```
    12/05/2019 13:25    SYRACUSE UNIVERSITY ANNEX PRODUCTION SYSTEM   PG     1
    RPT# 00135151 pbmovsov                                            cpyrept.

                                  EDD RETRIEVAL RECEIPT


            Order: 021380     Document Nbr: 1166361
              For: EDD
           Copied: 12/05/2019
          Shipped: 12/05/2019
       Deliver To: Internal ILLIAD user
     Item BarCode: 32911045066115
       Item Title: Sociological methods & research.
      Item Author:
Item Call Number: M1 .S687
    Item Vol/Part:


    Article Title: Robustness studies in covariance structure modeli
   Article Author: Hoogland & Boomsma
     Art Vol/Part: 26
    Article Issue:
         Beg Page: 329     End Page: 367          Total Pages: 39
       Other Info: 1166361
            Notes: 3|Exempt





            TOTAL COUNT: 1
```

In covariance structure modeling, several estimation methods are available. The robustness of an estimator against specific violations of assumptions can be determined empirically by means of a Monte Carlo study. Many such studies in covariance structure analysis have been published, but the conclusions frequently seem to contradict each other. An overview of robustness studies in covariance structure analysis is given, and an attempt is made to generalize findings. Robustness studies are described and distinguished from each other systematically by means of certain characteristics. These characteristics serve as explanatory variables in a meta-analysis concerning the behavior of parameter estimators, standard error estimators, and goodness-of-fit statistics when the model is correctly specified.

Robustness Studies in Covariance Structure Modeling

An Overview and a Meta-Analysis

JEFFREY J. HOOGLAND
ANNE BOOMSMA
University of Groningen

1. INTRODUCTION

Structural equation modeling (SEM) is focused on directed relationships between empirical phenomena, which are often represented by unobserved, latent variables (hypothetical constructs). Observed, measured variables can be used as indicators of these latent variables, thus defining the measurement part of a structural equation model, which can deal explicitly with measurement errors. Structural equation models represent the underlying structure among variables in terms of population covariances, which can be expressed as functions of unknown population parameters. The main purpose of covariance structure analysis (CSA) is to estimate the population parameters in these models using a sample of covariances based on N observations.

SOCIOLOGICAL METHODS & RESEARCH, Vol. 26 No. 3, February 1998 329-367

Several assumptions are implicitly made in SEM, particularly with respect to the estimation method (see Section 2) used to obtain estimates of model parameters, standard errors, and goodness-of-fit statistics. In practice, many of these assumptions turn out to be too restrictive. The effect of violation of assumptions depends on the robustness of an estimation method given a postulated covariance structure model and the distributional properties of the sample data. A statistical procedure is called robust if its performance is relatively insensitive to departures from the underlying assumptions used to derive that procedure (Box 1953:318).

Robustness properties of estimators can be obtained empirically by means of a Monte Carlo study. In a specific Monte Carlo design, several aspects, such as the postulated model and the distribution of the observed variables, must be specified a priori. A drawback is that the Monte Carlo results are conditional on that design, and generalizations are therefore only justified when there is a clear trend. The number of robustness studies in SEM is quite impressive, but the robustness issues investigated and the approach followed differ substantially (see Section 3). More important, the conclusions from these studies are often contradictory and in general hard to summarize.

This article gives an overview of robustness studies concerning covariance structure models. One of the main purposes of this overview is the generalization of robustness properties of estimators of parameters, estimators of standard errors, and goodness-of-fit statistics. Frequently, generalizations are possible by identifying causes for contradicting conclusions across robustness studies.

A collection of robustness studies will be described and systematically distinguished from each other by means of characteristics regarding the model, the simulated data, the estimation methods, and the obtained results. The robustness studies investigated and characteristics that are used to describe them systematically are specified in Section 4.

One of the things that can cause differences in conclusions across robustness studies are the assessment criteria used. Ideally, these criteria should have no influence on the conclusions of a Monte Carlo study. Therefore, they are redefined in Section 6 and applied to the reported results of any robustness study whenever possible. A meta-analysis on robustness results, as described in Section 6, serves to indicate which estimation methods can be used, or should be preferred, to obtain precise and reliable estimates of parameters, estimates of standard errors, and goodness-of-fit statistics. The meta-analysis is mainly verbal and judgmental because the robustness results are influenced by many characteristics.

A systematic comparison of robustness studies may establish which aspects of estimation in covariance structure modeling have had too little, or too much, attention. This leads to guidelines for future research in this area. In Section 7, the conclusions of the overview are given and topics for future research are discussed.

2. *ESTIMATION METHODS*

The fundamental hypothesis in covariance structure modeling is

$$\Sigma = \Sigma(\boldsymbol{\theta}), \tag{1}$$

where $\Sigma(\boldsymbol{\theta})$ ($k \times k$) is the population covariance matrix of k observed variables written as a function of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}$(t × 1) is a vector of the model parameters. The sample estimator of Σ in a sample of size N is

$$\mathbf{S} = \mathbf{Z}'\mathbf{Z}/(N-1), \tag{2}$$

where $\mathbf{Z}$ is an ($N \times k$) matrix of deviation (from the means) scores of the observed variables. Given a specified model $\Sigma(\boldsymbol{\theta})$, the unknown parameters of $\boldsymbol{\theta}$ are estimated in such a way that the discrepancy between the implied covariance matrix $\hat{\Sigma} = \Sigma(\hat{\boldsymbol{\theta}})$ and the sample covariance matrix $\mathbf{S}$ is as small as possible given some criteria, where $\hat{\boldsymbol{\theta}}$ is the vector of parameter estimates. A discrepancy function $F(\mathbf{S}, \Sigma(\boldsymbol{\theta}))$ is needed to quantify the fit of a model to the sample data. Four estimation methods and their discrepancy functions are described now.

- Maximum Likelihood (ML)

The most widely used fitting function is the ML function, which can be derived by assuming that the observed variables **z** ($k \times 1$) are multinormally distributed. ML parameter estimates are obtained by minimizing the discrepancy function

$$F_{ML} = \log |\Sigma(\theta)| + \mathrm{tr}[\mathbf{S}\Sigma^{-1}(\theta)] - \log |\mathbf{S}| - k. \tag{3}$$

- Generalized Least Squares (GLS)

GLS-estimates of model parameters are obtained by minimizing the discrepancy function

$$F_{GLS} = \mathrm{tr}[\{(\mathbf{S} - \Sigma(\theta))\, \mathbf{V}^{-1}\}^2]/2, \tag{4}$$

where $\mathbf{V}^{-1}$ is a positive definite weight matrix for the residual matrix $(\mathbf{S} - \Sigma(\theta))$.

A sufficient condition for the underlying distributional assumption to hold is that the observed variables do not have excessive kurtosis; that is, the kurtosis of each observed variable equals zero,[1] the kurtosis of a normal distribution (Browne 1974; Bollen 1989:114).

- Asymptotically Distribution Free (ADF)

The distributional assumptions of the ADF estimator are valid under very general conditions (Browne 1984). Its discrepancy function can be written as

$$F_{ADF} = [\mathrm{s} - \sigma(\theta)]'\mathbf{W}^{-1}[\mathrm{s} - \sigma(\theta)], \tag{5}$$

where **s** is a vector of the $k(k + 1)/2$ nonduplicated elements of **S**, $\sigma(\theta)$ is the corresponding vector of $\Sigma(\theta)$, and $\mathbf{W}^{-1}$ is an optimal weight matrix. The matrix $\mathbf{W}^{-1}$ must be optimal in the sense that **W** has to be a consistent estimator of the matrix *N*U, where the elements of U are the asymptotic covariances between s_{ij} and s_{gh} for each i, j, g, and h. In general,

$$U_{ij,gh} = ACOV(s_{ij},s_{gh}) = N^{-1}(\sigma_{ijgh} - \sigma_{ij}\sigma_{gh}), \tag{6}$$

where σ_{ijgh} is the fourth-order multivariate moment of z_i, z_j, z_g, and z_h around their mean, and σ_{ij} is the population covariance of z_i and z_j around their mean.

In practice, the elements σ_{ijgh} in (6) are estimated by

$$s^*_{ijgh} = \frac{1}{N}\sum_{p=1}^{N} (z_{pi} - \bar{z}_i)\,(z_{pj} - \bar{z}_j)\,(z_{pg} - \bar{z}_g)\,(z_{ph} - \bar{z}_h), \tag{7}$$

and the elements σ_{ij} are estimated by

$$s^*_{ij} = \frac{1}{N}\sum_{p=1}^{N} (z_{pi} - \bar{z}_i)\,(z_{pj} - \bar{z}_j), \tag{8}$$

where $\bar{z}_i$ is the mean of z_{pi}.

- Elliptical (E)

A special case of F_{ADF} follows when the observed variables z are assumed to have a multivariate elliptical distribution (Browne 1984:74-5). Elliptical distributions have zero skewness but can have a homogeneous kurtosis κ that may deviate from the normal kurtosis. In that case,

$$\sigma_{ijkl} = (\kappa + 1)(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}), \tag{9}$$

where $\kappa = \kappa_i = (\sigma_{iiii} / 3\sigma_{ii}^2) - 1$, i = 1, 2, . . . , k, is the homogeneous kurtosis.

Elliptical parameter estimates are obtained by minimizing an elliptical discrepancy function

$$F_E = [2(\kappa + 1)]^{-1}\mathrm{tr}[(\mathbf{S} - \Sigma(\theta))\mathbf{V}^{-1}]^2 - \rho[\mathrm{tr}(\mathbf{S} - \Sigma(\theta))\mathbf{V}^{-1}]^2, \tag{10}$$

where $\rho = \kappa / [4(\kappa + 1)^2 + 2k\kappa(\kappa + 1)]$. There are several estimators available for the kurtosis parameter κ (Harlow 1985), and several possible choices for **V** in (10). When $\Sigma(\hat{\theta})$ is used for **V**, the estimation method is called an Elliptical Reweighted Least Squares (ERLS) estimator (Bentler 1995).

