

# Joint estimation of survey error components in multivariate statistics

Daniel L. Oberski

Tilburg University, Netherlands and Universitat Pompeu Fabra, Spain

## Abstract

We outline a procedure for simultaneously estimating the “design” effects of different survey error components, in the context of structural equation models. The effects of clustering, measurement error, and non-normality, are jointly estimated for an example multivariate model involving reciprocal effects, instrumental variables, correlated error terms, and measurement error. The example is estimated on real data from the European Social Survey 2008.

It is shown how estimates of the effects of these different survey error components can be obtained. In the example given, it is also shown that the relative sizes of these effects are very different than commonly found in the estimation of means and totals. In particular, measurement error is an important factor in our example.

Finally, it is remarked that our general knowledge of the relative importance of different survey error components for multivariate statistics could be greatly increased by the application of the method discussed in this paper to a cross-section of real analyses.

## Introduction

The total survey error literature is rich in discussions of the effects of different error components on the bias and variance of mean estimators (starting points for the abounding literature are Cochran, 1977; Groves, 1989). Effects on parameters of multivariate models are also discussed, though less often (e.g. Scott & Holt, 1982; Lyberg, 1997; Biemer, Groves, & Lyberg, 2004). Such discussions, however, focus mostly on bias in the multivariate statistics due to survey errors, a notable exception being Kish and Frankel (1974). Here we will attempt to take also into account the effect on the variance of the estimates, with a focus on the simultaneous estimation of the effect of different error sources.

The same error components that affect means may also affect regression coefficients and other multivariate statistics such as factor loadings, and latent variable variances. The influences and relative influences they have, however, cannot be expected to be similar. For example, the design effect due to clustering for a mean is

$$1 + (c - 1)\rho, \tag{1}$$

where  $\rho$  is the intraclass correlation (icc) and  $c$  is the common cluster size (e.g. Cochran, 1977). Compare this with the same design effect for a simple regression coefficient, which approximately equals

$$1 + (c - 1)\rho_\epsilon\rho_x, \quad (2)$$

where  $\rho_\epsilon$  is the icc for the residuals and  $\rho_x$  the icc for the independent variable (Scott & Holt, 1982): clearly the design effect due to clustering for the regression coefficient is in general smaller than the design effect on means.

Random measurement error, meanwhile, does not add systematic errors and can therefore be ignored in the analysis of bias in means and totals<sup>1</sup>. On the other hand, regression coefficients are well-known to be biased by measurement error if left uncorrected (Fuller, 1987, 3). For example, in a simple linear regression of a dependent variable  $Y$  on an observed independent variable  $X$ , denote the linear regression slope of  $Y$  on  $X$  by  $\gamma$ . If  $X$  is not a perfect measure, but contains measurement error,  $X$  has a reliability  $\rho_{xx}$ , sometimes termed 'reliability ratio' in the survey error literature (e.g. Groves, 1989). If the true regression slope is denoted by  $\beta$ , the relationship between the regression slope of observed variables, the true slope, and the reliability is (Fuller, 1987, 5):

$$\gamma = \rho_{xx} \cdot \beta. \quad (3)$$

Thus the bias in the regression slope uncorrected for measurement error is multiplicative in the reliability. Since reliabilities of 0.7 are not uncommon (Saris & Gallhofer, 2007b; Alwin, 2007), large biases can occur.

Measurement error in the independent variable, when left uncorrected, will also decrease the explained variance and thus increase the variance of the estimated (standardized) coefficient. Since the mean square error (MSE) of the regression coefficient is the sum of bias squared and variance, it follows that measurement error influences the MSE through both bias and variance increase.

Depending on sample size, cluster size, and effect size, the effect on the mean square error of measurement error may exceed that of clustering or vice versa. Thus, careful attention should be given to the relative sizes of the effect of different error sources: they cannot be assumed similar to those for means and totals.

We will in turn discuss different error components, both sampling and nonsampling, and how their impact on multivariate statistics may be measured. We do this in a structural equation model context developed in the following section. Due to the generality of structural equation models, our discussion will cover (multivariate) regression, factor analysis, longitudinal models, and models with ordinal, count, and censored variables among others (Muthén, 1984).

## Structural equation models

Suppose a sample of  $J$  clusters has been drawn from a population, and in total  $n$  persons have been selected within the clusters by random sampling with probability weights  $w_i$ . In this way,  $n$  observations of the measures  $y$  are observed. These observed variables are

<sup>1</sup>Obviously measurement error biasing effects also exist in the form of systematic errors ('relative bias'); these do affect mean estimation (see also Biemer & Trewin, 1997).

assumed to be imperfect measures of a vector of variables  $\eta$ , with possibly correlated random measurement error  $\epsilon$ . Systematic stochastic measurement errors such as method and style effects can also be incorporated through the latent variable structure (see e.g. Werts & Linn, 1970; Jöreskog, 1970).

A structural equation model (SEM) can then be specified as

$$\eta = B_0\eta + \zeta, \quad (4)$$

$$y = \Lambda\eta + \epsilon, \quad (5)$$

$$\forall_{k \in K} \forall_{l \in L} (E(\epsilon_k, \zeta_l) = 0), \quad (6)$$

with  $\Phi_{K \times K}$  the covariance matrix of the latent variable disturbance term vector  $\zeta$  and  $\Psi_{L \times L}$  the covariance matrix of the measurement error variables  $\epsilon$ . This specification is known as the ‘‘LISREL all-y model’’ (Jöreskog, 1970). Other well-known model formulations are the Bentler-Weeks model (Bentler & Weeks, 1980) and the RAM model (McArdle & McDonald, 1984). All three models can be re-written into equivalent specifications to fit the form of the other models. The parameters of the model are collected into a vector  $\theta$ .

We assume there is a matrix of observed variances and covariances  $S$  on the  $p$  observed variables that converges in probability to a population covariance matrix  $\Sigma$ . The  $p(p+1)/2$  unique elements of  $S$  can be collected into a vector  $s := \text{vech}S$ .

