(y_i - \bar{y})')$ and $N = \sum_{i\in\mathcal{S}} w_i$ which equals the population size. The objective now is to find expressions for $\text{acov}(\bar{y}_1, \bar{y}_2)$ and $\text{acov}(s_1, s_2)$.

To simplify matters, we make the following assumptions which are satisfied in the application of Section 3:

1. It holds that $1 \ll n_1 \ll n$, i.e., the audit sample is much smaller than the original sample, but still sufficiently large to ignore terms of the order $O(1/n_1)$.
2. The inclusion probabilities for the second phase do not depend on the realisation of the first-phase sample. This means that we can consider the second-phase weights v_{gi} as fixed numbers.[7)]
3. It holds that $N_g = N$ for each group. (This is only approximately true in practice.)
4. We already have procedures in place to estimate the asymptotic (co)variances of $\bar{y}_g$ and s_g within each separate group, as well as those of $\bar{y}$ and s based on the entire first-phase sample.

First consider the covariances of the estimated group means. For all $i \in \mathcal{S}$, let $a_{1i} = 1$ if the unit is selected in the audit sample and $a_{1i} = 0$ otherwise. Conditionally on the first-phase sample, it holds that:

$$\begin{aligned}
\operatorname{acov}(\bar{y}_1, \bar{y}_2 | \mathcal{S}) &= \frac{1}{N^2} \operatorname{acov}\left(\sum_{i \in \mathcal{S}} w_{1i} a_{1i} y_i, \sum_{i \in \mathcal{S}} w_{2i} (1 - a_{1i}) y_i \,\middle|\, \mathcal{S} \right) \\
&= \frac{-1}{N^2} \operatorname{acov}\left(\sum_{i \in \mathcal{S}} w_{1i} a_{1i} y_i, \sum_{i \in \mathcal{S}} w_{2i} a_{1i} y_i \,\middle|\, \mathcal{S} \right) \\
&= \frac{-1}{N^2} \operatorname{acov}\left(\sum_{i \in \mathcal{S}} w_{1i} a_{1i} y_i, \sum_{i \in \mathcal{S}} w_{1i} \frac{v_{2i}}{v_{1i}} a_{1i} y_i \,\middle|\, \mathcal{S} \right) \\
&= -\operatorname{acov}(\bar{y}_1, \bar{y}_1^* | \mathcal{S}),
\end{aligned}$$

with $\bar{y}_1^* = \frac{1}{N} \sum_{i \in \mathcal{S}_1} w_{1i} y_i^*$ and $y_i^* = (v_{2i}/v_{1i}) y_i$. Assumption 3 was used in the first line. In the second line and in the definition of y_i^* we used assumption 2 that v_{1i} and v_{2i} are fixed.

The unconditional covariance can be obtained as:

$$\operatorname{acov}(\bar{y}_1, \bar{y}_2) = E\left\{ \operatorname{acov}(\bar{y}_1, \bar{y}_2 | \mathcal{S}) \right\} + \operatorname{acov}\left\{ E(\bar{y}_1 | \mathcal{S}), E(\bar{y}_2 | \mathcal{S}) \right\}.$$

Using $E(a_{1i} | \mathcal{S}) = 1/v_{1i}$, it is not difficult to show that $E(\bar{y}_g | \mathcal{S}) = \bar{y}$ for $g = 1, 2$. Thus, the second term evaluates to $\operatorname{avar}(\bar{y})$.

Using the same conditioning argument as before, we also find

$$\begin{aligned}
\operatorname{acov}(\bar{y}_1, \bar{y}_1^*) &= E\left\{ \operatorname{acov}(\bar{y}_1, \bar{y}_1^* | \mathcal{S}) \right\} + \operatorname{acov}\left\{ E(\bar{y}_1 | \mathcal{S}), E(\bar{y}_1^* | \mathcal{S}) \right\} \\
&= -E\left\{ \operatorname{acov}(\bar{y}_1, \bar{y}_2 | \mathcal{S}) \right\} + \operatorname{acov}(\bar{y}, \bar{y}^*),
\end{aligned}$$

with $\bar{y}^* = \frac{1}{N} \sum_{i \in \mathcal{S}} w_i y_i^*$. Combining these expressions, we obtain:

$$\operatorname{acov}(\bar{y}_1, \bar{y}_2) = \operatorname{acov}(\bar{y}, \bar{y}^*) - \operatorname{acov}(\bar{y}_1, \bar{y}_1^*) + \operatorname{avar}(\bar{y}). \tag{19}$$

According to assumption 4 above, the three expressions on the right-hand-side can be estimated by known procedures, because they depend on either the entire first-phase sample or the first group alone.

