1

The Robustness of LISREL Modeling Revisited

January 10, 2001

Anne Boomsma[1] and Jeffrey J. Hoogland[2]

ABSTRACT Some robustness questions in structural equation modeling (SEM) are introduced. Factors that affect the occurrence of nonconvergence and improper solutions are reviewed in detail. Recent research on the behaviour of estimators for parameters, standard errors and model fit, under conditions of (non)normality, is summarized. It is emphasized that both model and sample data characteristics affect the statistical behaviour of these estimators. This knowledge may be used to set guidelines for a combined choice of sample size and estimation method. It is concluded that for large models, under a variety of nonnormal conditions, (robust) maximum likelihood estimators have relatively good statistical properties compared to other estimators (GLS, ERLS, ADF or WLS). The cumulative theoretical knowledge about robust (asymptotic) estimators and corrective statistics and the availability of practical guidelines from robustness research together, may enhance statistical practice in SEM and hence lead to more sensible and solid applied research.

1 Introduction

In maximum likelihood (ML) estimation of structural equation models the following statistical assumptions are made: (1) the sample observations, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, are independently distributed, where $\mathbf{x}_i$, $i = 1, \ldots, N$, is a random vector of k observed variables, (2) each vector $\mathbf{x}_i$ has a multivariate normal distribution, $\mathbf{x}_i \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0)$, (3) the hypothesized model $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ is approximately correct, where $\boldsymbol{\theta}$ is a vector of t model parameters, (4) a sample covariance matrix $\mathbf{S}$ is analyzed, and (5) the sample size N is very large, because, given the foregoing assumptions, asymptotic properties of parameter, standard error and model-fit estimators can be derived.

Each of these five assumptions may raise its own, specific robustness questions, comprehensively expressed as: What are the consequences with respect to the model estimates (i.e., parameters, standard errors of param-

[1]Department of Statistics and Measurement Theory, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands; e-mail: *a.boomsma@ppsw.rug.nl*

[2]Statistics Netherlands (CBS), P.O. Box 4000, 2270 JM Voorburg, The Netherlands; e-mail: *jhgd@cbs.nl*

eter estimates, and the fit of the model) if any of these assumptions are violated? In this chapter, only the robustness against small sample size and nonnormality is revisited. It is examined what can be learned from certain aspects of research in that area since the 1980s, and what guidelines regarding choice of sample size and SEM estimators can be offered to applied statisticians, today. Other, equally important robustness issues in SEM are not treated here.

2 Robustness against Small Sample Size

When researchers have a small sample, say $N < 200$, and do want to apply structural equation modeling (SEM), there are two persistent estimation problems likely to occur: nonconvergence and improper solutions. For both problems there is no really satisfying solution, given that the sample size cannot be increased and the user is stuck with his measurement instruments. In this section, some old results on both issues are revisited and some new ones are added. It is emphasized that the applied researcher should be well aware of the various factors having an impact on these two problems, and that some SEM estimators are more robust than others against the effects of small sample size in these matters.

2.1 Nonconvergence

In the 1980s nonconvergence (NC) meant that the iterative maximum likelihood estimation procedure in LISREL– the only one available in those days – did not converge within 250 iterations (or more). Empirical evidence from research in that decade (e.g., Boomsma, 1982, 1983, 1985) generally revealed that three main factors affect the occurrence of NC in factor models and, as a consequence, in structural models as well: sample size, size of factor loadings, and number of indicators per factor. The effect of sample size is most evident and probably well-known, but the impact of the other two explanatory variables is far less acknowledged.

Sample size. There is a primary effect of sample size N on the occurrence of nonconvergence. In general, if the model is correct, NC decreases with N, and for $N > 200$ there are hardly any problems (Boomsma, 1983). The larger the amount of independent sample information the better the chances to find a solution. This is illustrated in Table 1.1, where U and C represent uncorrelated and correlated two-factor models, respectively, with 3 or 4 indicators per factor, and small (S), medium (M), and large (L) factor loadings (see Boomsma, 1983, for details). In Table 1.1, and elsewhere in this chapter, *NR* denotes the number of Monte Carlo replications.

Size factor loadings. NC decreases with larger loadings; compare, for example, Models 3US, 3UM and 3UL in Table 1.1. In terms of population

TABLE 1.1. Percentage of nonconvergence; ML estimation, $NR = 300$.

	Sample Size				
Model	**25**	**50**	**100**	**200**	**400**
3US	48	28	13	2	
3CS	57	36	15	3	
3UM	12	1			
3CM	11	1			
3UL					
3CL	1				
4US	27*	8	1		
4CS	29*	8			
4UM	1				
4CM	2				
4UL					
4CL					

Note: Nearest integers; a blank entry means zero; *$NR = 100$.

covariances σ_{ij} it implies that NC increases as σ_{ij} gets closer to zero. Boomsma (1985) gave a partial explanation for this phenomenon by inspecting sign patterns of the observed sample covariances s_{ij} linked to same factor: for some models, inadmissible sign patterns had good predictive value for NC. For an uncorrelated two-factor model with three indicators per factor (Model 3US) and $N = 50$, a prediction rate of 99% was found.

Number of indicators per factor. If the number of indicators per factor (the NI/NF ratio) increases, NC decreases, that is, the larger the NI/NF ratio the better; compare, for example, Models 3US and 4US in Table 1.1.

The three factors that influence NC have in common that increasing information, in the form of independent observations, more reliable measurements, and a broader empirical enhancement in measuring latent variables ('validity'), decreases the occurrence of NC.

These results for factor models from the 1980s were recently confirmed and generalized by research of Marsh, Hau, Balla, and Grayson (1998); see also Marsh and Hau (1999). Marsh and his colleagues agree with Boomsma's (1982) recommendations to have at least $N = 100$ for NI/NF = 3 or 4, and that $N > 200$ is generally safer. Their generalization was that NI/NF = 2 requires at least $N = 400$, and that for NI/NF = 6 or 12 a sample as small as $N = 50$ is sufficient. These additional results support a *More is Better* conclusion – as they called it – for both N and NI/NF. The general implication of this research is that there is a mutual compensatory effect of N and NI/NF: a higher NI/NF ratio may compensate for small N, and larger N may compensate for a small NI/NF ratio.

As a consequence Marsh et al. (1998) deduced that it would be unwise to blindly follow general guidelines (rules of thumb) for the minimum number of observed variables k, given a sample size N and the size of the model, quantified by t, the number of parameters to be estimated – or vice versa. Such guidelines focus on minimum ratios of N/k or N/t. For example, an admittedly 'oversimplified guideline,' often referred to in the literature, is that of Bentler (1995, p. 6): $N/t = 5$ for normal or elliptical theory, and $N/t = 10$ for arbitrary distributions. Such prescriptions require NI/NF to be minimal if N small, which could be catastrophical because it would increase nonconvergence, and improper solutions as well.