The implied variance-covariance matrix by the model above is then

$$\Sigma(\theta) = B^{-1}\Lambda\Phi\Lambda'B^{-1} + \Psi, \quad (7)$$

where  $B := I - B_0$ . We collect the unique elements of  $\Sigma(\theta)$  into a vector  $\sigma(\theta) := \text{vech}\Sigma(\theta)$ .

Given the above assumptions, the parameters of the model can be consistently estimated by minimizing the weighted least squares fitting function

$$F = (s - \sigma(\theta))'V(s - \sigma(\theta)), \quad (8)$$

where  $V$  is a positive definite, possibly stochastic, weight matrix (Satorra, 1989).

The weighted least squares fitting function will be equivalent to maximum likelihood estimation if  $V$  is chosen as the inverse of the model-implied fourth-order moments under normality. Thus the discussion given here encompasses normal-theory maximum likelihood as well as generalized least squares and ‘asymptotic distribution free’ estimation methods, among others.

Consistency is not affected by the choice of  $V$ , as long as  $V$  does not violate identification conditions. Only one choice of  $V$  is asymptotically optimal, however; namely that  $V$  such that  $V$  converges in probability to  $\Gamma^{-1}$ , where  $\Gamma$  is (a function of) the matrix of fourth-order moments of  $y$ . The consistency of the estimates also does not depend on any assumption about the distribution of  $y$ .

Under the model the asymptotic variance of the estimates  $\hat{\theta}$  is

$$\text{avar}(\hat{\theta}) = n^{-1}(\Delta'V\Delta)^{-1}\Delta'V\Gamma V\Delta(\Delta'V\Delta)^{-1}, \quad (9)$$

where  $\Delta$  is the first derivative of the implied covariance matrix  $\Sigma(\theta)$  with respect to the parameters  $\theta$ , and  $\Gamma$  the matrix of fourth-order moments (e.g. Satorra, 1989). An expression for  $\Delta$  in terms of the parameters of the model was given by Neudecker and Satorra

(1991). We will employ this expression to calculate the asymptotic variance under different conditions.

The matrix  $\Gamma$  in equation 9 will play an important part in the discussion that follows: it is the matrix of fourth-order moments of  $y$ , and the primary means by which survey error components affect the variance of the estimates  $\hat{\theta}$ . In general, if there is no clustering a consistent estimate of  $\Gamma$  under general conditions is given by

$$\hat{\Gamma} = n^{-1} \sum_{i=1}^n (b_i - \bar{b})(b_i - \bar{b})', \quad (10)$$

where  $b_i = D^+ \text{vec}(y_i - \bar{y})(y_i - \bar{y})'$  (e.g. Fuller, 1987, 332) and  $D^+$  is the Moore-Penrose inverse of the duplication matrix (Magnus & Neudecker, 2002).

With clustering and weighting the estimate of  $\Gamma$  can be obtained by first aggregating to the level of the clusters while using the sampling weights. Then the estimate becomes

$$\hat{\Gamma}^{(c)} = \frac{J}{n^2(J-1)} \sum_{j=1}^J (b_j - \bar{b})(b_j - \bar{b})', \quad (11)$$

where  $b_j$  is the weighted sum of all  $b_i$ 's in cluster  $j$ , replacing  $\bar{y}$  with the weighted sample mean (Muthén & Satorra, 1995).

The matrices  $\hat{\Gamma}$  and  $\hat{\Gamma}^{(c)}$  provide consistent estimates of the fourth-order moments regardless of the distribution of  $y$ . If, however,  $y$  can be assumed to have a multivariate normal distribution, then the  $\Gamma$  matrix can be consistently estimated by

$$\hat{\Gamma}^* = 2D^+(S \otimes S)D^{+'}. \quad (12)$$

The fourth order moments are then a function only of the variances and covariances.

As we have remarked earlier, the choice of  $V$  determines the estimation procedure. By replacing  $\Gamma$  in equation 9 with  $\hat{\Gamma}^*$ ,  $\hat{\Gamma}$ , or  $\hat{\Gamma}^{(c)}$ , normal-theory variances, variances robust to non-normality, or cluster and weighting-corrected variances are obtained. Here it should be noted that the cluster-corrected variances of Muthén and Satorra (1995) are also robust to non-normality.

The elements of the measurement error variance matrix  $\Psi$  can, if identification conditions have been met, be estimated simultaneously with the 'structural' parameters  $B_0$  and  $\Phi$ . However, in practical applications, not enough information may be available to estimate the measurement error from the sample or doing so simultaneously might lead to very large models (Saris & Gallhofer, 2007a). In such cases an estimate of the measurement error in the measures  $y$  may be obtained from other sources, such as published evaluations of psychometric properties of scales or meta-analyses of measurement error estimates (Saris & Gallhofer, 2007b; Alwin, 2007). Correction for measurement error can then proceed by fixing the elements of  $\Psi$  (the 'single indicators' approach)<sup>2</sup>.

Now that we have developed some results for the general structural equation model we will discuss an application of a structural equation model from the literature. We then proceed to separately estimate the magnitude of the error components in the variance of parameters of an analysis of this real data set.

<sup>2</sup>An alternative method that will not be discussed here is the so-called 'covariance reduction' approach (Saris & Gallhofer, 2007a).