[7)] In the application of Section 3 this property holds because stratified simple random sampling is used in both phases and the stratification for the second phase is a coarser version of the stratification used in the first phase. It does *not* hold in general for two-phase sampling (Särndal et al., 1992, Section 9.2).

For the estimated group covariances, we first note that, under assumption 3, it is possible to rewrite s_g as follows:

$$\begin{aligned} s_g &= \frac{1}{N}\sum_{i\in\mathcal{S}_g} w_{gi}\mathrm{vech}((y_i - \bar{y} + \bar{y} - \bar{y}_g)(y_i - \bar{y} + \bar{y} - \bar{y}_g)') \\ &= \frac{1}{N}\sum_{i\in\mathcal{S}_g} w_{gi}d_i - \mathrm{vech}((\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})'). \end{aligned}$$

From this, it follows that

$$E(s_g|\mathcal{S}) = s - E\left\{\mathrm{vech}((\bar{y}_g - \bar{y})(\bar{y}_g - \bar{y})')|\mathcal{S}\right\} = s - \mathrm{vech}(\mathrm{var}(\bar{y}_g|\mathcal{S})). \qquad (20)$$

Regarding the second term: in the special case that simple random sampling without replacement is used in both phases, it is not difficult to show that

$$\mathrm{vech}(\mathrm{var}(\bar{y}_g|\mathcal{S})) = \frac{1}{n_g}\left(1 - \frac{n_g}{n}\right)\frac{n}{n-1}s.$$

In particular, for the first group with $1 \ll n_1 \ll n$, it follows that this term is approximately equal to $(1/n_1)s$ and thus much smaller than the first term in (20) by assumption 1. Under the assumption that $\mathrm{vech}(\mathrm{var}(\bar{y}_1|\mathcal{S}))$ is of the same order (or even smaller) for other practical sampling designs, it is acceptable to write $E(s_1|\mathcal{S}) \approx s$.

Furthermore, it holds approximately that

$$s \approx \frac{n_1}{n}s_1 + \frac{n - n_1}{n}s_2.$$

Hence, we obtain:

$$\begin{aligned} \mathrm{acov}(s_1, s_2) &\approx \mathrm{acov}\left(s_1, \frac{n}{n-n_1}s - \frac{n_1}{n-n_1}s_1\right) \\ &= \frac{n}{n-n_1}\mathrm{acov}(s_1, s) - \frac{n_1}{n-n_1}\mathrm{avar}(s_1). \end{aligned}$$

By assumption 4, we have a procedure for estimating the second term on the right. For the first term, we find:

$$\begin{aligned} \mathrm{acov}(s_1, s) &= E\left\{\mathrm{acov}(s_1, s|\mathcal{S})\right\} + \mathrm{acov}\left\{E(s_1|\mathcal{S}), E(s|\mathcal{S})\right\} \\ &\approx 0 + \mathrm{avar}(s), \end{aligned}$$

since the first component is zero and $E(s_1|\mathcal{S}) \approx s$. Hence:

$$\mathrm{acov}(s_1, s_2) \approx \frac{n}{n-n_1}\mathrm{avar}(s) - \frac{n_1}{n-n_1}\mathrm{avar}(s_1). \qquad (21)$$

Using (19) and (21) we obtain approximate estimates for the off-diagonal blocks in (12) when the above assumptions 1-4 are satisfied, as was the case in Section 3.