When the parameters are estimated by some estimation method, an important question is whether the model fits the data well. The most commonly used test of overall model fit is the chi-square goodness-of-fit test. Consider the discrepancy function F_{ADF} as defined by (5). The chi-square test statistic is easily computed as $T_{ADF} = (N - 1)F_{ADF}$ (Browne 1982:97). Asymptotically T_{ADF} has a χ^2 distribution if the following assumptions hold:

1. the null hypothesis H_0: $\Sigma = \Sigma(\theta)$ holds exactly,
2. θ is identified,
3. a covariance matrix **S** is analyzed,
4. **W** is optimal.

When assumptions 1, 2, and 3 hold and the distributional assumptions regarding the observed variables are satisfied, any of the other discrepancy functions multiplied by $(N - 1)$ are also asymptotically chi-square distributed.

When the distributional assumptions concerning the estimation method are satisfied, the standard errors of the parameter estimates can be obtained from the asymptotic covariance matrix of $\hat{\theta}$

$$ACOV(\hat{\theta}) = \frac{1}{N-1} (\Delta' \mathbf{U}^{-1} \Delta)^{-1}, \tag{11}$$

where $\Delta = (\partial\sigma(\theta)/\partial\theta')|_{\theta=\hat{\theta}}$ is the matrix of partial derivatives of the model with respect to the parameters, and **U** is the matrix that consists of the elements defined by (6). By means of estimators of the matrices in (11), which are consistent when the distributional assumptions for the specific estimation method hold, standard error estimates are obtained.

3. ROBUSTNESS QUESTIONS

The seriousness of a departure from model assumptions depends on the robustness of an estimation method against such departures, conditional on the model under study. The issues that are considered below are robustness against distributional violations, small sample size, analysis of correlation rather than covariance matrices, model misspecification, and nonlinear structural equations.

- Distributional Violations

When the observed variables have excessive kurtosis, the ML, GLS, and ERLS estimates of the standard errors and the associated chi-square statistic may be incorrect (Bentler and Dudgeon 1996). A possible solution is to transform the observed variables (Meijerink 1995).

Another possibility is to adjust an estimator of standard errors or a chi-square statistic for excessive kurtosis. The following adjustments can be made after the model parameters have been estimated.

(a) Browne (1984:76) gives "corrections for kurtosis" for the ML and GLS chi-square statistic and the asymptotic covariance matrix.
(b) A so-called robust sample covariance matrix, being robust against distributional misspecification, is defined by Browne (1984:67).
(c) Satorra and Bentler (1988) proposed a correction to the standard errors and a scaled test statistic.

- Small Sample Size

A small sample size may cause problems because the statistical properties of estimators of parameters, asymptotic covariances, and test statistics are asymptotic properties. For instance, the estimators of the parameters discussed in Section 2 are asymptotically unbiased under very general conditions when (1) holds, but the estimators can be biased for finite sample sizes. The ADF estimation method especially may give problems because fourth-order moments have to be estimated. When the sample size is small, this results in unstable estimates of the elements of $\mathbf{W}^{-1}$.

Given that $\Sigma(\theta)$ holds, the larger the sample, the better the performance of the specific estimation method is expected to be because **S** converges to Σ as N grows large. When the distributional assumptions are satisfied, the estimates of the standard errors should be better for large sample size because the matrices in (11) will be estimated more accurately. The chi-square statistic concerning a specific estimation procedure is merely asymptotically χ^2 distributed. For finite sample size, the χ^2 distribution is therefore an approximation of the true distribution, which can be expected to be worse the smaller the sample size.

- Analysis of Correlation Rather Than Covariance Matrices

In practice, the scales of the observed variables are often arbitrary. Therefore, in many applications, there is a tendency to use a correlation matrix **R** instead of a covariance matrix **S**. However, this is not an adequate procedure for all models. The use of **R** may even alter the model under study (Cudeck 1989).

When a correlation matrix **R** is analyzed, two concepts play an important role: scale invariance of a model and scale freeness of a parameter (Browne 1982). When a model is scale invariant, parameter estimates and the chi-square statistic are unaffected when a correlation matrix is analyzed.[2] Estimated standard errors will only be correct for scale-invariant models when the associated model parameters are scale free. Browne (1982:94) proposed corrections for the standard errors when correlations are analyzed. These corrections are available in RAMONA (Browne, Mels, and Coward 1994) and SEPATH (Steiger 1995).

- Model Misspecification

Even if all available theoretical information is implemented in specifying a model, the common experience is that the model is still misspecified. Herting and Costner (1985) describe several types of possible specification errors. Jöreskog (1993) gives a general strategy to find a suitable model (see Hayduk 1996, chap. 2, for a discussion).

If the model is poorly specified, nonconvergence or improper solutions may occur. The estimates of the path coefficients will be biased when there is a misspecification in the structural part of a model. Moreover, because the estimation methods of Section 2 are full information methods, which means that all equations are estimated simultaneously, all model parameter estimates may be biased then, including parameters concerning the measurement part of a model. A possible solution to this problem is the use of a limited information method such as the two-stage least squares (2SLS) method (Bollen 1996). When this estimation method is used, a misspecification in one part of the model does not necessarily affect other parts of the model.

- Nonlinear Structural Equations

It is conceivable that there are nonlinear relationships between latent variables. In that case, the general linear structural equation model, as defined by the LISREL model (Jöreskog 1973) or the Bentler and Weeks model (Bentler and Weeks 1980), does not hold. Examples of nonlinear relationships are quadratic and interaction terms. It is not possible to transform latent variables to make relationships approximately linear as can be done with observed variables. However, in principle, methods are available that can handle quadratic and interaction terms for latent variables, for example, the product indicators technique of Kenny and Judd (1984). Jöreskog and Yang (1996) and Yang (1997) present extensive studies of the Kenny and Judd model. Bollen (1995), Ping (1995), and Klein et al. (1997) have suggested alternative approaches.

4. COLLECTION OF ROBUSTNESS STUDIES

Table 1 gives an overview of a collection of robustness studies in SEM that investigated robustness issues discussed in Section 3 by means of Monte Carlo methods.[3] The effect of sample size is covered when at least two different sample sizes are studied and the effect of analysis of **R** when both the sample covariance matrix **S** and the sample correlation matrix **R** are used to estimate the model parameters.

It can be concluded from Table 1 that the effects of nonlinearity and analysis of **R** have received relatively little attention.

TABLE 1: Overview of Robustness Studies

		Model	*Robustness Against*				
Author(s)	*Year*	*Number*	*SS*	*DV*	*MM*	*R*	*NL*
Anderson and Gerbing	1984	1	X				
Babakus, Ferguson, and Jöreskog	1987	2-3	X	X			
Baldwin	1986		X		X		
Bearden, Sharma, and Teel	1982	4-5	X				
Benson and Fleishman	1994	6-9	X	X			
Boomsma	1983	10-28	X	X		X	
Brown	1990	29-30		X			
Browne	1984	31-32		X			
Chou, Bentler, and Satorra	1991	33-34	X	X			
Curran, West, and Finch	1996	35	X	X	X		
Dolan	1994	36	X	X			
Ethington	1987	37		X			
Gallini and Mandeville	1984		X		X		
Gerbing and Anderson	1985	38	X				
Harlow	1985	39-40	X	X			
Harlow, Chou, and Bentler	1986	41-42	X	X			
Henly	1993	43	X	X			
Hu, Bentler, and Kano	1992	44-46	X	X			
Jaccard and Wan	1995		X	X			X
Kaplan	1989		X		X		
Klein et al.	1997			X			X
Lance, Cornwell, and Mulaik	1988				X		
Lee, Poon, and Bentler	1995	47	X	X			
Meijer and Mooijaart	1992	48	X	X			
Muthén and Kaplan	1985	49		X			
Muthén and Kaplan	1992	50-53	X	X			
Ping	1995		X	X			X
Potthast	1993	54-57	X	X			
Reddy	1992		X		X		
Satorra and Bentler	1988	58		X			
Sharma, Durvasula, and Dillon	1989	59	X	X			
Tanaka	1984	60	X	X			
Yang	1997		X	X			X
Yung and Bentler	1994	61-62	X	X			

NOTE: The numbers in the third column correspond to a model study. The last five columns indicate whether the effect of sample size (SS), distributional violations (DV), model misspecification (MM), analysis of R (R), or nonlinearity (NL) are investigated.

5. CHARACTERISTICS OF ROBUSTNESS STUDIES

The robustness studies, and the models investigated within each study, that can be used for the meta-analysis are described systematically by means of certain characteristics. The characteristics can be categorized as model, data, estimation, simulation, and research summary characteristics. Each aspect is discussed briefly now.

MODEL CHARACTERISTICS

Covariance structure models can be subdivided in pure measurement models and structural models (with directed relationships between indicators and/or between latent variables). In a robustness study, the population model is often a Confirmatory Factor Analysis (CFA) model. A CFA model can be described by the number of factors, the number of indicators per factor, the correlations between the factors, and the factor loadings.

Important model characteristics are (a) whether the postulated model holds in the population; (b) the degrees of freedom of the model, which has a direct influence on the distribution of a goodness-of-fit statistic; (c) the number of parameters to be estimated, which is an indicator of the model complexity; and (d) whether models are invariant under a constant scaling factor (ICSF). A covariance structure model is ICSF when for each α there exists a θ^* such that $\Sigma(\theta^*) = \alpha^2\Sigma(\theta)$ (Browne 1982).

Table 2 contains characteristics of CFA models, from which it can be observed that most models are ICSF, the number of factors in a model is at most 4, the number of observed variables ranges from 4 to 16, the number of parameters to be estimated is at most 33, and the degrees of freedom of a model is always less than 100.

Table 3 contains characteristics of structural models. It can be seen that only a few structural models have been investigated, with at most 48 degrees of freedom.