Since models are often misspecified by lack of solid theoretical knowledge and reliable and valid measurement models, some attention to the possible effects of misspecification on nonconvergence is needed. Chou and Bentler (1995, p. 42) suggest that NC is caused by model misspecification and poor starting values. Clearly, reality is not that simple, and it could be doubted whether there is enough substantial, and unequivocal evidence for the statement. Boomsma (1985), for example, concluded that there was hardly any effect of starting values on NC if correct models are being analyzed. But what about the effects of misspecification? Luijben (1989, p. 69) and Camstra (1998, p. 89f., p. 114) found no strong indications for increasing NC with larger misspecifications. However, in a small study on factor models, Hendriks (1999) observed more NC and more local minima with increasing misspecification – and even more so with increasing N (using the LISREL program). When Hendriks and Boomsma re-analyzed the samples raising these problems, it was found that the use of 'arbitrary' starting values removed nonconvergence and local minima. It was concluded that under misspecified model conditions default starting values may cause such estimation problems, and its occurrence should be regarded as a first symptom of model-data discrepancy. More work has to be done in this area to fully understand the nature of these phenomena. In particular, the generalizability of these preliminary findings needs to be scrutinized.

2.2 Improper Solutions

Improper solutions (IS) of an estimated structural equation model refer to cases where one or more variance estimates have negative values – also referred to as Heywood cases. There is ample empirical evidence that the real danger for the occurrence of IS is a small sample, but there are additional factors that matter, similar to those that affect NC. Following Boomsma (1985), the main factors can be itemized as follows.

Sample size. With increasing N there are less IS; see Table 1.2.

Population covariances. Two cases can be distinguished here. On the one hand, across comparable models (for example, Models 3US, 3UM and

TABLE 1.2. Percentage of improper solutions; ML estimation, $NR = 300$.

	Sample Size				
Model	**25**	**50**	**100**	**200**	**400**
3US	51	41	22	11	3
3CS	47	33	18	6	3
3UM	38	21	11	1	
3CM	44	25	6	1	
3UL	26	9		1	
3CL	24	7	1		
4US	47*	19	3		
4CS	37*	15	4	1	
4UM	22	5			
4CM	27	1			
4UL	7	1			
4CL	7				

Note: Nearest integers; a blank entry means zero; *$NR = 100$.

3UL in Table 1.2) with increasing factor loadings there are less IS. On the other hand, within a single model (see Boomsma, 1985, Table 6) with increasing population values of variances parameters there are less IS, i.e., indicators with the largest loadings in a model show more IS. The latter phenomenon can be referred to as the Close to Zero case (Van Driel, 1978): as population variances get closer to zero, the probability of obtaining negative estimates of those variances increases.

Number of indicators per factor. Factor models with more indicators evoke less IS, hence, with a larger NI/NF ratio there are less IS; compare, for example Models 3UM and 4UM in Table 1.2.

It can be noticed that the same or similar factors affect the occurrences of NC and IS. Both problems are symptoms of empirical underidentification (cf. Rindskopf, 1984). Given small amounts of information (roughly to be translated in terms of independent observations, reliability and validity of measurements), inconsistencies between an hypothesized model and the empirical data are more likely to occur. In model estimation this may trigger either no solution at all, or an inadmissible solution.

In the 1980s, with slow computers and expensive computer time, mostly small models were being studied, and only the ML estimation procedure was available. Today, computers are much faster and cheaper in use, and meanwhile new estimators for SEM are available. Two questions to be answered next are (1) Can familiar small-model results be generalized to large models? and (2) Are these other estimators less incommoded with NC and IS than ML estimators? But other questions arise as well.

3 Comparing Estimators under Nonnormality

After a meta-analysis of robustness research in SEM, Hoogland and Boomsma (1998) concluded that it was necessary to know more about the robustness of nonnormality in large models, say models with a number of observed variables k larger than six or eight. Therefore, Hoogland (1999) studied a variety of such large models using Monte Carlo methods. In this and following sections an account is given of part of that research, aimed at the comparison of the behaviour of four estimators under eleven different conditions of nonnormality and varying sample size.

The estimators being compared were maximum likelihood (ML), generalized least squares (GLS), elliptical reweighted least squares (ERLS), and the asymptotically distribution-free (ADF) estimator, in LISREL also known as the weighted least squares (WLS) estimator.

The eleven distributional conditions (DCs) of the observed variables in the models under study can be characterized by their skewness γ and kurtosis κ. A rough summary of these conditions is given in Table 1.3, showing minimum, mean and maximum values of skewness and kurtosis over k variables. Condition A represents the normal case (no skewness, no kurtosis), conditions B through E are slightly nonnormal, and from condition F to K nonnormality further increases. See Hoogland (1999) for further details.

The procedure by which variables with specific skewness and kurtosis were generated is that of Vale and Maurelli (1983), as implemented in the EQS program (Bentler, 1995) that was also used for model estimation.

The sample size of the generated samples was $N = 200, 400, 800, 1600$. In special cases for ADF a sample size as large as $N = 4500$ was used. In the

TABLE 1.3. Distributional conditions A through K.

	Skewness γ			Kurtosis κ		
DC	min.	mean	max.	min.	mean	max.
A						
B				$-.5$	$-.5$	$-.5$
C	$-.4$	$-.6$	$-.8$			
D				1	1	1
E				-1	-1	-1
F				2	2	2
G	.4	.6	.8	2	2	2
H	-2		2	-1	3.5	8
I				6	6	6
J				2	6	8
K	1.2	1.6	2	6	6	6

Source: Hoogland (1998, Table 5.1, p. 69).
Note: A blank entry means zero.

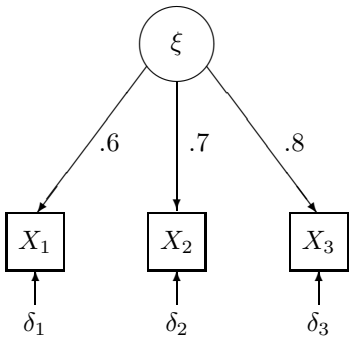

FIGURE 1.1. A single factor with three indicators in Model 35L7.

Monte Carlo study $NR = 300$ admissible replications were used throughout.

Both factor models and structural models were studied. The factor models varied by three characteristics: (1) the number of indicators per factor, (2) the number of latent variables, and (3) the size the factor loadings λ. Hence, model complexity (number of estimated parameters t) and the number of degrees of freedom (df) varied implicitly. In this chapter, results are mainly illustrated for a typical measurement model, namely Factor Model 35L7. The research conclusions in this chapter are drawn generally, however, based on results obtained for all models that were investigated.

Factor Model **35L7** represents a population model with **3** indicators for each of the **5** correlated factors and an average factor **L**oading, $\overline{\lambda}$, of 0.**7**. A single factor and its indicators from this five-factor model is shown in Figure 1.1; each factor in Model 35L7 is similar, and the correlations, ϕ, among factors are all equal to 0.3. For identification purposes the variances of the latent variables were standardized to one. Hence, Model 35L7 has $k = 15$ observed variables, the number of free parameters $t = 40$, and $df = 80$. This is a relatively large model.

4 Nonconvergence and Heywood Cases Revisited

The four estimation methods clearly differ in the probability by which nonconvergent and improper solutions are obtained under different distributional conditions and for different sample sizes. Table 1.4 shows results for Factor Models 35L7 and 35L5 (three indicators for each of the five 0.3 correlated factors with an average factor loading of 0.5) for $N = 200$.