## Application of a structural equation model to real data

Saris and Gallhofer (2007a) provide an analysis of a structural equation model of social and political trust with corrections for measurement error. A simplified adaptation of their model is shown in figure 1. The model shown in the figure can be expressed as:

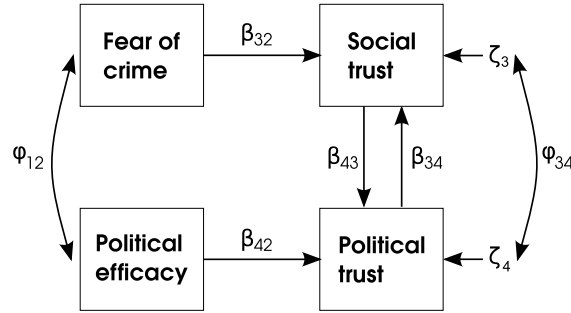


Figure 1. Structural equation model adapted from Saris & Gallhofer (2007a).

$$\text{SocTrust} = \beta_{34} \text{PolTrust} + \beta_{31} \text{Fear} + \zeta_3 \quad (13)$$

$$\text{Poltrust} = \beta_{43} \text{SocTrust} + \beta_{42} \text{Efficacy} + \zeta_4 \quad (14)$$

$$E(\phi_1 \zeta_3) = E(\phi_1 \zeta_4) = E(\phi_2 \zeta_3) = E(\phi_2 \zeta_4) = 0. \quad (15)$$

It can be seen that the model cannot be estimated with ordinary linear regression because it contains a reciprocal effect between social and political trust, which is identified by the “instruments” fear of crime and political efficacy. In line with standard practice in econometrics we allow for the possibility of a covariance between the disturbance terms  $\zeta_3$  and  $\zeta_4$ .

Whether this model is reasonable is not the topic being discussed here. We will assume that an interest exists in estimating the model among users of a survey and show how the effect of different survey error components on the parameters of interest can be estimated using the theory outlined in the previous sections.

The variables shown in the model are not observed variables, but rather they are constructs defined as influencing the answers to certain survey questions. For each of these constructs we can obtain at least two measures from the European Social Survey (ESS) round 4, conducted in 2008 (Jowell, Roberts, Fitzgerald, & Eva, 2007). As an illustration we will select one country only, Denmark, because it had a simple random sampling design, simplifying the discussion needed below. We stress, however, that our methods can be equally easily applied to designs with unequal inclusion probabilities.

The variables used to measure each of the four constructs can be found in the appendix. An estimate of the composite scores was constructed by taking the simple sum of indicators. Table 2 shows the histograms and summary statistics for the resulting sum scores.

In order to estimate the reliability of each construct and the associated error variance, we estimated a four-factor model using the software EQS 6.1 (Bentler, 1995). We obtained the standard errors of the reliability and error variance through a non-parametric bootstrap

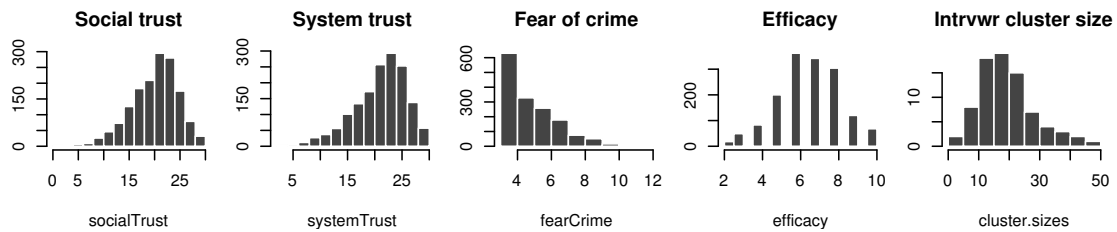


Table 1: Histograms of the observed variables and number of interviews per interviewer (cluster size) in Denmark.

Variable	Mean	Std dev	Skewness	Excess kurtosis	icc	$\hat{\rho}_x$ (s.e.)	$\hat{\psi}$ (s.e.)
socialTrust	20	4.7	-0.70	-0.80	0.25	0.73 (0.01)	6.0 (0.22)
systemTrust	21	4.8	-0.79	-0.54	0.11	0.77 (0.01)	6.3 (0.24)
fearCrime	5	1.7	0.77	-0.31	0.17	0.57 (0.02)	1.3 (0.04)
efficacy	7	1.7	-0.25	0.05	0.11	0.64 (0.03)	1.2 (0.07)

Table 2: Summary statistics for the Denmark dataset. The total sample size was 1610. The icc shown is the intra-interviewer correlation coefficient.

<sup>3</sup>. The resulting reliability and error variance estimates are shown in the last two columns of table 2.

The data collection mode was computer-assisted personal interviewing in the home of the respondent. Each interviewer conducted at least 4 and at most 48 interviews. Below we will investigate the effects on the variance of the estimates of the model due to correlation between the answers of different respondents interviewed by the same interviewer. The sixth column of table 2 shows the univariate intra-interviewer correlation coefficients<sup>4</sup>. It can be seen that considerable intra-interviewer correlations exist. The average number of interviews per interviewer was 20; the distribution of the number of interviews per interviewer ranged between 4 and 48 and is shown in the rightmost histogram.

Using the statistical software R 2.11.0 (R Development Core Team, 2010) and the package OpenMx (Boker et al., 2010), we estimated the model<sup>5</sup> shown in figure 1. The parameter estimates are shown in table 4.

The parameters of most interest to substantive researchers are the direct effects of social trust on political trust and vice versa, as well as the so-called 'total effects'. The direct effect of social trust on political trust is stronger than the converse effect. The total effects of social trust and political trust can be calculated as 1.0 and 0.40, respectively. This implies that for a given amount of change in social trust, political trust, which was measured on the same

<sup>3</sup>The procedure we used is similar to that of Raykov (2009), except that in contrast with the approach discussed there, we also allow for randomness in the loadings in estimating the error variance and reliability.

<sup>4</sup>The intra-interviewer correlation coefficient was estimated using R 2.11.0 by fitting a multilevel linear model with a random interviewer intercept to each variable. The icc was then estimated as the square root of the ratio of the random intercept variance to the residual variance.