DATA CHARACTERISTICS

In a robustness study, a distributional condition is mostly represented by the univariate skewness and/or kurtosis for each observed variable. We consider two distributional conditions to be comparable when they can be given the same qualitative description—for instance, two distributional conditions of continuous variables both with zero homogeneous skewness and positive heterogeneous kurtosis. A dis-

TABLE 2: Characteristics of CFA Models

Model	*Exact*				*Factor Correlations*		*Factor Loadings*			
Number	*Fit*	*ICSF*	f	k	*Minimum*	*Maximum*	*Minimum*	*Maximum*	t	df
6	no	yes	1	4			.38	.5	8	2
7	no	yes	1	4			.74	.8	8	2
8	no	yes	1	8			.38	.5	16	20
9	no	yes	1	8			.74	.8	16	20
15, 27	yes	yes	2	8	.3	.3	.6	.8	17	19
16	yes	yes	1	6			.3	.7	12	9
17	yes	yes	2	6	.0	.0	.4	.6	12	9
18	yes	yes	2	6	.0	.0	.6	.8	12	9
19	yes	yes	2	6	.0	.0	.8	.9	12	9
20	yes	yes	2	6	.3	.3	.4	.6	13	8
21	yes	yes	2	6	.3	.3	.6	.8	13	8
22	yes	yes	2	6	.3	.3	.8	.9	13	8
23	yes	yes	2	8	.0	.0	.4	.6	16	20
24	yes	yes	2	8	.0	.0	.6	.8	16	20
25	yes	yes	2	8	.0	.0	.8	.9	16	20
26, 29	yes	yes	2	8	.3	.3	.4	.6	17	19
28, 30	yes	yes	2	8	.3	.3	.8	.9	17	19
31	yes	yes	1	8	.5	.5	1.0	1.0	2	34
32	yes	yes	1	8			.71	.71	16	20
33, 39	yes	yes	2	6	.24	.24	.37	.9	13	8
34, 40	yes	no	2	6	.24	.24	.37*	.9*	7	14
36	latent	yes	1	8			.7	.9	16	20
41	yes	yes	4	12	.25	.25	.4	.9	30	48
42	yes	no	4	12	.25	.25	.4*	.9*	18	60
44, 61	yes	yes	3	15	.3	.5	.7	.8	33	87
45	yes	no	3	15	.3*	.5*	.7	.8	27	93
46, 62[a]	yes	yes	3	15	.3	.5	.7	.8	33	87
48	yes	yes	1	4			.6	.8	8	2
49	yes	yes	1	4			.7	.7	8	2
50	latent	yes	2	6			.7	.7	13	8
51	latent	yes	3	9			.7	.7	21	24
52	latent	yes	3	12			.7	.7	27	51
53	latent	yes	3	15			.7	.7	33	87
54	latent	yes	1	4	.3	.3	.71	.71	4	2
55	latent	yes	2	8	.3	.3	.71	.71	9	19
56	latent	yes	3	12	.3	.3	.71	.71	15	51
57	latent	yes	4	16	.3	.3	.71	.71	22	98
58	yes	no	1	4					7	3
59	yes	yes	2	8	.4	.4	.71	.71	5	31
60	no	yes	2	6	.28	.28			13	8

NOTE: The second column indicates whether the model fits exactly in the population; the term *latent* means that the model fits the latent variables instead of the observed variables. Subsequent columns indicate whether a model is ICSF (invariant under a constant scaling factor), number of factors (*f*), number of observed variables (*k*), range of factor correlations, range of factor loadings, number of parameters to be estimated (*t*), and degrees of freedom (*df*) of the model. A blank means that the specific entry is not applicable or not available. An asterisk means that the specific number is fixed.

a. With dependent latent variables.

TABLE 3: Characteristics of Structural Models

Model Number	*Exact Fit*	*ICSF*	f	k	t	df
4	yes	yes	2	6	13	8
5	yes	yes	4	12	30	48
10, 13	yes	yes	3	6	17	4
11, 14	yes	no	2	10	17	17
37	yes	yes	5	10	30	25

tributional condition is considered to be extreme when the observed variables have the largest absolute values of skewness and kurtosis among conditions with the same qualitative description. These *extreme* distributional conditions play an important role in Section 6.

Furthermore, it is shown to be useful to distinguish the following *distributional characteristics* for the description of distributional conditions: normal, skewed, platykurtic, leptokurtic, both leptokurtic and skewed, both platykurtic and skewed, categorical, and categorical in combination with another aspect of nonnormality.[4]

A summary of the following data characteristics of robustness studies was made: the number of categories of discrete variables, the sample sizes, whether a correlation or a covariance matrix was analyzed, and the minimum, mean, and maximum skewness and kurtosis of the observed variables for each distributional characteristic.[5] In most robustness studies, a covariance matrix was analyzed with the observed variables having a continuous distribution. The absolute value of the mean skewness was at most 3.0; the mean kurtosis values ranged from –1.3 to 21.0. It should be noted, however, that most skewnesses and kurtoses being reported were the levels aimed at, not the levels that were actually obtained in, the data generation process.

ESTIMATION CHARACTERISTICS

The estimation methods that are studied are an important characteristic of a robustness study.[6] The ML method is investigated most often; GLS, ERLS, and ADF methods are also studied frequently. The GLS, ERLS, and ADF estimators differ across studies depending on the specific estimator used for $\mathbf{V}$ in (4), κ and $\mathbf{V}$ in (10), and $\mathbf{W}$ in (5).

SIMULATION CHARACTERISTICS

The number of replications, denoted as *NR*, is an important simulation characteristic. The larger *NR*, the better the sampling distributions of the statistics of interest can be approximated by an empirical, simulated sampling distribution. This is important because the behavior of a statistic is assessed by means of characteristics of such empirical sampling distributions.

For specific replications, an estimation method may not converge. Exclusion of nonconvergent replications could imply that the results are biased. When a replication leads to convergence, but certain parameters have improper solutions, it must be decided whether that replication is retained or not.[7] Regretfully, most robustness studies did not report whether improper solutions occurred, let alone whether they were either included or excluded.

Vale and Maurelli (1983) developed a method to generate multivariate nonnormal data given the population means, covariance matrix, and the skewness and the kurtosis of the univariate distributions of the observed variables. Their method is often used in robustness studies when the first four moments of a distribution are the only moments of interest. Because data generation methods may differ in their performance, it is important to keep track of the method being used and its performance—for example, whether the intended levels of skewness and kurtosis correspond to the observed levels of skewness and kurtosis.

RESEARCH SUMMARY CHARACTERISTICS

Research summary characteristics indicate which results are obtained and how the quality of these results is assessed. In contrast to the characteristics mentioned so far, they do not influence the results directly, but their choice might affect the conclusions of a robustness study. A *research summary statistic*, abbreviated as r.s.s., serves to summarize the results, with the aim to assess the quality of specific estimators.

The following research summary statistics are frequently encountered in robustness studies:

- percentage of replications that lead to nonconvergence;
- percentage of convergent replications that give improper solutions;
- bias of parameter estimates;
- standard deviation of parameter estimates;
- bias of standard error estimates;
- rejection rate, mean, and standard deviation of a chi-square statistic;
- *p*-value of the Kolmogorov-Smirnoff test for a χ^2 distribution.

With respect to the bias of standard error estimates, one aspect needs further attention. This bias can only be computed when the population values of the standard errors have been estimated. However, there are different methods to estimate these population values: (a) Compute the so-called theoretical standard errors, given that the model holds exactly. This is done by estimating (11) by means of the population values for the parameters, which is only correct when the distributional assumptions for the specific estimation method are satisfied. (b) Use the *empirical standard errors*. The empirical standard error of a specific parameter θ_i is defined as

$$\left[\sum_{j=1}^{NR}(\hat{\theta}_{ij} - \bar{\hat{\theta}}_i)^2/(NR-1)\right]^{½}, \qquad (12)$$

where $\bar{\hat{\theta}}_i$ is the mean of the estimates for parameter i across NR replications.

The choice for the method by which the population standard errors are estimated is not irrelevant. Incorrect estimates of the theoretical standard errors might be obtained when the underlying estimation method is subject to distributional violations. Statistic (12) is not derived under distributional assumptions, but the quality of these empirical standard errors is highly dependent on the number of replications.

6. *META-ANALYSIS ON ROBUSTNESS RESULTS*

In this section, Monte Carlo results based on different levels of skewness and kurtosis of the observed variables, different sample sizes, and different models are compared as objectively as possible. The results concern the performance of ML, GLS, ERLS, and ADF estimation methods when the model is correctly specified. In Section 6.1, the assessment criteria being used are defined. In Section 6.2, robustness results are presented concisely. In Section 6.3, estimators are ranked according to their relative performance. In Section 6.4, the findings of the meta-analysis are discussed.

6.1. CRITERIA FOR ASSESSING PERFORMANCE OF ESTIMATORS

Formal procedures are developed to compare results concerning the performance of estimators across robustness studies systematically. These procedures are based on the following line of reasoning.

Imagine a robustness study that investigated the behavior of an estimator for a number of finite sample sizes, several distributional conditions, and a specific model. We use the term *model study* to refer to a robustness study for one specific model. The performance of an estimator in a specific cell of the Monte Carlo design is quantified by means of a research summary statistic (see Section 5). To simplify the assessment of an estimator, we define criteria that dichotomize the performance of an estimator as *acceptable* or *unacceptable* for each cell of the Monte Carlo design. Given that the performance of an estimator will in general be better with increasing sample size, what is essential is the smallest sample size for which an estimator has acceptable performance given a specific distributional condition. This sample size will be called the *sufficient sample size* for that distributional condition. The necessary conditions for the performance of an estimator to be called acceptable are given below.

Some robustness studies have investigated many distributional conditions for a specific model, which gives as many sufficient sample sizes. We therefore try to summarize the sufficient sample sizes for specific distributional conditions. This summary is based on the finding that the performance of an estimator generally decreases when the skewness and kurtosis of the observed variables deviate more from zero.

In Section 5, it was explained that distributional conditions can be described by means of distributional characteristics (e.g., leptokurtic).

Suppose that for a specific model, the larger the kurtoses of a leptokurtic condition, the worse the performance of the ML parameter estimator. Suppose further that the sufficient sample size is the same for each leptokurtic condition and the normal condition. In that case, it is not very informative to present the sufficient sample size for each leptokurtic condition. In principle, it is sufficient to tabulate this sample size for the normal condition and the extreme leptokurtic conditions.