Clearly, ML and ERLS give the least problems; GLS is worse than ML regarding IS. For a sample size as common as $N = 200$, ADF is a disaster with more than one-third of the solutions being improper; and for some factor models percentages even raise to 40% IS if $N = 200$. It is noted that

TABLE 1.4. Percentages of nonconvergence (NC) and improper solutions (IS); Factor Models 35L5 and 35L7, $N = 200$, $NR = 300$.

	Model 35L5						Model 35L7					
	ML/ERLS		GLS		ADF		ML/ERLS		GLS		ADF	
DC	NC	IS	NC	IS	NC	IS	NC	IS	NC	IS	NC	IS
A	.4	5.0	.2	10.0	5.2	35.4		.3		1.2		10.7
B	.2	4.8	.2	10.1	4.8	33.3				.9		9.6
C		5.2	.2	10.4	3.1	32.9				1.2		11.2
D	.2	4.9	.2	9.4	4.1	34.8		.3		1.4	.3	12.7
E	.4	4.1		10.0	2.8	30.8				.6		7.1
F	.4	4.6	.4	8.0	5.0	35.8		.3		1.7	.9	12.6
G		6.2		9.9	3.9	36.0		.3		1.2	.3	11.1
H		6.4		8.4	2.9	35.2		.6		1.1	.3	13.5
I	.3	4.7	.2	8.5	4.7	40.5		.5		1.4	1.4	17.1
J	.2	5.3		9.0	5.5	38.3		.8		1.6	1.9	17.4
K		8.1	.4	13.3	2.9	33.1		.6		1.1	.9	12.5

Note: A blank entry means zero.

their are hardly any effects of the degree of nonnormality on NC and IS. For larger sample sizes the problems are negligible, except for ADF.

If the results for Models 35L5 ($\overline{\lambda} = 0.5$) and 35L7 ($\overline{\lambda} = 0.7$) are compared (see Table 1.4), it is clear that NC and IS both decrease with increasing values of factor loadings, or increasing population covariances σ_{ij}.

The importance of the factors studied for their effect on the occurrence of NC and IS are recapitulated as follows.

- **Sample size.** As expected, a larger sample size N gives less NC and IS, for all estimation methods. For $N > 200$ there were hardly any problems, except for ADF as far as Heywood cases are concerned (see Hoogland, 1999, p. 97). For $N \geq 400$ the percentage of IS is at most 1.6% for ADF.
- **Size factor loadings**. As expected, models with larger loadings have less NC and IS, for all estimation methods.
- **Number of indicators per factor**. As expected, models with a larger NI/NF ratio have less NC and IS, for all estimation methods.
- **Model complexity**. The more parameters to be estimated, i.e., the larger t, the more NC and IS will occur. This was expected from an information point of view: if the estimation requirements are larger, more information is needed. Thus, if the sample size and the factor loadings are about the same, more trouble is expected regarding NC and IS as model complexity increases. However, no strong effect was found. For ML it was only relevant in DC K. For ADF the effects are notably, however, and

mainly due to the fact that increasing t is associated with an increased number of observed variables k, and the latter affects the size of the weight matrix $\mathbf{W}^{-1}$. Due to larger unstability in ADF estimation of $\mathbf{W}^{-1}$ with increasing k, the number of NC/IS problems increases as well.

- **Degree of nonnormality**. This factor has no systematic effects on NC and IS, except for ADF. For the latter method it holds that when the kurtosis increases IS increases. As for explanations, it should be realized that ADF requires estimates of 4th order moments, and standard errors of sample moments are a function of moments of population density functions and sample size (cf. Kendall & Stuart, 1958, p. 243).

5 Bias of Estimators and Minimum Sample Size

Consider the following question, which is implicitly of major practical interest. What is an acceptable size of the bias of estimators of parameters, of corresponding standard errors and of model fit? There is no easy and unequivocal answer to this question. First, criteria for acceptable bias are subjective, although in the literature some conventions are noticeable. Second, specific amounts of bias might be judged differentially disturbing for dissimilar types of parameters. The criteria that were used in defining acceptable bias are given below; see Hoogland (1999, p. 30ff.) for details on the foundations of the decisions involved.

Bias of Parameter Estimators
The relative bias of an estimator $\hat{\theta}_j$ for population parameter θ_j is defined as

$$B(\hat{\theta}_j) = \frac{\overline{\hat{\theta}}_j - \theta_j}{\theta_j} \ , \quad j = 1, 2, \ldots, t \ , \tag{1.1}$$

where $\overline{\hat{\theta}}_j$ is the mean of the parameter estimates over 300 replications.

In comparing the bias of parameter estimators the criterion of the *mean absolute relative bias* (MARB) was used. The MARB of a parameter estimator should be less than 0.025; in formula

$$\text{MARB}(\hat{\theta}_j) = \frac{1}{t} \sum_{j=1}^{t} \mid B(\hat{\theta}_j) \mid \ < \ 0.025 \ , \quad j = 1, 2, \ldots, t \ . \tag{1.2}$$

Bias of Standard Error Estimators
The relative bias of estimators for the standard error of parameter estimates, $\hat{se}(\hat{\theta}_j)$, is defined as

$$B[\hat{se}(\hat{\theta}_j)] = \frac{\overline{\hat{se}}(\hat{\theta}_j) - \text{SD}(\hat{\theta}_j)}{\text{SD}(\hat{\theta}_j)} \ , \quad j = 1, 2, \ldots, t \ , \tag{1.3}$$

where $\overline{\hat{se}}(\hat{\theta}_j)$ is the mean of the estimated standard errors and $\mathrm{SD}(\hat{\theta}_j)$ is the standard deviation of the parameter estimates, both calculated over $NR = 300$ replications.

The criterion for an acceptable bias of standard error estimators for $\hat{\theta}_j$ was that the *mean absolute relative bias* of the estimators should be less than 0.05, i.e.,

$$\mathrm{MARB}[\hat{se}(\hat{\theta}_j)] = \frac{1}{t}\sum_{j=1}^{t} \mid B[\hat{se}(\hat{\theta}_j)] \mid \; < \; 0.05 \; , \quad j = 1, 2, \ldots, t \; . \tag{1.4}$$

Bias of the Chi-Square Test Statistic

To evaluate the behaviour of the chi-square model test statistic T, which equals $(N-1)$ times the minimum value of the fit function of the model, two criteria are used in this chapter. [Other criteria, not reported here, were employed as well; see Hoogland (1999, p. 31f., p. 62f.) for details.]

First, the *rejection frequency* (RF) of the model over $NR = 300$ replications, given a significance level $\alpha = 0.05$. Given NR and α, this rejection frequency has a binomial distribution, i.e., $RF \sim \mathrm{Bin}(NR, \alpha)$. A 99% prediction interval for RF, in our case [7, 27], was used as a criterion for acceptable behaviour of the fit statistic T. For adequate behaviour of T, the observed value of the RF should lie within this prediction interval.

Second, the *mean value of chi-square test statistic* over $NR = 300$ replications, denoted as $\overline{T}$. Using a Student t test statistic,

$$t^* = \frac{(\overline{T} - df)\,\sqrt{NR-1}}{\mathrm{SD}(T)} \; , \tag{1.5}$$

the bias of the chi-square test statistic T was taken to be acceptable if the null hypothesis $H_0 : E(T) = df$ is not rejected at $\alpha = 0.01$.