<sup>5</sup>In formulating this model it was simpler to parameterize the error covariance as an effect of a latent variable with fixed loadings. This model is mathematically equivalent to the model shown in the figure but implies that the error variance parameters  $\phi$  of social and political trust equal  $\phi_i - \phi_{ij}$ .

scale, is expected to increase by the same amount. The reverse is not the case, as social trust can be expected to increase by only 40% of the change in political trust. This is largely in correspondence with the literature on social capital (e.g. Putnam, 2001).

We do not comment on whether the model discussed is correct. This must be assessed by thorough investigation into a combination of appropriate theoretical considerations and model fit to observed data, which is outside the scope of this paper. We only show how the contribution of different survey error components to the variance of the coefficients can be estimated, also for models involving such complexities as instrumental variables, reciprocal effects, and a correlated error term.

## Estimation of sampling and non-sampling errors in structural equation models

In the previous section we presented the structural equation modeling framework as well as an parameter estimates of a model found in the literature using real data. We will now discuss how different survey error components affect the variance of those estimates, and proceed to separately estimate their effects in the analysis presented.

### Complex sampling and (interviewer) clustering

Muthén and Satorra (1995) provided an in-depth discussion of the estimation of structural equation models under complex sampling. Our discussion of this topic largely follows their results.

Unequal selection probabilities necessitate the estimation of the covariance matrix by a weighted estimator. We will denote this estimator as  $S^{(c)}$ . The variance of the parameter estimates can be obtained by the normal variance estimator if an adjustment is made to the estimated  $\Gamma$  matrix of fourth-order moments. The effect of unequal sampling weights on the variance, operating through the  $\Gamma$  matrix, is therefore in general multiplicative. Indeed, all survey error components that affect the fourth-order moments have a multiplicative effect on the variance of the estimates.

Clustering does not affect the estimator needed for the estimation of  $\Sigma$ . The variance of the estimates is affected, however. Muthén and Satorra (1995) discuss two separate ways to take clustering into account: one design-based and one model-based. The model-based method proceeds by specifying a random effects (multilevel) model with the clusters (for example, PSU's or interviewers) as second-level units. Their design-based solution provides an adjustment to the  $\Gamma$  matrix that takes both clustering and stratification into account. Their method also allows for complex sampling designs on levels lower than the PSU through aggregation to the level of the PSU's.

In the discussion that follows we will adopt the design-based approach to variance estimation in the presence of clustering. The resulting variance estimator can be seen as obtained through the Taylor linearization method (Muthén & Satorra, 1995, 284).

## Non-normality

The asymptotic distribution of estimators of the mean of a variable are free of the distribution of that variable (e.g. Neyman, 1934). Such is not the case, however, for regression coefficients and other parameters of multivariate models. The variance and the form of the distribution of these parameters depends on the fourth-order moments of the observed variables (Satorra, 1989; Muthén & Satorra, 1995).

It has been shown that even under non-normality, normal-theory maximum likelihood estimates of structural equation models are still consistent (e.g. Satorra, 1989). Indeed, this result holds for the entire family of minimum distance-estimators discussed in the preceding section. Thus, non-normality does not cause asymptotic bias in the estimates.

When the  $y$  vector does not follow a multivariate normal distribution, this does affect the variance of structural equation model parameters. The  $\Gamma^*$  matrix of equation 12, which is a function purely of the observed variances and covariances, no longer provides a consistent estimate of the matrix of fourth-order moments. In this case the general  $\Gamma$  matrix of equation 10 must be used; or, in the case of complex samples,  $\Gamma^{(c)}$  of equation 11. It should be noted that the default behavior of all commonly used structural equation modeling software is to provide the variance estimators assuming normality.

Thus, the effect of non-normality is in general to change the matrix used as an estimate of  $\Gamma$  used in equation 9. Therefore non-normality, similarly to complex sampling, has a multiplicative effect on the variances and covariances of the parameter estimates.

## Measurement error and its estimation

The effect that measurement error has on regression coefficients is well-known (Fuller, 1987; Biemer & Trewin, 1997). For general structural equation models the effect will depend on the structure of the model, which can be deduced from the matrix of first derivatives of the population covariances with respect to the parameters. For regression with a single predictor, the regression coefficient will be biased downwards. In multiple regression the bias is not necessarily downwards. Bias in multiple regression coefficients can be upwards or downwards, depending on the correlations between the predictors and the relative amount of measurement error in each of them.

If the measurement error was correctly estimated, either in the model itself or in an earlier analysis, the estimates are consistent even under non-normality, adding no asymptotic bias component to the total error. This is because the measurement error has already been corrected for. As will be shown, however, the correction does add variance.

Structural equation models can be used to simultaneously estimate measurement errors and correct for them. In fact, the distinguishing characteristic of structural equation models is that they are a marriage of psychometric (factor analysis) and econometric (regression) models (Jöreskog, 1978).

In practice, however, simultaneous estimation of measurement and 'structural' parameters may be impractical or impossible.

The inclusion of both a measurement and structural part in the model simultaneously may cause a model to become prohibitively large. In the example discussed above the model size would not be exceedingly large as the four constructs are estimated from 11 indicator variables. This is close to the minimum needed of 8 indicators. On the other side of the extreme, educational and psychological scales may have hundreds of indicators each.

Often, moreover, repeated measures or validation data are not available in the same study as that used for the estimation of the structural model, although the measurement properties of the variables used have been estimated in other studies. In such cases simultaneous estimation is impossible, and the structural model must be corrected for measurement error using the estimates from previous studies. Another approach is the prediction of measurement error from meta-analyses of reliability based on characteristics of the question (Oberski, Saris, & Kuipers, 2004; Saris & Gallhofer, 2007a).

## Nonresponse, coverage, and survey mode

To the extent that other error components such as nonresponse, coverage, and mode bias the covariance matrix, the parameters  $\theta$  will be correspondingly biased.