When some extreme conditions in a model study can be described by the same distributional characteristic, it often turns out that the results of an estimator are more or less comparable for these conditions. We will therefore only tabulate sufficient sample sizes for each distributional characteristic.

An important goal of the meta-analysis is to explain differences in conclusions across robustness studies. One of the things that can cause differences in conclusions are the assessment criteria used. Assessment criteria determine whether the realizations of a research summary statistic are acceptable; they therefore determine whether the performance of an estimator is acceptable. Because the assessment criteria should not be a cause for differences in conclusions, they are redefined and applied to the reported results in a robustness study whenever possible. The conclusions of the author(s) of a study are therefore not taken for granted.

Below, we describe procedures used to assess the performance of estimators for specific models and distributional characteristics. When the data needed to apply these procedures are not available, the conclusions of the author(s) will be followed.

We study three dependent variables explicitly in the meta-analysis. The characteristics of the model studies are used as explanatory variables. The dependent variables are the sufficient sample size regarding

(a) the bias of parameter estimates,
(b) the bias of estimated standard errors, and
(c) the rejection rate of the chi-square statistic at the 0.05 level.

These three variables are examined for each model study with a correctly specified model, each distributional characteristic (see Section 5), and the ML, GLS, ERLS, and ADF estimation method.

Parameter Estimates

The research summary statistic used to evaluate the behavior of an estimator for parameter θ_i is its *relative bias*

$$B(\hat{\theta}_i) = \frac{\bar{\hat{\theta}}_i - \theta_i}{\theta_i}, \quad i = 1, 2, \ldots, t, \tag{13}$$

where θ_i is the population value of the *i*th parameter ($\theta_i \neq 0$), $\bar{\hat{\theta}}_i$ is the mean of the estimates for the *i*th parameter across the *NR* replications, and *t* is the number of parameters to be estimated.

Criteria for acceptability. Now, consider the extreme distributional conditions with a specific distributional characteristic. If for one of these distributional conditions, $|B(\hat{\theta}_i)| < 0.05$ for $i = 1, 2, \ldots, t$, then all parameter estimates are considered to be *acceptable* for that condition.[8] Only if the parameter estimates are acceptable for each of these distributional conditions, their estimator is by definition acceptable for the specific distributional characteristic.

Standard Error Estimates

The research summary statistic used to determine the behavior of a standard error estimator regarding parameter θ_i is its *relative bias*

$$B(\hat{se}_{\hat{\theta}_i}) = \frac{\bar{\hat{se}}_{\hat{\theta}_i} - \hat{se}_{\theta_i}}{\hat{se}_{\theta_i}}, \quad i = 1, 2, \ldots, t, \tag{14}$$

where $\hat{se}_{\theta_i}$ is an estimate of the population value of the standard error of $\hat{\theta}_i$, and $\bar{\hat{se}}_{\hat{\theta}_i}$ is the mean of the estimated standard errors of $\hat{\theta}_i$ across the *NR* replications.

Criteria for acceptability. Consider the extreme distributional conditions with a specific distributional characteristic. The estimates for the standard errors are by definition *acceptable* for a distributional condition, if $| B(\hat{se}_{\hat{\theta}_i}) | < 0.1$ for $i = 1, 2, \ldots, t$, and, in addition, the mean absolute relative bias $\frac{1}{t}\sum_{i=1}^{t} | B(\hat{se}_{\hat{\theta}_i}) |$, abbreviated as the m.a.r.b., is smaller than 0.05.[9] When the estimated standard errors are acceptable for each considered distributional condition, their estimator is by definition acceptable for the distributional characteristic.

The Direction of Bias

When the bias of parameter estimates, or standard error estimates, is unacceptable for some N, it is interesting to know whether it is systematically positive or negative. The following procedure is applied to determine whether the bias of parameter estimates or estimated standard errors has a systematic direction. We evaluate the bias for sample sizes smaller than the sufficient sample size. Let b represent the acceptance boundary for individual parameters—that is, $b = 0.05$ for parameter estimates—and $b = 0.1$ for standard error estimates. For each combination of N and a distributional characteristic, the following rules are applied:

- bias is *negative*, if $\sum_{i=1}^{t} B(\cdot) < -bt^{1/2}$
- bias is *positive*, if $\sum_{i=1}^{t} B(\cdot) < bt^{1/2}$
- bias is *varying in sign* (across parameters), otherwise.

The bias for the specific model and distributional characteristic is defined as positive (negative) if the bias for each combination is positive (negative). In all other cases, the bias is by definition varying in sign.

Chi-Square Statistic

Rejection rate. The emphasis is on the rejection rate of the chi-square statistic at the 0.05 level.[10] The rejection rate equals the number of replications for which the model is rejected (the reject frequency, denoted as RF) divided by the total number of replications included in the analysis (denoted as NRI).

Criteria for acceptability. Consider the extreme distributional conditions with a specific distributional characteristic. Whether the rejection rate is acceptable for a specific sample size and one of these conditions depends on a statistical test. Given NRI and the significance level α, the associated RF is binomially distributed with parameters NRI and α when the model is true. From the binomial distribution, 99 percent confidence intervals for RF in the population can be computed (Sachs 1974:258). Because the main interest is in the sufficient sample size for each distributional characteristic, the mean of the RFs across the extreme distributional conditions, abbreviated as MRF, is computed.[11]

MRF is not directly viewed as acceptable when it falls inside the 99 percent confidence interval for its population value because the power of this test is rather low for the number of replications in most robustness studies. It is viewed as *acceptable* when it falls inside this confidence interval for the sample size under consideration and for each of the larger sample sizes.

However, MRF is not always viewed as unacceptable when this criterion is not met. When one solely relies on the statistical test, it could happen by chance that an MRF falls outside the 99 percent confidence interval for its population value. However, it can be inferred largely from other research summary statistics, such as the mean and standard deviation of the chi-square statistic and the corresponding MRF for other sample sizes, whether this occurs. In the meta-analysis, such a subjective correction of the objective criterion of acceptability was applied twice.

The rejection rate is (un)acceptable for a distributional characteristic when the associated MRF is (un)acceptable. When the rejection rate of a goodness-of-fit statistic is unacceptable, we speak of a positive bias of the rejection rate if a model is rejected more often than expected and of a negative bias of the rejection rate if a model is rejected less often than expected.

Mean of the chi-square statistic. Because assessing the performance of a chi-square statistic solely on its rejection rate at the 0.05 level is somewhat unreliable, an alternative dependent variable was also studied. This dependent variable is the sufficient sample size for an acceptable mean of the chi-square statistic across the *NR* replications. When the mean (*M*) and standard deviation (*SD*) of the chi-square statistic across the replications are available, it is tested whether *M* differs significantly from its expected value (the number of degrees of freedom of the model). This is done by means of the test statistic

$$T_{St} = \frac{M - df}{SD/(NR - 1)^{½}}, \qquad (15)$$

which has a Student *t*-distribution with *NR – 1* degrees of freedom under the null hypothesis H_0: $\mu = df$, where μ is the mean of the chi-square statistic in the population.

Criteria for acceptability. The mean of the chi-square statistic is defined as unacceptable if the null hypothesis is rejected at the 0.01 level when a two-sided hypothesis test is performed; H_1: $\mu \neq df$. When T_{St} is too large (small), the bias of the chi-square statistic is positive (negative).

6.2. SUFFICIENT SAMPLE SIZES FOR ESTIMATORS

In Table 4, for each model and distributional characteristic, the sufficient sample size for acceptable bias of parameter estimates, acceptable bias of standard error estimates, and an acceptable rejection rate of the chi-square statistic at the 0.05 level are given for the ML estimation method. For sample sizes smaller than the sufficient sample size, Table 4 also indicates whether an ML estimator has a bias that is positive, negative, or varying in sign.

Results for the ML estimation method are discussed in Section 6.3. For the GLS, ERLS, and ADF estimation method, corresponding results are also discussed in Section 6.3.[12] Results for the ERLS estimation method are only studied when the Mardia-based estimator

TABLE 4: Sufficient Sample Sizes for the ML Estimation Method

Model Number	*Distributional Characteristic*	*Parameter Estimate*	*Standard Error*	*Chi-Square Statistic*
10	normal	800 ±	100[t] +	*25*
11	normal	400 ±	100[t] +	100 +
13	categorical[a]	*400*	*400*[t]	*400*
	skewed[c]	>400 ±	*400*[t]	*400*
14	categorical[a]	*400*	*400*[t]	*400*
	skewed[d]	*400*	*400*[t]	>400 +
15	categorical[a]	*400*	*400*[t]	*400*
	skewed[d]	*400*	*400*[t]	>400 +
16	categorical[a]	*400*	*400*[t]	*400*
	skewed[d]	*400*	*400*[t]	*400*
17	normal	>400 ±	>400[t] +	100 +
18	normal	200 –	200[t] +	*25*
19	normal	100 –	50[t] +	*25*
20	normal	400 ±	>400[t] +	*25*
21	normal	400 –	200[t] +	*25*
22	normal	100 –	50[t] ±	*25*
23	normal	100 ±	200[t] +	50 +
24	normal	100 –	50[t] +	50 +
25	normal	50 ±	*25*[t]	50 +
26	normal	200 –	200[t] +	*25*
27	normal	50 –	50[t] ±	50 +
27	normal	50 –	50[t] +	50 +
28	normal	*25*	*25*[t]	50 +

Model	*Distributional*	*Parameter*	*Standard*	*Chi-Square Statistic*		
Number	*Characteristic*	*Estimate*	*Error*	*Ordinary*	*Corrected*	*Scaled*
31	normal	*500*	*500*[t]	*500*	*500*	
	leptokurtic[b]	*500*	>500[t] –	>500 +	*500*	
32	normal	*500*	*500*[t]	*500*	*500*	
	leptokurtic[b]	*500*	>500[t] –	>500 +	*500*	
33	skewed		>400 ±	*200*		*200*
	platykurtic		>400 +	*200*		*200*
	leptokurtic		>400 –	*200*		*200*
34	skewed		>400 ±	*200*		*200*
	platykurtic		>400 +	*200*		*200*
	leptokurtic		>400 –	>400 +	400 +	
36	2 cat.	>400	>400	>400 +		
	3 cat.	>400	>400	400 +		
	5 cat.	>400	>400	*200*		
	7 cat.	*200*	>400	*200*		
	2 to 5 cat.[b]	>400	>400	>400 +		
	7 cat.[b]	*200*	>400	*200*		
37	normal	>500 –		*500*		
	categorical	>500 –		*500*		
	skewed[c]	>500 ±		*500*		