Required Minimum Sample Size for Acceptable Bias

Another important practical question, for applied researchers and statistician alike, is: What is the minimal sufficient sample size required to stay within acceptable ranges of bias for estimates of parameters, corresponding standard errors and model fit? The procedures that were followed to come to conclusions and recommendations regarding this question are too complex to be summarized in a few lines; they were described in detail by Hoogland (1999).

In the following sections guidelines are given with respect to the minimum sample size required for obtaining acceptable bias of estimators according to the criteria defined above. It should be noted here, that although the emphasis in this chapter is on bias of the estimators, their variances and mean squared errors (MSEs) were also considered. Where necessary, attention is also paid to these statistical properties of SEM estimators.

6 Parameter Estimators

Bias of Parameter Estimators

Four estimators were compared on the minimum sample size needed for an acceptable bias according to the mean absolute relative bias (MARB) criterion, as defined by (1.2). The results are summarized in Table 1.5. In the first column of this table the distributional conditions are categorized by degree of nonnormality, which is additionally quantified by ranges of $\overline{\overline{\kappa}}_a$ over sample size N, the mean of the average empirical kurtosis over 15 variables and 300 admissible replications, respectively. For the nonnormal distributional conditions I, J, and K, the values of $\overline{\overline{\kappa}}_a$ calculated over $k = 15$ variables and $NR = 300$ replications lie in the range $[4.5,\ 5.8]$, which is indicative of large positive kurtosis.

The numbers in this table – and in similar tables to follow – are the minimum sample size needed to obtain acceptable bias according to (1.2). Boldface values of N indicate absence of bias for that size; otherwise the sign to the right of the minimum required N denotes whether the bias is negative or positive for that size. From Table 1.5 is can be seen that ML and ERLS do well for $N \geq 200$, even under conditions of severe nonnormality. GLS and ADF underestimate parameters for sample sizes $N \geq 800$ (although results were quite model dependent). There were hardly any effects of nonnormality on the bias of parameter estimators, except for ADF, as illustrated below. In general it was found that the bias of ML decreases with a larger NI/NF ratio.

Figure 1.2 illustrates the effect of sample size on the MARB for Model 35L7 under the nonnormal, kurtotic condition K. It can be concluded that the MARB drops linearly with $1/\sqrt{N}$. ADF and GLS clearly behave worst. In general, considering the results for all models under study, ADF shows a strong effect of $\overline{\kappa}$ on its relative bias. For Model 35L7, to achieve acceptable bias in parameter estimation ADF would require $N = 3600$ for condition K.

Sample Size and Acceptable Bias for ADF

Based on all empirical findings from our research, some general conditions for almost unbiased $\hat{\boldsymbol{\theta}}_{ADF}$ estimators can be formulated. Within the design of our Monte Carlo study – that is, not considering distributional conditions

TABLE 1.5. Sufficient sample size for acceptable bias of parameter estimators; Factor Model 35L7, $NR = 300$.

DC	$\overline{\overline{\kappa}}_a$	ML	GLS	ERLS	ADF
A, B, C	−0.5, 0.0	**200**	800 −	**200**	800 −
D	0.9, 1.0	**200**	800 −	**200**	1600 −
E	−1.0, −1.0	**200**	800 −	**200**	800 −
F, G, H	1.7, 3.4	**200**	800 −	**200**	1600 −
I, J, K	4.5, 5.8	**200**	800 −	**200**	>1600 −

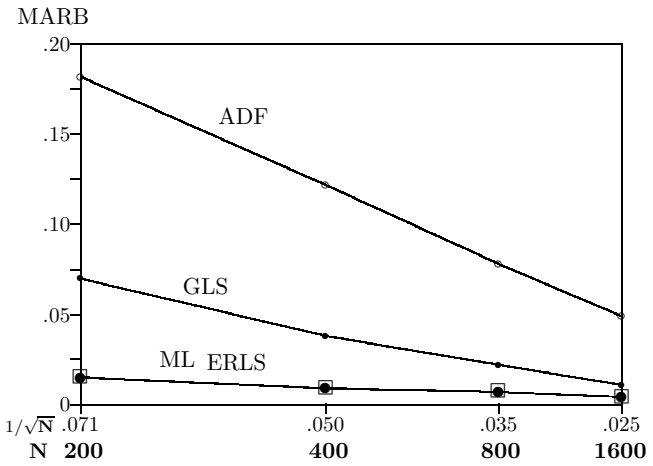

FIGURE 1.2. The MARB of parameter estimators under nonnormal condition K with $\overline{\gamma} = 1.6$, $\overline{\kappa} = 6.0$; Factor Model 35L7, $NR = 300$.

outside the range A through K (see Table 1.3), nor generalizing to any other DCs – for a given range of mean kurtosis $\overline{\kappa}$ among the investigated DCs, the required N for ADF is a linear function of the number of variables k. General requirements for acceptable parameter bias can be summarized, provisionally, as follows (cf. Hoogland, 1999, p. 140):

$$\begin{aligned} &\text{if} \quad -1 \leq \overline{\kappa} \leq 0 \quad \text{then} \quad N \geq 50k \ , \\ &\text{if} \quad\ \ 0 < \overline{\kappa} \leq 3 \quad \text{then} \quad N \geq 100k \ , \\ &\text{if} \quad\ \ 3 < \overline{\kappa} \leq 6 \quad \text{then} \quad N \geq 250k \ . \end{aligned}$$

These guidelines should be interpreted cautiously, because – apart from the kurtosis – they are formulated rather unconditionally; presently, the authors are studying their general validity in more detail, which may well lead to slightly refined recommendations. Nevertheless, the implications of these guidelines reaffirm the long-known fact that ADF needs very large sample sizes. What was not known as clearly, is that increasingly more observations are needed with increasing normality violations. The required sample size is a function of kurtosis and model complexity. Therefore, it is concluded that with a finite sample size the ADF estimator is not free from distributional effects! It is only an *asymptotically* distribution-free method.

Variance and Mean Squared Error of Parameter Estimators
Besides the bias, the variance and the mean squared error (MSE) of estimators are of statistical interest. The standard deviation (SD) of parameter estimators, for individual parameters denoted as $\mathrm{SD}(\hat{\theta}_j)$, was simultaneously

TABLE 1.6. Mean standard deviation of estimators: $100 \times \overline{\mathrm{SD}}(\hat{\boldsymbol{\theta}})$; Factor Model 35L7, $N = 200$, $NR = 300$

	Distributional Condition										
	A	B	C	D	E	F	G	H	I	J	K
ML	8	8	8	8	8	9	9	9	10	10	11
ADF	11	11	11	11	10	12	12	11	13	13	12

estimated for different types of model parameters. For all parameters $\boldsymbol{\theta}$ taken together, it was calculated as the mean SD of parameter estimates over t parameters in $NR = 300$ replications, and denoted as $\overline{\mathrm{SD}}(\hat{\boldsymbol{\theta}})$. Findings for ML and ADF, regarding all parameters, are given in Table 1.6. The results were about the same for separate types of parameters.