The theoretical effect of nonresponse bias on the covariances was discussed by Groves and Couper (1998, chapter 2): it is a function of the nonresponse bias, nonresponse rate, and the difference in variances between respondents and nonrespondents. A similar result can be developed for coverage errors. The theory therefore suggests that nonresponse and coverage errors can in principle bias multivariate parameter estimates.

One can conclude from the results developed that for a bias to exist, there must be an interaction between variables correlated with nonresponse and the variables under study. Such might for example occur when, in a simple regression of social trust on fear of crime, there would be an interaction with living in a city or not, so that the relationship between fear of crime and trust were different for city dwellers. Since urbanicity is a commonly found correlate of nonresponse (Groves & Couper, 1998) this would imply a bias caused by nonresponse in the simple regression coefficient. Thus, for nonresponse bias to occur in a regression coefficient or other function of covariances mean, a 'third order' interaction must exist, whereas for nonresponse bias in means and totals a bivariate relationship or 'second order' interaction with participation correlates suffices.

The general focus in studies examining such effects in real surveys is on the estimation of means (De Leeuw & Van der Zouwen, 1988; Groves, 2002; Groves & Peytcheva, 2008). Due to this focus very few studies examine the extent of such biasing effects on multivariate statistics. An exception for nonresponse is Voogt (2004), who used official record data and found nonresponse bias in political variables' means to be high but did not find bias for logistic regression coefficients. Recently Révilla and Saris (forth), comparing a web survey with the face-to-face ESS, investigated possible mode effects and found no differences in the correlations between repeated measures and other variables.

The record of nonresponse, coverage, and mode effects on multivariate statistics is thus rather incomplete. Theoretically biases can exist, but more is required than for means and totals. Two studies could not find any biases, but a generalization cannot be made. It is worthy of note, however, that the study of such biases requires a special data collection design (e.g. Biemer, 2001), which is not available to us in the example we discuss. Therefore we are forced to ignore the possible biasing effects of nonresponse, coverage, and survey mode in our subsequent discussion.

	Distribution of $y$	Clustering/weighting	Measurement error	$\Gamma$	$\hat{\Sigma}_\eta$	$\Sigma_\Psi$
1	Normal	-	-	$\hat{\Gamma}^*$	$S$	0
2	Normal	-	Fixed	$\hat{\Gamma}^*$	$B^{-1}\Phi B^{-1}$	0
	Normal	-	Estimated	$\hat{\Gamma}^*$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$
	Normal	Yes	-	$\hat{\Gamma}^{(c)*}$	$S^{(c)}$	0
	Normal	Yes	Fixed	$\hat{\Gamma}^{(c)*}$	$B^{-1}\Phi B^{-1}$	0
	Normal	Yes	Estimated	$\hat{\Gamma}^{(c)*}$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$
	Non-normal	-	-	$\hat{\Gamma}$	$S$	0
3	Non-normal	-	Fixed	$\hat{\Gamma}$	$B^{-1}\Phi B^{-1}$	0
	Non-normal	-	Estimated	$\hat{\Gamma}$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$
	Non-normal	Yes	-	$\hat{\Gamma}^{(c)}$	$S^{(c)}$	0
4	Non-normal	Yes	Fixed	$\hat{\Gamma}^{(c)}$	$B^{-1}\Phi B^{-1}$	0
	Non-normal	Yes	Estimated	$\hat{\Gamma}^{(c)}$	$B^{-1}\Phi B^{-1}$	$\hat{\Sigma}_\Psi$

Table 3: Different error components and their effect on the choice of estimators of the parameters  $\theta$  and their variance. The effect of weighting is not shown separately because in our subsequent example simple random sampling was employed. However, the procedure for examining its effect is identical to that for clustering. The final variance is always obtained by equation 9.

## Decomposition of the variance of multivariate statistics

The variance of the parameter vector  $\theta$  was given in equation 9. Clustering, unequal sampling weights, and non-normality are all factors that affect the choice of the matrix  $\Gamma$  necessary for obtaining correct variance estimates. Sampling weights and measurement error also affect the necessary choice of a covariance matrix estimator.

Each row of table 3 yields a different estimator of the variance (or standard errors) of the parameter estimates of the model. A simple model, based on the observation that the effects of the conditions are in general multiplicative, is then to assume that each of the variance vectors under the different conditions is the result of a multiplication of the effects of non-normality, clustering, measurement error and a general scaling constant of the variance.

$$\text{var}_{\text{condition}} = vNCM, \quad (16)$$

where  $N$ ,  $C$ , and  $M$  are the multiplicative 'design' effects of non-normality, clustering, and measurement error, respectively.

These effects can be estimated by estimating the four numbered rows shown in table 3. We then report the square roots of the encountered effects (defts), since these are on the scale of the parameters rather than their square.

We now apply this decomposition to the example multivariate analysis discussed earlier, reporting standard errors under the various conditions as well as the square root of the interviewer clustering, measurement error, and non-normality effects.

	Estimate	$\sigma_1(\hat{\theta})$	$\sigma_2(\hat{\theta})$	$\sigma_3(\hat{\theta})$	$\sigma_4(\hat{\theta})$	deft <sub>measerr</sub>	deft <sub>non-normality</sub>	deft <sub>clustering</sub>
soctrust $\rightarrow$ poltrust	0.77	0.08	0.16	0.17	0.17	1.92	1.06	1.02
efficacy $\rightarrow$ poltrust	0.51	0.08	0.16	0.18	0.23	1.94	1.11	1.30
poltrust $\rightarrow$ soctrust	0.30	0.11	0.19	0.21	0.25	1.77	1.10	1.17
fearcrim $\rightarrow$ soctrust	-0.68	0.12	0.22	0.25	0.25	1.88	1.10	1.03
$\phi$ (poltrust, soctrust)	6.54	1.60	3.17	3.52	4.20	1.98	1.11	1.19
$\phi$ (efficacy)	1.64	0.06	0.10	0.10	0.11	1.71	0.99	1.11
$\phi$ (fearcrime)	1.71	0.06	0.11	0.12	0.14	1.78	1.07	1.16
cov(efficacy, fearcrime)	-0.60	0.06	0.10	0.08	0.09	1.73	0.77	1.10

Table 4: Parameter estimates, standard error estimates under various conditions, and square root design effects (deft) for the example analysis.