TABLE 4 Continued

Model Number	*Distributional Characteristic*	*Parameter Estimate*	*Standard Error*	*Chi-Square Statistic*		
				Ordinary	*Corrected*	*Scaled*
39	normal	*200*	>400 ±	*200*	*200*	
	skewed	*200*	>400 –	*200*	*200*	
	platykurtic	*200*	>400 +	*200*	*200*	
	leptokurtic	400 –	>400 –	*200*	>400 –	
	leptokurtic[b]	400 –	>400 –	*200*	>400 –	
40	normal	*200*	>400 ±	*200*	*200*	
	skewed	*200*	400 ±	*200*	*200*	
	platykurtic	*200*	>400 +	*200*	*200*	
	leptokurtic	*200*	>400 –	>400 +	*200*	
	leptokurtic[b]	400 ±	>400 –	>400 +	*200*	
41	normal			*200*		
	leptokurtic[b]			400 +		
42	normal			*200*		
	leptokurtic[b]			>1200 +		
43	normal	600 ±	300	*75*	*75*	
	leptokurtic	2400 ±	>9600 –	>9600 +	300 +	
	leptokurtic[b]	600 ±	>9600 –	2400 +	>9600 –	
44	normal			500 +	500 +	
	nonnormal			500 +	500 +	
45	nonnormal			500 +	250 +	
46	leptokurtic			>5000 +	*150*	
49	categorical	*1000*	*1000*	*1000*		
	platykurtic[e]	*1000*	*1000*	*1000*		
	leptokurtic[c]	*1000*	>1000 –	*1000*		
	leptokurtic[e]	*1000*	>1000 –	>1000 +		
58	leptokurtic			*300*	>300 –	*300*

NOTE: When the sufficient sample size happens to be the minimum sample size investigated, it is printed in italics. When there is still relevant bias for the largest sample size investigated, the sample size is preceded by the > symbol.

a. Also platykurtic.

b. Also skewed.

c. Also categorical.

d. Also categorical and leptokurtic.

e. Also categorical and skewed.

t. With theoretical standard errors as population values.

of κ (Browne 1984) has been used because this estimator has the best performance among the available estimators of κ (Harlow 1985).

The sufficient sample size for an acceptable mean of the chi-square statistic and the direction of bias are not presented because the results are often comparable with those concerning the rejection rate of the chi-square statistic at the 0.05 level. They will be discussed, however, in Section 6.3.

TABLE 5: Performance of Estimators of Parameters and Standard Errors

Model Characteristics		*Data Characteristics*				
Number of Indicators/ Factors	*Mean Factor Loading*	*Mean Absolute Skewness*	*Mean Kurtosis*	*Estimation Method*	*Parameter Estimate*	*Standard Error*
2	.8	0.0	0.0	ML	500 ±	
3	.5	0.0	0.0	ML	600 ±	
3	.7	0.0	0.0	ML, ERLS	200 ±	600 ±
3	.7	≤1.5	≤5.0	ML, ERLS	400 ±	
3	.7	0.0	≤0.0	ADF	400 –	
3	.7	≥0.0	≥0.0	ADF	>400 –	
3	.85	0.0	0.0	ML	100 –	
4	.5	0.0	0.0	ML	150 ±	600
4	.7	0.0	0.0	ML	75 –	
4	.85	0.0	0.0	ML	50 ±	600

Tables 5 and 6 summarize the sufficient sample sizes for ML, GLS, ERLS, and ADF estimators under several conditions regarding the data and the model investigated. The direction of bias when the sample size is too small is also given. The lower bound for the mean kurtosis of the observed variables is always equal to –1.0. Again, the results of these tables are discussed in Section 6.3.

6.3. RANKINGS OF ML, GLS, ERLS, AND ADF ESTIMATORS

For each model study in which at least two estimation methods are studied, a ranking is made concerning the performance of the estimators of parameters and standard errors and the chi-square statistic. The ranking is made separately for each distributional characteristic.

For a specific distributional characteristic, an estimator gets a better ranking than another estimator when the values of the accompanying r.s.s. are better according to some rules that are given below. The relative performance of estimators is evaluated across sample sizes, as follows. It is evaluated first for the largest sample size smaller than or equal to 500.[13] When all sample sizes are larger than 500, the smallest sample size is chosen. The choice for the value of 500 seems to have a minor influence on the obtained rankings.

Each pair of estimators is evaluated on their relative performance, according to the following procedure. When one estimator performs

TABLE 6: Performance of Chi-Square Statistics

Model Characteristics						
Size	*ICSF*	*Categorical*	*Mean Absolute Skewness*	*Mean Kurtosis*	*Estimation Method*	*Chi-Square Statistic*
$df \leq 10$	yes	no	≤1.5	≤5.0	ML, ADF	200 +
$df \leq 20$	no	no	≤1.0	≤0.0	ADF	400 +
$df \leq 20$		no	≤.75	≤1.5	ERLS	200 –
$df \leq 24$		no	0.0	0.0	ML	100 +
$20 < df \leq 60$		no	≤.75	≤1.5	ML	200 +
$20 < df \leq 60$		no	≤.75	≤1.5	ERLS	400 –
$20 < df \leq 60$	yes	no	≤1.5	≤5.0	ADF	1000 +
$20 < df$	yes	yes			ADF	>1000 +
$60 < df < 100$		no	0.0	0.0	ML	>250 +
$60 < df < 100$		no	0.0	≤2.0	ML	500 +
$60 < df < 100$		no	0.0	0.0	ERLS	250 +
$60 < df < 100$		no			ADF	>1000 +
$df < 100$		no	0.0	≤2.0	GLS	150 +
$df < 100$		yes	0.0	0.0	GLS	500 +

NOTE: A blank means that the value of the specific entry does not matter.

worse than the other for a specific sample size, the relative performance of the former estimator is worse, and the procedure stops. However, when two estimators have comparable performance for that sample size, the following is done.

- When the two estimators both have acceptable performance, the relative performance on the next smaller sample size is investigated.
- When the two estimators both have unacceptable performance, the relative performance on the next larger sample size is investigated.

This procedure is repeated until the two estimators show a difference in performance, or until the performance has to be evaluated for a sample size already considered. In the latter case, the estimators are said to have comparable performance. When one estimator has a better performance than the other, the former gets a better ranking. When two estimators have a comparable performance, these estimators do not automatically get the same ranking. For example, when estimator A has a better performance than estimator B, and estimator C has a comparable performance with both estimators A and B, estimator C gets a better ranking than estimator B.

The problem now is to formulate rules on how to assess the relative performance of estimators adequately and unambiguously by means of an r.s.s.[14]

Parameter estimators and standard error estimators. Consider a specific model, sample size, and the extreme distributional conditions with a specific distributional characteristic. The rules for bias of parameter estimators and bias of standard error estimators are

- When for each condition the values of $B(\hat{\theta}_i)$ or $B(\widehat{se}_{\hat{\theta}_i})$, $i = 1, 2, \ldots, t$, as defined by (13) and (14), respectively, are acceptable for estimator A, whereas those values are unacceptable for estimator B for at least one condition, estimator A performs better.
- When for each condition the values of $B(\hat{\theta}_i)$ or $B(\widehat{se}_{\hat{\theta}_i})$ are acceptable for two estimators, these estimators have comparable performance.
- When for at least one condition the values of $B(\widehat{se}_{\hat{\theta}_i})$ are unacceptable for estimator A and the same holds for estimator B, estimator A performs worse than B when the mean of its m.a.r.b.s across the distributional conditions is at least 0.05 larger compared with that of B. Otherwise, the two estimators have comparable performance.

Example. Table 7 gives the relative bias of the estimated standard error for each of the $t = 13$ parameters for conditions B and D under model 39, which is an ICSF CFA model studied by Harlow (1985), and $N = 400$. The estimation methods considered are ML, ERLS, and ADF. For them, the values of the r.s.s. as defined by (14) are unacceptable, both because some values fall outside the range [–0.1, 0.1] and because the m.a.r.b. is larger than 0.05.

For $N = 400$, the ADF and ML standard error estimator, and conditions B and D, the sum of the r.s.s. is smaller than $-0.1t^{1/2} = -.36$. Because this also holds for $N = 200$, the bias for both estimators is negative under leptokurtic conditions. The sum of this r.s.s. is between $-0.1t^{1/2}$ and $0.1t^{1/2}$ for ERLS and conditions B and D. Irrespective of the value of that sum for $N = 200$, the bias of the ERLS estimator therefore varies in sign.

For model 39 and the leptokurtic distributional characteristic, the mean of the m.a.r.b. across the two conditions B and D is more than 0.05 smaller for ADF than for ERLS and ML. The ADF estimator of

TABLE 7: Relative Bias of Estimated Standard Errors for Model 39 and $N = 400$ for the ML, ERLS, and ADF Estimator

	ML		*ERLS*		*ADF*	
Parameter	*B*	*D*	*B*	*D*	*B*	*D*
λ_1	–0.15	–0.06	0.14	0.26	–0.05	–0.05
λ_2	–0.23	–0.20	0.04	0.09	–0.15	–0.16
λ_3	–0.09	–0.12	0.21	0.19	–0.06	–0.06
λ_4	–0.39	–0.39	–0.14	–0.15	–0.10	–0.09
λ_5	–0.40	–0.41	–0.15	–0.17	–0.10	–0.08
λ_6	–0.44	–0.49	–0.21	–0.28	–0.08	–0.08
ϕ	–0.02	–0.02	0.32	0.32	–0.08	–0.06
θ_1	–0.15	–0.02	0.14	0.32	–0.02	–0.02
θ_2	–0.09	–0.03	0.24	0.32	0.07	0.09
θ_3	–0.48	–0.54	–0.24	–0.33	–0.06	–0.08
θ_4	–0.33	–0.33	–0.06	–0.06	–0.08	–0.06
θ_5	–0.46	–0.49	–0.22	–0.27	–0.04	–0.03
θ_6	–0.30	–0.36	–0.06	–0.13	–0.05	–0.04
sum	–3.52	–3.47	1.69	0.86	–0.80	–0.72
m.a.r.b	0.27	0.27	0.17	0.22	0.07	0.07
mean m.a.r.b.	0.27		0.20		0.07	

standard errors has therefore the best ranking, as can be seen from Table 8.