The most important finding was that the average SD of $\hat{\boldsymbol{\theta}}$ for ML, GLS and ERLS turns out to be *equal up to two decimal places* for different types of parameters; ADF being close to that result. Thus, hardly any differences between estimators were found. In addition, it is noteworthy that only a small effect of nonnormality was observed; see Table 1.6.

It is concluded that the parameter estimators differ in bias but not in variance. It is therefore unnecessary to compare their MSEs.

7 Robust Standard Errors and Fit Statistics

Before turning to standard errors and model fit statistics, an intermezzo on robust alternatives for regular – but not robust – estimators is necessary, because in the following sections regular (uncorrected) and robust (corrective) estimators are being compared as well. The following two examples illustrate the need for robust estimators.

First, it is known from the literature (e.g., Boomsma, 1983) that ML estimators of standard errors and global model fit are not robust against nonnormality. So the question arises: What actions can be taken in the presence of nonnormal data?

Second, in the previous section the bad behaviour of the asymptotic distribution-free method was exemplified. It is also known for at least fifteen years that ADF does not perform well when models are complex and when the sample size is small; see, e.g., Muthén and Kaplan (1985, 1990). The main reason for that is that ADF estimation involves the estimation and inversion of a matrix of 4th-order sample moments, resulting in some weight matrix $\mathbf{W}^{-1}$. If that matrix is large, depending on k, and if N is small, the estimated weight matrix is unstable, with the painful consequence that parameter estimates, and even more so the estimates of standard errors and model fit, are unreliable to an unacceptable degree.

Some analytical and statistical relief for the nonrobustness of ML against nonnormality and for the nonrobustness of ADF against small sample size, is offered by the development of robust (asymptotic) inferential statistics, and of what might be called robust, corrective statistics.

- *Robust (asymptotic) inferential statistics* do not require 4th-order sample moments and therefore do not suffer from the pitfalls of having an unstably estimated weight matrix. Also, under specific conditions of *stochastic independence*, these robust estimation methods produce asymptotic standard errors and model test statistics which are valid for any distribution of the data. This leads to the almost paradoxical consequence that normal ML theory works well under nonnormality. Reviews of this type of work, that has a long history and broad backgrounds, were given by Satorra (1990, 1992).
- *Robust, corrective statistics*, for example robust standard error estimators and robust model test statistics are supposed to be more robust against violations of the assumptions of large sample size and nonnormality than regular estimators. An overview of these robust, corrective statistics was given by Satorra and Bentler (1994); corresponding LISREL formulas can be found in Jöreskog, Sörbom, Du Toit, and Du Toit (1999).

In this chapter three robust statistics are considered that might help to improve regular estimates of standard errors or model fit. Notice that they all do require estimates of 4th-order moments of observed variables.

1. **RML**. Robust ML standard errors, which were calculated under nonnormality assumptions with the EQS program (Bentler, 1995); see Browne (1984), or Bentler and Dijkstra (1985) for theoretical details.
2. **SML**. The Scaled ML test statistic T_{SB} of Satorra and Bentler (1988, 1994), which has asymptotically a correct mean. Earlier research on small models showed promising behaviour of this T_{SB} statistic (see, for example, Satorra & Bentler, 1988; Chou, Bentler, & Satorra, 1991; Hu, Bentler, & Kano, 1992; and Curran, West, & Finch, 1996).
3. **YBA**. Yuan and Bentler's (1997) corrected ADF test statistic T_{YB}. So far, the statistical properties of this statistic have not been supported by much additional empirical research.

8 Standard Error Estimators

Bias of Standard Error Estimators

A comparison of the bias of standard error estimators in Model 35L7 for $N = 800, 1600$ is given in Table 1.7. It can be seen that there are no

TABLE 1.7. MARB of standard error estimators: 100 × MARB of $\hat{se}(\hat{\theta}_j)$; Factor Model 35L7, $N = 800, 1600$, $NR = 300$.

	$N = 800$					$N = 1600$				
DC	ML	RML	GLS	ERLS	ADF	ML	RML	GLS	ERLS	ADF
A	3	3	4	3	14	3	3	3	3	7
B	4	3	4	3	14	5	3	4	4	7
C	4	3	5	4	15	5	3	5	4	8
D	7	4	7	5	14	6	3	7	4	7
E	6	3	6	5	14	7	3	7	6	7
F	11	4	11	7	14	11	3	11	7	7
G	12	4	13	7	14	12	3	12	7	7
H	14	4	14	14	15	13	3	13	14	8
I	24	5	24	16	16	24	4	24	16	8
J	24	5	24	15	16	23	4	23	15	8
K	28	4	28	15	17	27	4	27	15	9

large differences between the traditional uncorrected ML, GLS, and ERLS (DCs A through E). ADF performs relatively poorly in small samples if nonnormality is small (DCs A through G); for $N = 1600$ ADF outperforms ML if nonnormality is large. It should be noted that for $N = 800$ the differences between ADF and other methods were not very large, but for $N = 200, 400$ ADF standard error estimators were rather bad indeed.

Of specific interest is the comparison of ML and its robust counterpart RML. From Table 1.7 it is clear that RML is an improvement over ML – and the other methods – for DCs F through K, by reducing the MARB of the standard error estimator considerably.

In Figure 1.3, for the kurtotic distributional condition K the MARB of five standard error estimators is plotted as a function of sample size N.

Within the research framework, it can generally be concluded that the bias of RML is overall smallest if $N \geq 400$ and $|\overline{\kappa}| \geq 1$. The RML estimator is quite robust against nonnormality. A second general conclusion is that the behaviour of ADF is relative good if $N \geq 1600$ (although results were quite model dependent). Figure 1.3 shows a clear effect of sample size for ADF: the MARB is decreasing linearly with $1/\sqrt{N}$. For an acceptable bias of estimated standard errors in Model 35L7, however, ADF still would need $N = 4500$ observations (cf. Hoogland, 1999, Table 6.6).

The question to be answered now is what a sufficient sample size is for standard error estimators to have an acceptable bias according to the MARB criterion (1.4). In Table 1.8 the results for Model 35L7 are given, from which it can be concluded that RML outperforms all other estimators, except ML and ERLS under pretty normal conditions. Still, with increasing nonnormality some hundreds of observations are needed for RML.

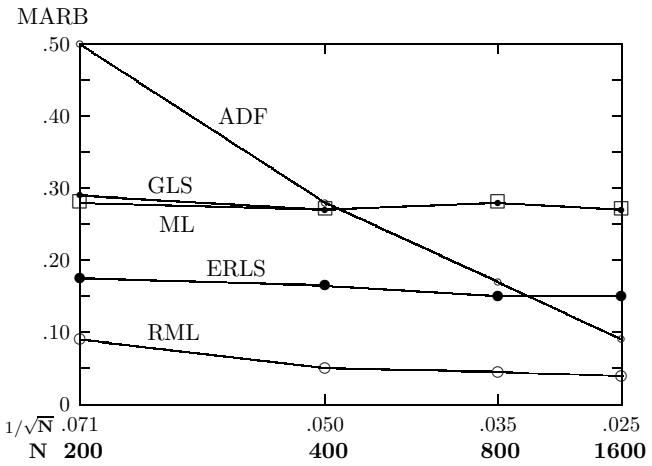

FIGURE 1.3. The MARB of standard error estimators under nonnormal condition K with $\overline{\gamma} = 1.6$, $\overline{\kappa}_a = 6.0$; Factor Model 35L7, $NR = 300$.