## Estimation of error components in the example

The analysis of the structural equation model of political and social trust presented earlier can be used to show how the effects of different survey error components can be estimated.

Table 4 shows the standard error estimates under different conditions. Four conditions are shown, corresponding to the numbered rows of table 3. From these standard error estimates the 'design effects' of measurement error, non-normality, and clustering can be estimated<sup>6</sup>. The square roots of these design effects (defts) are shown in the last three columns. These show the percentage increase in the standard error due to each factor. It can be seen that, in general, measurement error is the primary concern for the variance of the estimates, as it almost doubles the standard errors of the structural parameters of the model. The smallest measurement error deft is a 71% increase in the standard error for the variance of the independent variable efficacy.

The effect of interviewer clustering is less than the effect of measurement error but also considerable, with defts ranging between 1.03 for the standard error of the effect of 'fear of crime' on 'social trust' and 1.30 for the effect of 'political efficacy' on 'political trust'. The relative sizes of the defts of clustering may appear surprising considering equation 2; since the icc of social trust is rather large (0.25), in a simple regression we would expect the effect of social trust as a predictor to have the largest design effect. However, there are two important differences between the model of Scott and Holt (1982) leading to equation 2 and our model: the cluster sizes are not equal for all interviewers as shown in table 2, and we are not dealing with simple regression but with a complex structural equation model. This shows that approximations such as equation 2 cannot always be used in more complex situations.

Finally, it can be seen in table 4 that non-normality in general increases the variance

<sup>6</sup>We have designated the within-interviewer clustering effect an 'interviewer effect'. However, because the sample was not interpenetrated, some correlation between interviewer and region may exist.

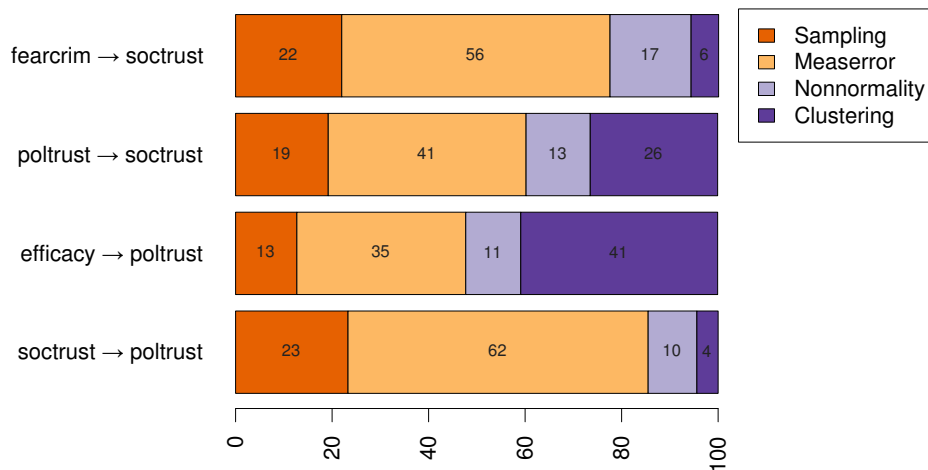


Figure 2. Percentage of variance of the regression coefficients of the model (see fig. 1) contributed by each error component.

of the estimates. This is not always the case in our sample, however, as the 'robustness effect' for the covariance between efficacy and fear of crime is smaller than unity. This most likely reflects the fact that the  $\Gamma$  matrix used is not the true population matrix, but a sample-based estimate. It may also reflect the fact that the second order sample moments may be negatively correlated. In general percentage increase in the standard errors due to non-normality is modest compared with the effects of interviewer clustering, and even more so when compared with the effects of measurement error.

The previous sections showed that the effects of measurement error, interviewer clustering, non-normality, and the sampling design in general are multiplicative in the variance of the estimates. For this reason the design effect or its square root is the more appropriate summary of their effects on the variance. However, for a given solution one can also calculate the percentage of variance due to each factor. Here a caveat should be added that such a measure is necessarily conditional on the sample, sample size, model, and parameter values encountered and cannot easily be generalized. However, it can be instructive for a given analysis to examine how much of the variance in the regression coefficients can be attributed to sampling, measurement error, non-normality, and interviewer clustering.

Figure 2 shows the amount of variance in the four structural (regression) coefficients of the model attributable to each of the four survey error sources studied here. Due to the simple random sampling scheme employed for the data collection, the factor "sampling" indicates purely the sample size. It can be seen that only a fifth of the total variance is attributable to sampling for all four parameters. It is also clear that measurement error is another important source of uncertainty about the parameter estimates. Interviewer clustering contributes in about equal parts with measurement error for two of the parameters, while not apparently playing any large role in the other two. The percentage of variance due to non-normality is slightly less than that simply due to sampling.

From the above discussion it is clear that sampling is by no means the only or even the most important factor contributing to the inferential uncertainty about the parameters in this example. It also shows that the error sources are far from equal and show a different

	$\hat{\theta}_{\text{correct}}$	$\hat{\theta}_{\text{naive}}$	$\text{bias}^2$	$\sigma^2(\hat{\theta}_{\text{naive}})$	$\sqrt{\text{MSE}}$	% MSE, bias	% MSE, var.
soctrust $\rightarrow$ poltrust	0.77	0.83	0.00	0.02	0.16	15%	85%
poltrust $\rightarrow$ soctrust	0.30	0.44	0.02	0.04	0.25	32%	68%
efficacy $\rightarrow$ poltrust	0.51	0.28	0.05	0.02	0.26	76%	23%
fearcrim $\rightarrow$ soctrust	-0.68	-0.31	0.13	0.01	0.38	89%	10%

Table 5: If the model is not corrected for measurement error the parameter estimates of interest will be biased. The Mean Square Error (MSE) of the naive (not corrected for measurement error) estimate then equals  $\text{bias}^2 + \sigma^2(\hat{\theta}_{\text{naive}})$ . The last two columns show the approximate percentage of mean square error in the naive estimates due to the bias and the variance, respectively. (Percentages not adding up to 100 are due to rounding errors.)

pattern from that typically found in the estimation of means and total. This finding cannot be generalized to other models, but does show that in an analysis of real data using a model found in the literature such differences can occur.