Chi-square statistics. Consider the extreme distributional conditions with a distributional characteristic for a specific model and sample size. First, MRF across these distributional conditions is computed for each estimation method. The following rules are then used for the rejection rate of the chi-square statistics at the 0.05 level.

- When MRF falls inside the 99 percent confidence interval for its population value for estimation method A, while MRF falls outside the corresponding confidence interval for estimation method B, estimation method A performs better.
- When two estimation methods both result in an MRF that falls outside the 99 percent confidence interval, method A performs worse than method B when MRF for method A is more than two times further away from the expected value (in the same direction) compared with MRF for method B.

TABLE 8: Rankings of Estimators for Several Models and Distributional Characteristics

Model	*Distributional*	*ML*			*GLS*			*ERLS*			*ADF*			*CML*	*SML*
Number	*Characteristic*	*p*	*s*	*g*	*p*	*s*	*g*	*p*	*s*	*g*	*p*	*s*	*g*	*g*	*g*
31	normal	1	1	1							1	1	1	1	
	leptokurtic[a]	1	2	3							2	1	1	1	
32	normal	1	1	1							1	1	1	1	
	leptokurtic[a]	1	2	3							1	1	1	1	
33	skewed		1	1								1	1		1
	platykurtic		2	1								1	1		1
	leptokurtic		2	1								1	1		1
34	skewed		1	1								1	3		1
	platykurtic		2	1								1	1		1
	leptokurtic		2	2								1	2		1
39	normal	1	1	1				1	1	1	3	1	1	1	
	skewed	1	2	1				1	1	1	3	3	4	1	
	platykurtic	1	1	1				1	1	1	3	1	1	1	
	leptokurtic	1	3	1				1	2	3	3	1	1	3	
	leptokurtic[a]	1	3	1				1	2	3	3	1	1	3	
40	normal	1	1	1				1	1	1	3	1	4	1	
	skewed	1	1	1				1	1	1	3	3	1	1	
	platykurtic	1	3	1				1	1	1	3	1	1	1	
	leptokurtic	1	3	3				1	2	1	3	1	3	1	
	leptokurtic[a]	1	3	3				1	2	2	3	1	3	1	
43	normal	1	1	1	1	1	1				1	1	4	1	
	leptokurtic	1	2	3	1	2	3				1	1	1	2	
	leptokurtic[a]	1	2	3	1	2	2				1	1	1	4	
44	normal			3			1			1			5		3
	nonnormal			2			1			5			4		2
45	nonnormal			3			1			5			4		2
46	leptokurtic			4			4			3			2		1

NOTE: An estimator that is ranked as best is given the number 1. A higher number indicates a lower ranking position. The three entries in the third to eighth column correspond subsequently to the rankings in the performance of the estimators of parameters (p), estimators of standard errors (s), and goodness-of-fit statistic (g). The last two columns refer to the rankings of the corrected (CML) and scaled (SML) ML chi-square statistic.

a. Also skewed.

- When one of these two conditions is not met, the two estimation methods have comparable performance.

TABLE 9: Leptokurtic Conditions of Model 40

Condition	*Kurtoses*					
A	1	1	1	1	1	1
B	6	6	6	6	6	6
C	–1	0	1	1	2	3
D	2	5	8	6	7	8

SOURCE: Harlow (1985). Reprinted by permission.

TABLE 10: Reject Frequency (RF), Mean Reject Frequency (MRF), Number of Replications Included in the Analysis (NRI), and Expected Reject Frequency (E) of Four Chi-Square Statistics for Model 40, Conditions B and D, and Two Sample Sizes *N*

	N = 400						N = *200*					
	B		*D*				*B*		*D*			
χ^2	*RF*	*NRI*	*RF*	*NRI*	*MRF*	*E*	*RF*	*NRI*	*RF*	*NRI*	*MRF*	*E*
ML	32	100	32	100	32.0	5.0	38	100	43	100	40.5	5.0
CML	6	100	4	100	5.0	5.0	4	100	4	100	4.0	5.0
ERLS	3	82	3	80	3.0	4.1	3	69	4	71	3.5	3.5
ADF	20	100	19	100	19.5	5.0	42	97	40	96	41.0	4.8

Example. Consider model 40, a non-ICSF CFA model with six observed variables studied by Harlow (1985), for the leptokurtic distributional characteristic. Harlow investigated four leptokurtic conditions with zero univariate skewness. The kurtosis of the six variables in these four conditions is given in Table 9. In this set of conditions, B and D are the two extreme conditions; therefore, no further attention is paid to A and C. The goodness-of-fit reject frequencies for ML, corrected ML (CML), ERLS and ADF, for the two extreme leptokurtic conditions B and D, are given in Table 10.

The total number of replications was fixed at 100. The nonconvergent replications and replications with improper solutions were excluded from the analysis. The number of replications that is used can therefore differ across conditions and across estimation methods. With *NR* = 100, the 99 percent confidence interval for RF in the population is approximately [1.1, 13.5]. The 99 percent confidence interval for MRF in the population is approximately [1.9, 10.4].

For both sample sizes, the performance of the ML and ADF chi-square statistic is unacceptable because the corresponding MRFs are too large. Neither for $N = 400$ nor for $N = 200$ is the difference between the mean reject frequency for ML and its expected value two times larger than the difference between the mean reject frequency for ADF and its expected value. The ML and ADF chi-square statistics are therefore considered to have comparable performance; CML and ERLS perform well for both sample sizes. The ranking of the chi-square statistics for model 40 and the leptokurtic distributional characteristic can be found in Table 8.

6.4. FINDINGS FOR ML, GLS, ERLS, AND ADF METHODS

In this section, the performance of ML, GLS, ERLS, and ADF estimation methods is assessed and summarized with regard to bias of parameter estimates, bias of standard error estimates, and mean and rejection rate of the chi-square statistic. For each of these aspects, the behavior of estimators is discussed first under normality, then under nonnormality, and finally the performance of adjusted estimators is treated. When no reference is given, the results follow from the meta-analysis.

Due to space limitations, the rankings of the parameter estimators, the standard error estimators, and the chi-square statistics are given in Table 8 for a subset of the model studies only.[15] From the complete set of rankings, it can be observed that the ML, GLS, and ERLS parameter estimators generally have comparable performance. The ADF parameter estimator performs worse when the investigated sample size is small. The ADF standard error estimator is superior when the observed variables have an average kurtosis larger than three and the sample size is at least 400. The rankings of the chi-square statistics show a varying pattern because they depend on many characteristics. For instance, the performance of the ML chi-square statistic is worse than ADF when the model is ICSF with less than 35 degrees of freedom, the sample size is at least 400, and the observed variables are both skewed and leptokurtic.

Parameter Estimates

For multinormally distributed variables, there is empirical evidence that the quality of ML parameter estimates depends on the model under study. For the 12-factor models studied by Boomsma (1983), the bias of ML parameter estimates is smaller for models with more indicators per factor and higher factor loadings. This trend is also visible for other models with multinormally distributed variables.

The bias of ML parameter estimates increases when the levels of univariate skewness and kurtosis deviate increasingly from normal theory values. A larger sample size is a remedy to obtain unbiased parameter estimates. The effect of categorization of variables on the bias of ML estimates is still obscure. The bias seems to be somewhat worse for categorical variables when the univariate skewness of the variables varies in sign (Ethington 1987).[16]

Harlow (1985) concludes that ML and ERLS parameter estimates were comparable. Henly (1993) notes a striking resemblance between ML and GLS estimates. Muthén and Kaplan (1985) also did not find discernible differences between ML and GLS. In general, however, these parameter estimators were not compared very often. The bias of ML, GLS, and ERLS estimates is negative, or varying in sign, as can be observed from Tables 4 and 5.

ADF needs a larger sample size to estimate the parameters properly. Except for a multiplicative model (Henly 1993), unacceptable bias of ADF parameter estimates is negative. This bias was most substantial for distributional conditions with large positive kurtoses. The ADF parameter estimator is therefore not distribution free.

Standard Error Estimates

For normally distributed variables, Boomsma (1983) gave empirical evidence that the bias of ML standard error estimates depends on model characteristics. This bias becomes smaller when the value of factor loadings and the number of indicators per factor increase.

Important differences in the sufficient sample size for acceptable bias of ML standard error estimates can also be explained by taking into consideration the way in which the population values for the standard errors are obtained (see Table 4). The theoretical standard errors are unrealistically near to the mean of the estimated standard errors. When the theoretical standard errors are used as population values, the sufficient sample size for acceptable bias of ML estimates is often much smaller than 500 for multinormally distributed variables.

On the other hand, when the empirical standard errors represent the population values, it can be concluded from Table 4 that sample sizes as large as 400 give unacceptable ML standard error estimates. Under multinormality, a sample size of 600 seems large enough to give acceptable bias of ML estimates (see Table 5). For smaller sample sizes, the bias is either negative or varying in sign.

The degree of skewness has a minor influence on the bias of ML standard error estimates for continuous distributions. For discrete distributions, the bias seems larger when variables are skewed. The bias increases when the absolute value of the kurtosis increases. There is a remarkable effect of the sign of the kurtosis: The bias of ML estimates is positive for platykurtic distributions and negative for leptokurtic distributions. This bias is often worst when the underlying distribution is highly leptokurtic.