GLS (nonnormal DCs) and ADF (overall) need very large number of observations. General rules of thumb – applicable within our research design – for a required minimum sample size for an almost unbiased $\hat{se}(\hat{\theta}_j)$ with ADF are: if $\overline{\kappa} = 0$ then $N \geq 10k(k+1)$, and if $0 < \overline{\kappa} < 5.7$ then $N \geq 15k(k+1)$, but under all circumstances $N \geq 400$ is required.

The conclusions regarding the bias of standard error estimators can be summarized as follows.

- **'Normal' conditions** ($|\overline{\kappa}| < 0.5$). ML, RML, and ERLS are acceptable at a minimum sample size $N = 400$, if $\overline{\lambda} \geq 0.7$. ADF needs a much larger sample size than ML. Smaller population loadings require larger N.

TABLE 1.8. Sufficient sample size for acceptable bias of standard error estimators; Factor Model 35L7, $NR = 300$.

DC	$\overline{\overline{\kappa}}_a$	ML	RML	GLS	ERLS	ADF
A, B	−0.5, 0.0	**200**	**200**	400 −	**200**	⊘ −
C	0.0, 0.0	400 −	**200**	⊘ −	**200**	⊘ −
D	0.9, 1.0	⊘ −	400 −	⊘ −	400 −	⊘ −
E	−1.0, −1.0	⊘ +	**200**	⊘ +	⊘ ±	⊘ −
F, G	1.7, 2.0	⊘ −	400 −	⊘ −	⊘ −	⊘ −
H	2.7, 3.4	⊘ −	400 −	⊘ −	⊘ ±	⊘ −
I, J, K	4.5, 5.8	⊘ −	800 −	⊘ −	⊘ −	⊘ −

Note: ⊘ a sample of size $N > 1600$ is required.

- **Nonnormal conditions** ($|\overline{\kappa}| \geq 1$). The RML estimator is preferred, but it needs at least $N \geq 400$. For other estimators, including ADF, $N = 1600$ is insufficient.
- **All conditions** (DCs A through K, that is). Underestimation of $\hat{se}(\hat{\theta}_j)$ is expected when N is too small ($\kappa > 0$). Inflation of $\hat{se}(\hat{\theta}_j)$ is expected if $\kappa \leq -1$; distributional conditions with large negative kurtosis were outside the range of the research design.

  Most frequently, progressive testing of model parameters is expected to occur when $\kappa > 0$, i.e., the null hypothesis $H_0 : \theta_j = 0$ is rejected too often. Since parameter estimation is often unbiased, users should be encouraged to look primarily at the substantive relevance of the estimated values of model parameters.

Variance of Standard Error Estimators

When comparing ML with its robust counterpart RML in estimating standard errors of parameter estimates, the variance and the mean squared error (MSE) of these estimators were also investigated. RML did not improve as fast over ML as when only bias is considered. This is due to the fact that the variance of individual RML parameter estimators can be five times larger than that of ML estimators. In Table 1.9 the average MSE of ML and RML standard error estimators are compared. The mean MSE is calculated over all $t = 40$ parameters $\boldsymbol{\theta}$. The numbers in the table are $10,000$ times the difference $\overline{\text{MSE}}[\hat{se}(\hat{\boldsymbol{\theta}}_{\text{ML}})] - \overline{\text{MSE}}[\hat{se}(\hat{\boldsymbol{\theta}}_{\text{RML}})]$. Hence, it can be noted that the actual MSEs are relatively small, and so are the differences between the two estimators. For instance, for $N = 200$ and normal condition A, the value of $-.4$ in Table 1.9 reflects the actual difference in mean MSE over 40 model parameters: $0.00006 - 0.00010$.

From Table 1.9 it is obvious that regarding all parameters, the differences in mean MSE depends on the sample size N and the degree of nonnormality.

TABLE 1.9. Differences in average MSE of the ML and RML standard error estimators over all parameters: $10,000 \times (\overline{\text{MSE}}_{\text{ML}} - \overline{\text{MSE}}_{\text{RML}})$; Factor Model 35L7, $NR = 300$.

	Distributional Condition										
N	A	B	C	D	E	F	G	H	I	J	K
200	$-.4$	$-.2$	$-.4$	$-.6$		$-.8$	$-.8$	-1.4		$-.2$	2.3
400	$-.1$		$-.1$	$-.1$	$-.1$	$-.1$	$-.1$	$-.1$	1.4	1.2	2.3
800						.1	.1	.3	1.3	1.1	2.0
1600						.1	.1	.3	.9	.8	1.2

Note: A blank entry means zero.

With increasing N and increasing nonnormality, RML behaves increasingly better relative to ML. In terms of $\overline{\text{MSE}}$, for $N = 200$ RML is better than ML for DC K only; for $N = 400$ it is better for DCs I through K; and for $N = 1600$ it is better for all distributional conditions. In conclusion, RML has to be preferred over ML when serious deviations from normality occur and N is not too small.

9 Chi-Square Model Test Statistics

In the comparison of estimators for the chi-square model fit statistic T, the model rejection frequency RF, the bias of T, and its variance are examined subsequently. The relevant statistics were defined in Section 5.

Model Rejection Frequency
From the results in Table 1.10 it can be observed that under *normal conditions* for $N = 200$, (1) GLS and YBA are conservative estimators, (2) ADF is very progressive (rejecting the correct model far too often), and (3) YBA gives an adequate correction to ADF. With the larger sample size $N = 1600$, (1) SML is not better than ML, nor is it for $N = 200$, and (2) YBA is still better than ADF, in general showing slight overrejection.

In contrast, under *nonnormal conditions* (see the bottom panel of Table 1.10), the following conclusions can be drawn. First, SML shows appropriate behaviour (progressive if $N = 200$), and the same holds for YBA (conservative if $N = 200$). Second, ADF strongly improves with increasing N. Third, ML and GLS have a larger rejection frequency when (a) the model has larger loadings, and (b) there are more indicators per factor.

TABLE 1.10. The chi-square model rejection frequency at level $\alpha = 0.05$: % Reject − 5; normal condition A (top panel), and nonnormal condition K with $\overline{\gamma} = 1.6$, $\overline{\kappa} = 6.0$ (bottom panel), $N = 200, 1600$, $NR = 300$

DC A	$N = 200$						$N = 1600$					
Model	ML	SML	GLS	ERLS	ADF	YBA	ML	SML	GLS	ERLS	ADF	YBA
34L5	−2	1	−3	−2	28	−4	1	2	2	1	7	2
44L5	3	4	−2	−1	94	−4	1	1		−1	12	1
35L5	−3	1	−4	−3	83	−5	2	1	2		5	1
35L7	3	6	−1	2	89	−3	1	2	1	1	7	1

DC K	$N = 200$						$N = 1600$					
Model	ML	SML	GLS	ERLS	ADF	YBA	ML	SML	GLS	ERLS	ADF	YBA
34L5	8	1	1	−5	24	−5	6	1	6	−5	3	2
44L5	15	5	3	−4	95	−5	12	1	10	−5	8	1
35L5	10	3	−1	−5	82	−5	9	1	7	−5	6	−1
35L7	24	5	6	−5	88	−4	23	3	20	−5	6	

Note: A blank entry means zero.