So far we have assumed that the practitioner obtains estimates of the measurement error variances and correct for them. If the correction is applied and assuming that the model is correct, the asymptotic bias is zero, even under non-normality and clustering. However, such corrections are not always applied. Therefore we will briefly show the effect that not correcting for measurement error has for this example.

Table 5 repeats, in the first column, the consistent estimates of the four regression coefficients while correcting for measurement error. The second column shows the estimates obtained by incorrectly assuming no measurement error. The square of the difference between these two is shown in the column labeled “ $\text{bias}^2$ ”. As discussed above, the mean square error of the naive estimates will equal the sum of the bias squared and the variance of the estimates. The last column shows the amount each of bias and variance contribute to the mean square error.

Without correction for measurement error the estimates of the model are biased. Therefore the root mean square error shown in the column  $\sqrt{\text{MSE}}$  in table 5 is a function of both this bias and the variance of the estimate. This root mean square error without correction for measurement error can be compared with the column labelled  $\sigma_4$  in table 4. This is so because the model in that table has been estimated with correction for measurement error so that the estimates are unbiased (asymptotically and under the null hypothesis). In that case the root mean square error of an estimate will equal the right-most standard error shown in table 4.

When thus comparing the root mean square errors with and without correction for measurement error, it is clear that the MSE of this model without correction is larger for all parameters than the MSE of the corrected model estimates. The bias caused by ignoring measurement error causes the MSE to exceed that of the unbiased estimates in this example.

Both from the point of view of obtaining unbiased estimates and that of minimizing the mean square error it is therefore necessary to correct for measurement error.

## Discussion and conclusion

This paper studied the effects of total survey error sources on multivariate statistics.

The first sections developed the structural equation modeling framework, which allows for the formulation of a wide range of common and specialized multivariate models; multiple regression, factor analysis, instrumental variables, multilevel models, and longitudinal models are among some of the wide variety of possible models that can be formulated within this framework. An example analysis with a model involving reciprocal effects, correlated errors, and measurement error demonstrated the use of structural equation models.

A review of existing studies showed that both theoretical and empirical considerations suggest such effects may differ greatly in relative size from their effects on the more commonly discussed estimation of means and totals.

It was discussed how each of the error sources (interviewer) clustering, unequal probability sampling, non-normality, and measurement error can be taken into account in structural equation models. It was shown exactly how each source influences the variance of the estimates of such models.

The relative effects of each error source on the estimates of our example analysis were then shown in terms of (root) 'design effects' (defts) and percentages. It was clear from this exercise that in the example given, estimated on real data from the 2008 European Social Survey, the effects of measurement error were the most pronounced, leading almost to a doubling of the standard errors relative to the variance the estimates would have had if there had been no measurement error. Clustering was another important factor, with non-normality leading to relatively smaller differences.

This result might lead one to think that it may not be worthwhile to correct for measurement error. However, the bias introduced by assuming no measurement error in the same analysis caused the mean square error to exceed that of the corrected model for all estimates. Therefore even if the goal is to obtain estimates that have the smallest mean square error, but that are not necessarily unbiased, the choice of preference should be the measurement error-corrected estimator.

One limitation of our example is that we have spoken of interviewer effects, while the interviewers were not randomly assigned to respondents (interpenetrated). Therefore it is possible that the clustering effects found were not (solely) due to the interviewer, but, for example, due to region. For a design allowing for the separation of such effects, see Bassi and Fabbri (1997). Another limitation is that in our example it was not possible to simultaneously estimate the effects of nonresponse, noncoverage/overcoverage, and survey mode. Again, a special study design is necessary for the study of such effects. The procedure presented in this paper, however, can easily be extended to encompass such effects.

A more fundamental caveat should be added about the use of the term "design effect". This term usually is taken to mean the variance of an estimator under the sampling scheme used, relative to the variance under simple random sampling. For measurement error, clustering, and non-normality, we have employed the same term, but the comparison is not with simple random sampling *per se* but with a design without measurement error, clustering, or non-normality. Thus, there is a strong analogy with the pure "design effect" but the two measures are not exactly the same.

We hope to have shown that it is possible to simultaneously estimate the effect on multivariate statistics of different survey error sources. We have given one example analysis.

The relative importance of different survey error sources in general, across different studies, remains a topic of considerable interest. The methods discussed in this paper could be applied to enable such a study.