Table I.1 Fit measures for the final model (with adjusted $\hat{\Gamma}$)

NACE group	45112	45190	45200	45400	50100	50300	52100	52290
X^{2*}_{SB}	60.1	72.0	41.1	167.1	88.2	58.6	76.1	95.9
df^*	62	61	61	62	62	61	62	62
p value	0.546	0.158	0.976	0.000	0.016	0.565	0.108	0.004
$\hat{c}^*_{\mathrm{SB}}$	35.1	14.7	43.8	2.5	6.8	23.8	12.9	15.9
CFI^*_{SB}	1.000	0.987	1.000	0.929	0.919	1.000	0.931	0.973
TLI^*_{SB}	1.001	0.988	1.034	0.931	0.922	1.009	0.934	0.974
$RMSEA^*_{\mathrm{SB}}$	0.000	0.047	0.000	0.248	0.098	0.000	0.057	0.054

We applied the above adjustment to $\hat{\Gamma}$ in the application of Section 3. Table I.1 shows the resulting robust fit measures. As can be seen by comparison to Table 3.5, for most NACE groups the effect of the adjustment was very small. To the extent that there was an effect, it mostly improved the fit of the model. The effects on standard errors were also small (not shown). Therefore, for simplicity, we ignored this adjustment in the rest of this paper.

II Parameter estimates

Table II.1 Parameter estimates for the final model (trade sector)

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
λ_1	0.83	0.02	0.90	0.05	0.74	0.07	0.81	0.08
λ_2	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_3	1.03	0.01	0.95	0.03	1.30	0.32	0.98	0.01
λ_4	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_5	1.03	0.01	0.97	0.02	1.22	0.23	1.01	0.01
λ_6	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_7	0.79	0.01	0.89	0.02	1.29	0.19	0.80	0.04
λ_8	1.02	0.01	0.95	0.02	1.23	0.20	0.99	0.03
λ_9	1.01	0.01	1.01	0.00	1.21	0.19	0.98	0.02
λ_{10}	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
θ_{11}	1.24	0.57	9.57	2.32	2.77	1.33	1.02	0.42
θ_{22}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{33}	0.04	0.01	0.36	0.13	0.05	0.05	0.00	0.00
θ_{44}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{55}	0.03	0.02	0.41	0.18	0.05	0.04	0.00	0.00
θ_{66}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{77}	1.00	0.20	0.57	0.19	0.04	0.01	0.04	0.01
θ_{88}	0.06	0.02	0.41	0.19	0.05	0.03	0.01	0.00
θ_{99}	0.87	0.21	0.00	0.00	0.06	0.03	0.00	0.00
$\theta_{10,10}$	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
τ_1	1.04	0.17	0.81	0.37	1.17	0.30	1.04	0.15
τ_2	0^a	–	0^a	–	0^a	–	0^a	–
τ_3	−0.01	0.04	0.03	0.03	−0.00	0.06	0.00	0.00
τ_4	0^a	–	0^a	–	0^a	–	0^a	–
τ_5	−0.01	0.04	0.03	0.04	−0.01	0.08	−0.00	0.01
τ_6	0^a	–	0^a	–	0^a	–	0^a	–
τ_7	−0.04	0.04	−0.00	0.05	−0.04	0.08	0.01	0.03
τ_8	0.00	0.05	0.06	0.03	−0.02	0.08	0.00	0.01
τ_9	−0.01	0.05	−0.00	0.00	−0.04	0.08	0.00	0.01
τ_{10}	0^a	–	0^a	–	0^a	–	0^a	–
β_{31}	0.05	0.00	0.04	0.01	0.04	0.01	0.04	0.01
β_{32}	1.03	0.00	1.14	0.03	1.19	0.11	1.16	0.04
β_{41}	0^a	–	0.01	0.00	0.01	0.00	0^a	–
β_{43}	1.02	0.00	1.00	0.01	0.96	0.04	1.02	0.02
ψ_{11}	157	0.62	96.8	0.65	67.4	0.07	10.5	2.06
ψ_{22}	28.6	1.75	8.29	1.52	0.36	0.18	0.56	0.10
ψ_{12}	59.7	1.33	25.2	2.49	3.59	0.87	2.11	0.31
ψ_{33}	0.01	0.00	0.05	0.01	0.02	0.02	0.00	0.00
ψ_{44}	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00
α_1	3.44	0.21	4.06	0.40	3.02	0.36	1.31	0.22
α_2	1.21	0.06	0.94	0.08	0.15	0.02	0.29	0.06
α_3	0.02	0.01	0.02	0.03	0.02	0.02	0.00	0.01
α_4	0.03	0.01	−0.01	0.02	0.02	0.01	0.01	0.01