TABLE 1.11. Sufficient sample size for acceptable rejection frequency at level $\alpha = 0.05$; Factor Model 35L7, $NR = 300$

DC	ML		SML		GLS		ERLS		ADF		YBA	
A	**200**		400	+	**200**		**200**		⊘	+	**200**	
B	**200**		400	+	**200**		400	+	⊘	+	400	−
C	1600	+	400	+	**200**		**200**		⊘	+	**200**	
D	**200**		400	+	**200**		1600	−	⊘	+	400	−
E	400	+	400	+	**200**		1600	+	⊘	+	**200**	
F	400	+	400	+	**200**		⊘	−	⊘	+	400	−
G	800	+	400	+	**200**		⊘	−	⊘	+	**200**	
H	**200**		800	+	**200**		⊘	−	⊘	+	400	−
I	⊘	+	800	+	⊘	+	⊘	−	⊘	+	800	−
J	⊘	+	800	+	**200**		⊘	−	⊘	+	400	−
K	⊘	+	400	+	⊘	+	⊘	−	⊘	+	800	−

Note: ⊘ a sample of size $N > 1600$ is required.

Of course, where the whole distribution of T is involved, it is limited to consider only the behaviour of this statistic for overall model fit at a significance level at $\alpha = 0.05$, but in practice that tail area is still the most inspected one. Despite this restriction, next to other criteria the RF criterion was used to decide what a minimal sufficient sample size would be, i.e., the RF should be within a 99% prediction interval (see Section 5). On that limited basis first, some general conclusions can be drawn. The results for Model 35L7 are given in Table 1.11.

In general then, SML gives an appropriate correction to ML under conditions of nonnormality ($\overline{\kappa} = 6$ for DCs I, J, and K), but overall it needs at least $N \geq 400$. The correction of YBA on ADF is effective and suitable throughout; in general, however, ADF would need at least $N = 3600$ for

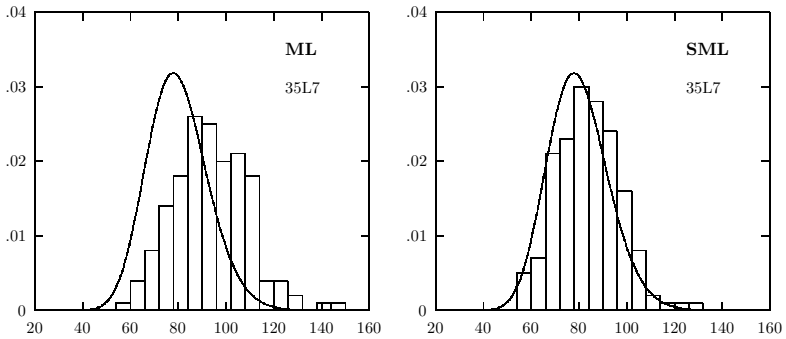

FIGURE 1.4. The distribution of the model fit statistic for ML and SML under nonnormal condition K with $\overline{\gamma} = 1.6, \overline{\kappa} = 6$; Factor Model 35L7, $df = 80$ $N = 400, NR = 300$.

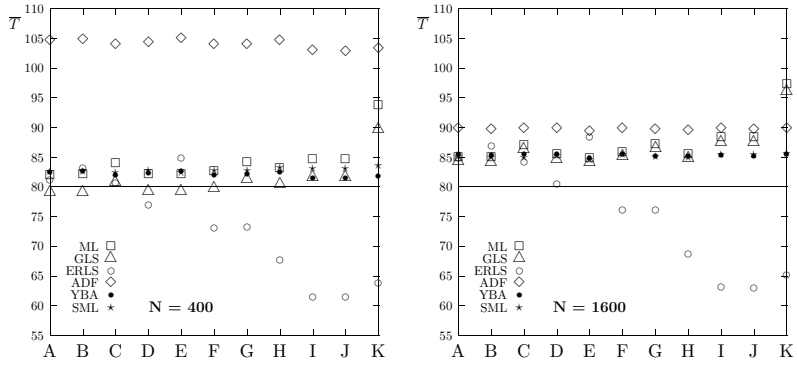

FIGURE 1.5. Mean value of model test statistic T; Factor Model 35L7, $df = 80$, $N = 400, 1600$, $NR = 300$.

$df = 80$; cf. Hoogland (1999, Table 6.7, p. 143). GLS behaves remarkably well, although results are rather model dependent (cf. Olsson, Troye, & Howell, 1999, for similar findings). Finally, ERLS is not robust at all, and not only with regard to the RF, as will be seen next.

The corrective effect of the scaled test statistic T_{SB} on the ordinary ML test statistic T is illustrated in Figure 1.4. The shift in the distribution to the correct mean value under nonnormal condition K is clearly visible.

Bias of the Chi-Square Model Test Statistic

In addition to an inspection of the right tail of the distribution of the chi-square model fit statistic T for different estimation procedures, it is of interest to check the expected value of T, or its bias. In Figure 1.5 it can be noticed that ERLS is severely underestimating the expected value of $df = 80$ under nonnormal conditions, and increasingly so with larger kurtosis. In contrast, ADF is overestimating the expected value, but improves rapidly as N gets larger.

Variance of the Chi-Square Model Test Statistic

Apart from the bias of the model fit statistic T, its variance was studied. For Model 35L7 the standard deviation of six different estimators are shown in Figure 1.6. Under the correct model the expected standard deviation $E[\mathrm{SD}(T)] = \sqrt{160} = 12.65$. In general it can be observed that ADF has too large a variance, and YBA slightly too small. Here, a reference is made to the idea of the adjusted test statistic $\overline{\overline{T}}$, introduced by Satorra and Bentler (1994, p. 408) to attain, asymptotically at least, both a correct mean and a correct variance. Finally, it was observed that for large deviations from normality (DCs I, J, and K) the variance of SML is slightly smaller than that of ML.

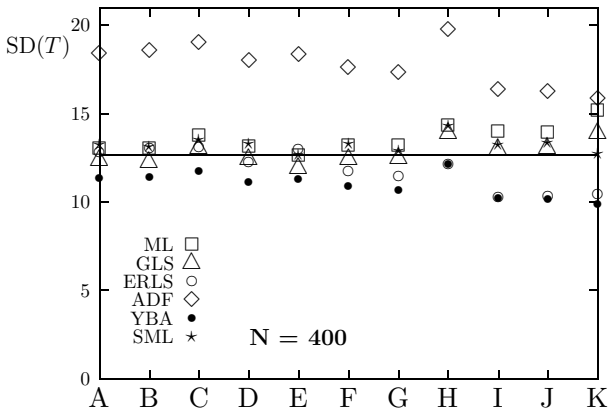

FIGURE 1.6. Standard deviation of model test statistic T; Factor Model 35L7, $df = 80$, $N = 400$, $NR = 300$.