## References

- Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement*. Wiley-Interscience.
- Bassi, F., & Fabbris, L. (1997). Estimators of nonsampling errors in interview-reinterview supervised surveys with interpenetrated assignments. *Survey Measurement and Process Quality*, 733–751.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.
- Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, 45(3), 289–308.
- Biemer, P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17(2), 295–320.
- Biemer, P., Groves, R., & Lyberg, L. (2004). *Measurement errors in surveys*. New York, NY: Wiley.
- Biemer, P., & Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, & C. D. et al. (Eds.), *Survey measurement and process quality* (pp. 603–631). New York, NY: John Wiley & Sons, Inc.
- Boker, S., Neale, M., Maes, H., Metah, P., Kenny, S., Bates, T., et al. (2010). Openmx: The openmx statistical modeling package [Computer software manual]. Available from <http://openmx.psyc.virginia.edu> (R package version 0.2.10-1172)
- Cochran, W. G. (1977). *Sampling technique*. New York: John Wiley & Sons.
- De Leeuw, E., & Van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. *Telephone survey methodology*, 283–299.
- Fuller, W. A. (1987). *Measurement error models*. New York: Wiley.
- Groves, R. (1989). *Survey errors and survey costs*. Wiley-Interscience.
- Groves, R. (2002). *Survey nonresponse*. Wiley-Interscience.
- Groves, R., & Couper, M. (1998). *Nonresponse in household interview surveys*. Wiley-Interscience.
- Groves, R., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72(2), 167.
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43, 443–477.
- Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2), 239.
- Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (2007). *Measuring attitudes cross-nationally: Lessons from the European social survey*. SAGE.
- Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1), 1–37.
- Lyberg, L. (1997). *Survey measurement and process quality*. New York, NY: Wiley-Interscience.
- Magnus, J. R., & Neudecker, H. (2002). *Matrix differential calculus with applications in statistics and econometrics, third edition*. John Wiley, Chichester.
- McArdle, J., & McDonald, R. (1984). Some algebraic properties of the Reticular Action Model for moment structures. *British Journal of Mathematical and Statistical Psychology*, 37(2), 234–251.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological methodology*, 25, 267–316.

- Neudecker, H., & Satorra, A. (1991). Linear structural relations: Gradient and Hessian of the fitting function. *Statistics and Probability Letters*, 11, 57–61.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558–625. Available from <http://www.jstor.org/stable/2342192>
- Oberski, D., Saris, W. E., & Kuipers, S. (2004). *SQP: survey quality predictor*. Available from <http://www.sqp.nl/>
- Putnam, R. (2001). *Bowling alone: The collapse and revival of American community*. Simon and Schuster.
- R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from <http://www.R-project.org> (ISBN 3-900051-07-0)
- Raykov, T. (2009). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 195–212.
- Révilla, M., & Saris, W. (frth). A comparison of surveys using different modes of data collection: European Social Survey versus LISS panel. *RECSM Universitat Pompeu Fabra working papers*. (<http://upf.edu/survey/>)
- Saris, W. E., & Gallhofer, I. (2007b). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, 1.
- Saris, W. E., & Gallhofer, I. N. (2007a). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience.
- Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, 54(1), 131–151.
- Scott, A. J., & Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77(380), 848–854.
- Voogt, R. (2004). *I'm not interested: Nonresponse bias, response bias and stimulus effects in election research*. Amsterdam: University of Amsterdam.
- Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, 74(3), 193–212.

## Questions used in the example analysis

Here we show the questions used to measure the four constructs analyzed above. All questions come from the European Social Survey Round 4 and were translated into each country's respective language (in our case Danish). The first item of political efficacy and the last two items of fear of crime were reverse-coded so higher scores indicated more efficacy or more fear, respectively.

### Social Trust

Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.

You can't	Most
be too	people
	can be



an institution at all, and 10 means you have complete trust. Firstly...

		No trust										Complete		
		at all										trust		
B4	...[country]'s parliament?	00	01	02	03	04	05	06	07	08	09	10		
B5	...the legal system?	00	01	02	03	04	05	06	07	08	09	10		
B6	...the police?	00	01	02	03	04	05	06	07	08	09	10		

### Fear of crime

How safe do you - or would you - feel walking alone in this area after dark? Do - or would - you feel... READ OUT...

...very safe,	1
safe,	2
unsafe,	3
or, very unsafe?	4

How often, if at all, do you worry about your home being burgled? Please choose your answer from this card.

All or most of the time	1
Some of the time	2
Just occasionally	3
Never	4

How often, if at all, do you worry about becoming a victim of violent crime? Please choose your answer from this card.

All or most of the time	1
Some of the time	2
Just occasionally	3
Never	4

### EQS input for reliability analysis

/TITLE

Factor analysis of construct indicators ESS round 4 Denmark

/SPECIFICATIONS

DATA='denmark.ess';  
 VARIABLES=46; CASES=1610; GROUPS=1;  
 METHOD=ML; ANALYSIS=COV; MATRIX=RAW;

/LABELS

...

/EQUATIONS

V2 = 1F1 + E2;  
V3 = \*F1 + E3;  
V4 = \*F1 + E4;  
V6 = 1F2 + E6;  
V7 = \*F2 + E7;  
V8 = 1F3 + E8;  
V9 = \*F3 + E9;  
V10 = \*F3 + E10;  
V17 = 1F4 + E17;  
V18 = \*F4 + E18;  
V19 = \*F4 + E19;  
  
F5 = V2 + V3 + V4;  
F6 = V6 + V7;  
F7 = V8 + V9 + V10;  
F8 = V17 + V18 + V19;

/VARIANCES

F1 = \*;  
F2 = \*;  
F3 = \*;  
F4 = \*;  
E2 = \*;  
E3 = \*;  
E4 = \*;  
E6 = \*;  
E7 = \*;  
E8 = \*;  
E9 = \*;  
E10 = \*;  
E17 = \*;  
E18 = \*;  
E19 = \*;

/COVARIANCES

F1,F2 = \*;  
F1,F3 = \*;  
F2,F3 = \*;  
F1,F4 = \*;  
F2,F4 = \*;  
F3,F4 = \*;

/PRINT

TABLE=EQUATION;  
COVARIANCE=YES;  
CORRELATION=YES;

```
/SIMULATION
  bootstrap = 1610;
  replication = 2000;
  seed = 123456789;

/OUTPUT
  parameters;
  standard deviation;

/END
```

All replications converged. The output from this analysis was read in using R 2.11.0, and the reliability calculated as

$$\frac{(\sum \lambda)^2 \text{var}(F_i)}{\text{var}(F_{i+4})}; i = 1, 2, 3, 4,$$

in each of the 2000 replications. The averages and standard deviations across replications are shown in table 2.