When the GLS, ERLS, or ADF estimation method is used, the bias of standard error estimates is also worst for leptokurtic distributions. The bias of GLS estimates is negative for these distributions. The effect of kurtosis on the ADF estimates is less extreme than with ML, GLS, and ERLS. The bias of ADF estimates is mostly negative, irrespective of the distribution of the observed variables.

When a population model is non-ICSF, (i.e., not invariant under a constant scaling factor), ML, ERLS, and ADF standard error estimates are considerably worse when the distribution of the observed variables is leptokurtic. For the ML method, this is also true for platykurtic distributions. The effect of a non-ICSF population model has not been investigated for the GLS estimation method.

Chou, Bentler, and Satorra (1991) studied the performance of the robust ML estimator of standard errors. When the observed variables had excessive kurtosis, these robust estimates were superior to the usual ML estimates, using empirical standard errors as a criterion.

Chi-Square Statistic

The performance of the chi-square statistic is only evaluated for correctly specified models, that is, when H_0 is true. The performance of the chi-square statistic under H_1 is not evaluated because robustness studies investigating the effect of model misspecification were not included in the meta-analysis (see Table 1).

The behavior of a chi-square statistic T depends on the size of the model (i.e., the number of degrees of freedom), as can be seen in Table 6. Especially the ADF chi-square statistic (T_{ADF}) is sensitive to the model size. T_{ADF} needs a sample size larger than 1,000 for models with more than 60 degrees of freedom. T_{GLS} and T_{ERLS} seem to be somewhat less sensitive to model size than T_{ML}.

For a specific model, the rejection rate and mean of T_{ML}, T_{GLS}, and T_{ERLS} are often worst when the underlying distribution has a large kurtosis. This is not the case for T_{ADF} when a model is ICSF because the performance seems to be independent of the distribution of the variables then. When a model is non-ICSF, the performance of T_{ML} and T_{ADF} becomes worse for the same leptokurtic conditions.

Hu, Bentler, and Kano (1992) showed that the scaled T_{ML} is preferable compared with T_{ML} for a large model, although the performance of the scaled T_{ML} can be somewhat worse for small sample size. They showed that T_{ML} and T_{GLS} completely break down when the latent variables and measurement errors are not independently distributed. This is not the case for the scaled T_{ML}. Yung and Bentler (1994) show that a bootstrap correction of additive bias on T_{ADF} yields the desired tail behavior for a sample size of 500, even if the latent variables and measurement errors are dependent.

7. CONCLUSIONS AND TOPICS FOR FUTURE RESEARCH

By means of a robustness study, in principle, guidelines can be given for the estimation method to be used under specific circumstances. For instance, when an estimator performs poorly in a robustness study for a specific model, distribution of the observed variables, and sample size, this estimator will probably perform worse in a real-life situation when a comparable model, distribution, and sample size are investigated. The reason is that many other factors, such as independency of observations, are often worse compared with a simulation study. In practice, the sample size required for an acceptable performance of an estimator in a robustness study should therefore be recommended as a lower bound.

To generalize findings concerning the performance of estimators across robustness studies, several causes for observed differences in conclusions across robustness studies were detected. For instance, in several robustness studies, the conclusions concerning the performance of standard error estimators were too optimistic because the estimates of standard errors were compared with the wrong type of population values.

In general, the ML, GLS, and ERLS parameter estimators seem to be comparable. Bias of these estimators increases when the levels of univariate skewness and kurtosis deviate more from normal theory values. When these levels increase, a larger sample size is a remedy to obtain acceptable parameter estimates. The ADF parameter estimator may need a larger sample size compared with the ML, GLS, and ERLS parameter estimators, especially when the model has more than 12 observed variables.

The standard error estimates are unreliable if the sample size is smaller than 500, regardless of the estimation method used. When the ML or GLS estimation method is used, the standard errors are underestimated when the variables have positive kurtosis and overestimated when the variables have negative kurtosis, regardless of the sample size. The ADF standard error estimates are superior when the observed variables have an average kurtosis larger than 3.0 and the sample size is at least 400.

The ML chi-square statistic rejects the true model too often when the sample size is smaller than five times the size of the model (i.e., the number of degrees of freedom of the model). When the observed variables have an average positive kurtosis as large as 5.0, the sample size may have to be increased up to 10 times the size of the model. Given that the model is correct, the GLS chi-square statistic may have an acceptable performance for a sample size that is two times smaller than the sample size needed for an acceptable performance of the ML chi-square statistic. The ADF estimation method is relatively insensitive for the distribution of the observed variables, but a disadvantage of this method is that a relatively large sample size is needed. When the sample size is smaller than 20 times the size of the model, the ADF chi-square statistic rejects the true model too often. Another disadvantage of the ADF estimation method is that the necessary computations are hardly feasible when there are more than about 30 observed variables.

In robustness studies, the size of the models is often small compared with the models applied in practice. Larger models should therefore be investigated in future studies, which will probably result in findings that are more disappointing regarding the chi-square statistic. Because of many obscurities, the effect of model characteristics on the performance of all estimators considered in this article is an important topic for future research.

There are several analytical adjustments available for estimators of standard errors and the chi-square statistic that can be applied after the model parameters have been estimated (Satorra and Bentler 1988). Because of the promising performance of adjusted ML estimators, they should be studied intensively in future simulations. Applying those adjustments to the GLS and ERLS estimators of standard errors and the associated chi-square statistics may also lead to substantial improvements in the estimation of structural equation models.

NOTES

1. We implicitly refer to the relative kurtosis, that is, the kurtosis minus 3.

2. The conditions for scale invariance are violated when, for example, factor loadings have more than one fixed nonzero value per latent variable, error variances have fixed nonzero values, or parameters are constrained to be equal to other parameters (Cudeck 1989:326).

3. A necessary condition to include a robustness study in the overview is that it reports results concerning the bias of parameter estimates, the bias of estimated standard errors, or distributional properties of the chi-square statistic. When some of those results are available for each model in a subset of all the models investigated, that subset will be included in the meta-analysis of Section 6. A further condition for including a study in the overview is that the behavior of estimators mentioned in Section 2, or analytical corrections of those estimators, was examined. Other conditions are that at least 20 replications have been conducted and that covariance matrices or correlation matrices based on product moment correlations were analyzed.

4. A distributional characteristic is platykurtic (leptokurtic) when the univariate skewnesses are zero, and the univariate kurtoses are mainly negative (positive). A leptokurtic distribution is

peaked with relatively long tails, whereas a platykurtic distribution is flat with relatively short tails, both relative to a normal distribution.

5. Tables that give these characteristics for each model study can be obtained from the first author.

6. Throughout this article, we assume that the choice of software available for estimation of parameters, standard errors, and chi-square test statistics (i.e., a specific version of EQS, LISCOMP, or LISREL) has no substantial influence on the research results. This assumption is relevant because these programs do not always use the same algorithm. We also assume that the choice of starting values, maximum number of iterations, convergence criteria, and the approach regarding improper solutions does not influence the research results.

7. A disadvantage of exclusion of replications with improper solutions, rather than inclusion, is that the sampling distribution of S differs more from the population distribution. A disadvantage of inclusion is that it gives problems of interpretation.

8. A boundary for acceptance of 0.05 is often used in robustness studies.

9. The chosen boundary value of 0.1 for an acceptable $B(\widehat{se}\hat{\theta}_i)$ for individual parameters is two times larger than for parameter estimates because the population values of the standard errors have to be estimated. A boundary of 0.1 is also frequently used in robustness studies. On the other hand, the additional acceptance boundary for the m.a.r.b. is imposed to compensate somewhat for the less stringent criteria of acceptability of estimated standard errors as opposed to parameter estimates. The boundary of 0.05 for the acceptance of the m.a.r.b. is chosen because even when the absolute values of the relative biases are smaller than 0.1, they are viewed as unacceptable when they are in general more often in the range [0.05, 0.10] than in the range [0.0, 0.05].

10. The main reason for considering the 0.05 level is that the rejection rate is mostly available for that level; moreover, the use of $\alpha = 0.05$ is very common in applications.

11. The mean reject frequency (MRF) in the population has a 99 percent confidence interval that is more narrow than the 99 percent confidence interval for a reject frequency (RF) in the population.

12. Tables with sufficient sample sizes for those estimation methods for each model and distributional characteristic can be obtained from the first author.

13. The number 500 has been chosen because an estimator A that performs well for moderate to large sample sizes is considered to be superior compared with an estimator B that performs badly, irrespective of the sample size, even if estimator B performs relatively well for a small sample size.

14. These rules are initially applied, but we shall rely on the conclusions of the author(s) of a robustness study when there is not enough information available to apply them.

15. The model studies that were selected contained relatively much information regarding the rankings of estimators.

16. Using a population methodology, Olsson (1979) already observed that the bias of estimated factor loadings is larger when categorical variables have fewer categories, and larger skewness, or when they are skewed in opposite direction.

REFERENCES

Anderson, James C. and David W. Gerbing. 1984. "The Effect of Sampling Error on Convergence, Improper Solutions, and Goodness-of-Fit Indices for Maximum Likelihood Confirmatory Factor Analysis." *Psychometrika* 49:155-73.

Babakus, Emin, Carl E. Ferguson, and Karl G. Jöreskog. 1987. "The Sensitivity of Confirmatory Maximum Likelihood Factor Analysis to Violations of Measurement Scale and Distributional Assumptions." *Journal of Marketing Research* 24:222-28.

Baldwin, Beatrice O. 1986. "The Effects of Structural Model Misspecification and Sample Size on the Robustness of LISREL Maximum Likelihood Parameter Estimates." Ph.D. dissertation, Lousiana State University.

Bearden, William O., Subhash Sharma, and Jesse E. Teel. 1982. "Sample Size Effects on Chi-Square and Other Statistics Used in Evaluating Causal Models." *Journal of Marketing Research* 19:425-30.

Benson, Jeri and John A. Fleishman. 1994. "The Robustness of Maximum Likelihood and Distribution-Free Estimators to Non-Normality in Confirmatory Factor Analysis." *Quality and Quantity* 28:117-36.

Bentler, Peter M. 1995. *EQS Structural Equations Program Manual*. Encino, CA: Multivariate Software.