Minimum Sample Size Required

For the investigation of the minimal sample size necessary for T to have an acceptable bias, the criterion of not rejecting $H_0 : E(T) = df$ was applied. In Table 1.12 results are presented for Model 35L7. In general it was concluded that SML is acceptable for $N \geq 800$ (overestimation), but it still needs $N = 1600$ under nonnormality. Also, ADF needs at least $N = 3600$ for $df = 80$, and YBA is acceptable at $N = 1600$ (overestimation). The required sample sizes are overall quite large, which may be partly due to the substantial power of the criterion test, given $NR = 300$ replications. It should also be emphasized that these sufficient sample size results were very model dependent.

TABLE 1.12. Sufficient sample size for acceptable mean value of model test statistic T; Factor Model 35L7, $NR = 300$.

DC	ML		SML		GLS		ERLS		ADF		YBA	
A	800	+	800	+	400	−	**200**		⊘	+	1600	+
B	800	+	800	+	400	−	⊘	+	⊘	+	1600	+
C	⊘	+	800	+	**200**		**200**		⊘	+	800	+
D	800	+	1600	+	400	−	⊘	−	⊘	+	1600	+
E	800	+	800	+	400	−	⊘	+	⊘	+	1600	+
F	1600	+	1600	+	400	−	⊘	−	⊘	+	1600	+
G	⊘	+	1600	+	1600	+	⊘	−	⊘	+	1600	+
H	1600	+	800	+	400	−	⊘	−	⊘	+	1600	+
I, J, K	⊘	+	1600	+	⊘	+	⊘	−	⊘	+	1600	±

Note: ⊘ a sample of size $N > 1600$ is required.

10 Discussion

In this chapter, some well-known facts were affirmed: (1) in samples of size $N \leq 200$, under both normal and nonnormal conditions, problems of nonconvergence and improper solutions still exist, (2) strong measurement instruments, both in terms of reliability and validity, may compensate for small sample size to reduce the number of nonconvergent and improper solutions, and (3) parameter estimation is least problematic, whereas estimation of standard errors and the chi-square test statistic for global model fit is of major concern when sample size is small and, more seriously, under violation of normality assumptions.

It was observed anew that ML is sensitive to violations of nonnormality, but here other estimation methods also turned out to be nonrobust. It was affirmed that ERLS was bad in estimating model fit, and that ADF is very demanding on N. Where GLS was known to behave rather well in small models, in larger models its behaviour was much worse; for example, it needs a much larger sample size for acceptable behaviour regarding the bias of parameter estimates than ML. It was partly known that robust, corrective statistics, like RML, SML, and YBA work rather well; such results were again found for the larger models that were studied here.

The key objective of robustness research is to offer practical guidelines for applied work, so as to prevent nonrobust analyses that would inevitably lead to wrong substantive inferences. Within that framework, a predominant question is: If structural models are to be analyzed, what estimation methods have to be preferred under what conditions? And one of the first specific questions in the planning of SEM research is: What is a minimum sample size needed for a precise and reliable analysis, conditional on the data and the model?

It was shown that answers to these questions are conditional on data and model characteristics alike. In practice, however, applied researchers often do not know the data and the model characteristics before data collection and analysis. In new areas of applied research, especially when measurement instruments are in a developing stage, little is known about distributional characteristics of observed variables. Also, in phases of model exploration there are uncertainties about the complexity of the 'final models,' about the number of reliable indicators and the size of factor loadings. It is evident that with better measurements and stronger theoretical foundations of model structures, it becomes much easier to make proper decisions on the choice of estimators and the planning of sample size.

Still, to advice applied statisticians on sample size and the choice of estimators under conditions of (non)normality much more knowledge and expertise is available than twenty years ago. On the one hand, there has been much theoretical progress, mainly due to the development of new estimators and ingenious robust statistics, including the emergence of general procedures for robust statistical inference. On the other hand, robustness

research has added practical guidelines (1) to accommodate for circumstances of finite, small sample size and nonnormalities, and (2) to discourage structural modeling when insufficient information is available.

Acknowledgments: The authors wish to express their gratitude to the editors for inviting them to make a scientific contribution in honor of Karl G. Jöreskog. They also thank Marijtje van Duijn for her valuable comments on the first draft of this chapter.

References

Bentler, P.M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P.M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P.R. Krishnaiah (Ed.), *Multivariate analysis — VI* (pp. 9–42). Amsterdam: North-Holland.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part I, pp. 149–173). Amsterdam: North-Holland.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation against small sample size and nonnormality*. Amsterdam: Sociometric Research Foundation. (Doctoral dissertation, University of Groningen, The Netherlands.)

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, *52*, 345–370.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Camstra, A. (1998). *Cross-validation in covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Chou, C.-P., & Bentler, P.M. (1995). Estimation and tests in structural equation modeling. In R.H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37–55). Thousand Oaks, CA: Sage.

Chou, C.-P., Bentler, P.M., & Satorra, A. (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*, 347–357.

Curran, P.J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

Hendriks, P. (1999). *Confirmatory factor analysis methods compared: The multiple group method and maximum likelihood confirmatory factor analysis*. Unpublished manuscript, Department of Psychology, University of Groningen, The Netherlands.

Hoogland, J.J. (1998). Robustness of estimators in covariance structure analysis: A Monte Carlo study with a large model. In J.J. Hox & E.D. de Leeuw (Eds.), *Assumptions, robustness, and estimation methods in multivariate modeling* (pp. 67–86). Amsterdam: TT-Publikaties.

Hoogland, J.J. (1999). *The robustness of estimation methods for covariance structure analysis*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367.

Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362.

Jöreskog, K.G., Sörbom, D., Du Toit, S., & Du Toit, M. (1999). *LISREL 8: New statistical features*. Chicago, IL: Scientific Software International.

Kendall, M.G., & Stuart, A. (1958). *The advanced theory of statistics: Vol. 1. Distribution theory*. London: Griffin.

Luijben, T.C.W. (1989). *Statistical guidance for model modification in covariance structure analysis*. Amsterdam: Sociometric Research Foundation. (Doctoral dissertation, University of Groningen, The Netherlands.)

Marsh, H.W., Hau, K.-T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181–220.

Marsh, H.W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R.H. Hoyle (Ed.), *Statistical strategies for small sample size* (pp. 251–306). Thousand Oaks, CA: Sage.

Muthén, B.O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–189.

Muthén, B.O., & Kaplan, D. (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, U.H., Troye, S.V., & Howell, R.D. (1999). Theoretic fit and empirical fit: The performance of maximum likelihood versus generalized least squares estimation in structural equation modeling. *Multivariate Behavioral Research*, *34*, 31–58.

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research*, *13*, 109–119.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity*, *24*, 367–386.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. In P.V. Marsden (Ed.), *Sociological methodology 1992* (pp. 249–178). Oxford: Blackwell.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the ASA* (pp. 308–313). Alexandria, VA: The American Statistical Association.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C.C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Vale, C.D., & Maurelli, V.A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465–471.

Van Driel, O.P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, *43*, 225–243.

Yuan, K.-H., & Bentler, P.M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association*, *92*, 767–774.