[a] parameter fixed a priori, value may depend on group as indicated

Table II.2 Parameter estimates for the final model (transportation sector)

	50100		50300		52100		52290	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
λ_1	0.74	0.09	0.57	0.09	0.79	0.07	0.80	0.03
λ_2	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_3	2.04	0.35	0.94	0.02	0.89	0.10	0.98	0.08
λ_4	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_5	1.10	0.07	0.90	0.02	1.04	0.09	1.02	0.06
λ_6	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_7	0.92	0.05	0.75	0.03	1.24	0.15	0.69	0.08
λ_8	1.12	0.06	0.89	0.02	1.09	0.08	1.01	0.06
λ_9	1.00	0.00	0.83	0.05	0.93	0.12	0.73	0.09
λ_{10}	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
θ_{11}	21.0	14.0	5.36	3.72	35.5	13.2	12.8	4.44
θ_{22}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{33}	9.03	6.61	0.06	0.04	12.0	6.98	7.32	2.32
θ_{44}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{55}	3.00	0.99	0.03	0.03	9.64	5.93	2.72	1.75
θ_{66}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{77}	3.42	1.61	0.62	0.29	16.5	6.04	61.1	9.53
θ_{88}	3.60	0.92	0.02	0.02	11.2	6.25	3.29	1.35
θ_{99}	0.03	0.03	0.98	0.48	8.84	5.47	32.4	9.71
$\theta_{10,10}$	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
τ_1	1.52	0.67	1.48	0.32	1.84	1.04	1.14	0.48
τ_2	0^a	–	0^a	–	0^a	–	0^a	–
τ_3	0.51	0.37	0.08	0.03	0.70	0.36	0.32	0.38
τ_4	0^a	–	0^a	–	0^a	–	0^a	–
τ_5	−0.86	0.27	0.07	0.03	0.45	0.39	0.20	0.39
τ_6	0^a	–	0^a	–	0^a	–	0^a	–
τ_7	0.15	0.16	0.18	0.03	0.03	0.55	1.14	0.44
τ_8	−0.69	0.21	0.07	0.03	0.32	0.41	0.24	0.41
τ_9	0.04	0.04	0.11	0.04	0.21	0.42	0.84	0.47
τ_{10}	0^a	–	0^a	–	0^a	–	0^a	–
β_{31}	0.20	0.05	0.07	0.03	0.08	0.01	0.07	0.00
β_{32}	2.06	0.42	1.02	0.05	0.95	0.07	1.05	0.01
β_{41}	0^a	–	0.01	0.00	0^a	–	0^a	–
β_{43}	1.03	0.06	1.01	0.01	1.05	0.04	1.04	0.01
ψ_{11}	202	0.41	120	0.78	920	4.43	704	2.62
ψ_{22}	5.83	3.58	6.10	2.52	39.7	16.9	94.9	20.4
ψ_{12}	12.56	8.83	17.5	6.94	128	34.2	120	16.3
ψ_{33}	6.68	3.54	0.17	0.06	3.26	1.83	0.26	0.11
ψ_{44}	0.84	0.45	0.02	0.01	0.57	0.38	0.16	0.08
α_1	4.20	0.97	2.72	0.40	12.8	1.46	9.42	0.63
α_2	0.27	0.17	0.11	0.04	1.04	0.40	2.70	0.44
α_3	1.06	0.41	0.12	0.08	0.56	0.21	0.09	0.06
α_4	−0.19	0.18	0.07	0.02	0.17	0.12	−0.02	0.06

^{a} parameter fixed a priori, value may depend on group as indicated