Bentler, Peter M. and Paul Dudgeon. 1996. "Covariance Structure Analysis: Statistical Practice, Theory, and Directions." *Annual Review of Psychology* 47:563-92.

Bentler, Peter M. and David G. Weeks. 1980. "Linear Structural Equations With Latent Variables." *Psychometrika* 45:289-308.

Bollen, Kenneth A. 1989. *Structural Equations With Latent Variables*. New York: John Wiley.

———. 1995. "Models That Are Nonlinear in Latent Variables: A Least Squares Estimator." *Sociological Methodology* 25:223-51.

———. 1996. "An Alternative Two Stage Least Squares (2SLS) Estimator for Latent Variable Equations." *Psychometrika* 61:109-21.

Boomsma, Anne. 1983. "On the Robustness of LISREL (Maximum Likelihood Estimation) Against Small Sample Size and Non-Normality." Ph.D. dissertation, Sociometric Research Foundation, Rijksuniversiteit Groningen, Amsterdam.

Box, George E. P. 1953. "Non-Normality and Tests on Variances." *Biometrika* 40:318-35.

Brown, Richard L. 1990. "The Robustness of 2SLS Estimation of a Non-Normally Distributed Confirmatory Factor Analysis Model." *Multivariate Behavioral Research* 25:455-66.

Browne, Michael W. 1974. "Generalized Least-Squares Estimators in the Analysis of Covariance Structures." *South African Statistical Journal* 8:1-24.

———. 1982. "Covariance Structures." Pp. 72-141 in *Topics in Multivariate Analysis*, edited by D. M. Hawkins. Cambridge: Cambridge University Press.

———. 1984. "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures." *British Journal of Mathematical and Statistical Psychology* 37:62-83.

Browne, Michael W., Gerhard Mels, and Mark Coward. 1994. "Path Analysis: RAMONA." Pp. 163-224 in *SYSTAT for DOS: Advanced Applications*. Evanston, IL: Systat.

Chou, Chih-Ping, Peter M. Bentler, and Albert Satorra. 1991. "Scaled Test Statistics and Robust Standard Errors for Non-Normal Data in Covariance Structure Analysis: A Monte Carlo Study." *British Journal of Mathematical and Statistical Psychology* 44:347-57.

Cudeck, Robert. 1989. "Analysis of Correlation Matrices Using Covariance Structure Models." *Psychological Bulletin* 105:317-27.

Curran, Patrick J., Stephen G. West, and John F. Finch. 1996. "The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis." *Psychological Methods* 1:16-29.

Dolan, Conar V. 1994. "Factor Analysis of Variables with 2, 3, 5 and 7 Response Categories: A Comparison of Categorical Variable Estimators Using Simulated Data." *British Journal of Mathematical and Statistical Psychology* 47:309-26.

Ethington, Corinna A. 1987. "The Robustness of LISREL Estimates in Structural Equation Models With Categorical Variables." *Journal of Experimental Education* 55:80-88.

Gallini, Joan and Garrett K. Mandeville. 1984. "An Investigation of the Effect of Sample Size and Specification Error on the Fit of Structural Equation Models." *Journal of Experimental Education* 51:9-19.

Gerbing, David W. and James C. Anderson. 1985. "The Effects of Sampling Error and Model Characteristics on Parameter Estimation for Maximum Likelihood Confirmatory Factor Analysis." *Multivariate Behavioral Research* 20:255-71.

Harlow, Lisa L. 1985. "Behaviour of Some Elliptical Theory Estimators With Nonnormal Data in a Covariance Structures Framework: A Monte Carlo Study." Ph.D. dissertation, University of California, Los Angeles.

Harlow, Lisa L., Chih-Ping Chou, and Peter M. Bentler. 1986. "Performance of Chi-Square Statistic With ML, ADF, and Elliptical Estimators." Presented at the Psychometric Society meeting, Toronto, Canada, June.

Hayduk, Leslie A. 1996. *LISREL Issues, Debates, and Strategies*. Baltimore, MD: Johns Hopkins University Press.

Henly, Susan J. 1993. "Robustness of Some Estimators for the Analysis of Covariance Structures." *British Journal of Mathematical and Statistical Psychology* 46:313-38.

Herting, Jerald R. and Herbert L. Costner. 1985. "Respecification in Multiple Indicator Models." Pp. 321-93 in *Causal Models in the Social Sciences*, edited by H. M. Blalock. New York: Aldine.

Hu, Li-Tze, Peter M. Bentler, and Yutaka Kano. 1992. "Can Test Statistics in Covariance Structure Analysis Be Trusted?" *Psychological Bulletin* 112:351-62.

Jaccard, James and Choi K. Wan. 1995. "Measurement Error in the Analysis of Interaction Effects Between Continuous Predictors Using Multiple Regression: Multiple Indicator and Structural Equation Approaches." *Psychological Bulletin* 117:348-57.

Jöreskog, Karl G. 1973. "A General Method for Estimating a Linear Structural Equation System." Pp. 85-112 in *Structural Equation Models in the Social Sciences*, edited by A. S Goldberger and O. D. Duncan. New York: Seminar.

———. 1993. "Testing Structural Equation Models." Pp. 294-316 in *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long. Newbury Park, CA: Sage.

Jöreskog, Karl G. and Fan Yang. 1996. "Nonlinear Structural Equation Models: The Kenny-Judd Model with Interaction Effects." Pp. 57-88 in *Advanced Structural Equation Modeling: Issues and Techniques*, edited by G. A. Marcoulides and R. E. Schumacker. Mahwah, NJ: Erlbaum.

Kaplan, David. 1989. "A Study of the Sampling Variability and Z-Values of Parameter Estimates from Misspecified Structural Equation Models." *Multivariate Behavioral Research* 24:41-57.

Kenny, David A. and Charles M. Judd. 1984. "Estimating the Non-Linear and Interactive Effects of Latent Variables." *Psychological Bulletin* 96:201-10.

Klein, Andreas, Helfried Moosbrugger, Karin Schermelleh-Engel, and Dirk Frank. 1997. "A New Approach to the Estimation of Latent Interaction Effects in Structural Equation Models." Pp. 479-86 in *Advances in Statistical Software 6*, edited by W. Bandilla and F. Faulbaum. Stuttgart: Lucius & Lucius.

Lance, Charles E., John M. Cornwell, and Stanley A. Mulaik. 1988. "Limited Information Parameter Estimates for Latent or Mixed Manifest and Latent Variable Models." *Multivariate Behavioral Research* 23:171-87.

Lee, Sik-Yum, Wai-Yin Poon, and Peter M. Bentler. 1995. "A Two-Stage Estimation of Structural Equation Models With Continuous and Polytomous Variables." *British Journal of Mathematical and Statistical Psychology* 48:339-58.

Meijer, Erik and Ab Mooijaart. 1992. "Factor Analysis Estimation Methods Compared Under Nonnormality Through a Monte-Carlo Study." Leiden Psychological Reports PRM 01-92.

Meijerink, Frits. 1995. *A Nonlinear Structural Equations Model.* Leiden: DSWO Press.

Muthén, Bengt and David Kaplan. 1985. "A Comparison of Some Methodologies for the Factor Analysis of Non-Normal Likert Variables." *British Journal of Mathematical and Statistical Psychology* 38:171-89.

———. 1992. "A Comparison of Some Methodologies for the Factor Analysis of Non-Normal Likert Variables: A Note on the Size of the Model." *British Journal of Mathematical and Statistical Psychology* 45:19-30.

Olsson, Ulf. 1979. "On the Robustness of Factor Analysis Against Crude Classification of the Observations." *Multivariate Behavioral Research* 14:485-500.

Ping, Robert A. 1995. "A Parsimonious Estimating Technique for Interaction and Quadratic Latent Variables." *Journal of Marketing Research* 32:336-47.

Potthast, Margaret J. 1993. "Confirmatory Factor Analysis of Ordered Categorical Variables With Large Models." *British Journal of Mathematical and Statistical Psychology* 46:273-86.

Reddy, Srinivas K. 1992. "Effects of Ignoring Correlated Measurement Error in Structural Equation Models." *Educational and Psychological Measurement* 52:549-70.

Sachs, Lothar. 1974. *Angewandte Statistik* (Applied Statistics). Berlin: Springer-Verlag.

Satorra, Albert and Peter M. Bentler. 1988. "Scaling Corrections for Statistics in Covariance Structure Analysis." UCLA Statistics Series 2.

Sharma, Subhash, Srinivas Durvasula, and William R. Dillon. 1989. "Some Results on the Behavior of Alternate Covariance Structure Estimation Procedures in the Presence of Non-Normal Data." *Journal of Marketing Research* 26:214-21.

Steiger, James H. 1995. "SEPATH." Pp. 283-487 in *STATISTICA 5.0.* Tulsa, OK: Statsoft.

Tanaka, Jeffrey S. 1984. "Some Results on the Estimation of Covariance Structure Models." Ph.D. dissertation, University of California, Los Angeles.

Vale, C. David and Vincent A. Maurelli. 1983. "Simulating Multivariate Nonnormal Distributions." *Psychometrika* 48:465-71.

Yang, Fan. 1997. "Non-Linear Structural Equation Models: Simulation Studies of the Kenny-Judd Model." Ph.D. dissertation, University of Uppsala.

Yung, Yiu-Fai and Peter M. Bentler. 1994. "Bootstrap-Corrected ADF Test Statistics in Covariance Structure Analysis." *British Journal of Mathematical and Statistical Psychology* 47:63-84.

Jeffrey J. Hoogland is a Ph.D. student in the Department of Statistics and Measurement Theory, University of Groningen. He graduated in statistics at the University of Amsterdam. His current research interests are covariance structure analysis, robustness questions, and simulation techniques.

Anne Boomsma is an associate professor in the Department of Statistics and Measurement Theory, University of Groningen. His current research interests are covariance structure analysis, statistical inference from iterative simulation (Markov Chain Monte Carlo Methods), bootstrap and jackknife analysis, regression analysis, and probability theory. His most recent publication is "Statistical Inference Based on Latent Ability Estimates" in Psychometrika *(1996).*