# Dimensionality assesment with factor analysis methods

Barendse, Mariska

# Chapter 4

# Structural Equation Modeling of Discrete Data: Model Fit After Pairwise Maximum Likelihood*


## Abstract

*Maximum likelihood factor analysis of discrete data rests on the assumption that the observed discrete responses are manifestations of underlying continuous scores that are normally distributed. As maximizing the likelihood of multivariate response patterns is computationally very intensive, the sum of the likelihoods of the bivariate response patterns is maximized instead. Little is yet known about how to assess goodness–of–fit when the analysis is based on such a pairwise maximum likelihood (PML) of two–way contingency tables. We propose new fit criteria for the PML method and conduct a simulation study to evaluate their performances in model selection. With large sample sizes (500 or more), PML performs as well as the robust weighted least squares analysis of polychoric correlations.*


---

*This Chapter is based on: Barendse, M. T., Ligtvoet, R., Timmerman, M. E. & Oort, F. J., (2014). *Structural Equation Modeling of Discrete data: Model Fit after Pairwise Maximum Likelihood.* Manuscript submitted for publication.

## 4.1 Introduction

Tests and questionnaires usually consist of items with discrete ordinal response scales. In the factor analysis of discrete item responses, multivariate normally distributed item scores are assumed to underly the discrete responses (e.g., Wirth & Edwards, 2007; Rhemtulla et al., 2012).

Let $X = (X_1, X_2, \ldots, X_k)$ denote the vector of the $k$ item score variable with discrete response scales, with realizations $x_i \in \{1, 2, \ldots, m\}$, so that each item $i$ has $m$ response options. The observed score $x_i$ on item $i$ is related to the unobserved score $x_i^*$ on the underlying continuum through

$$X_i = x_i \;\Leftrightarrow\; \tau_{x_i-1} < x_i^* \leq \tau_{x_i}, \tag{4.1}$$

where $\tau_{x_i-1}$ and $\tau_{x_i}$ are the threshold parameters for the category $x_1$ $(x_i = 1, 2, \ldots, m)$ of item $i$, where $\tau_0 = -\infty$ and $\tau_m = \infty$, by definition.

As the underlying continuous item score variable $X_i^*$ is not observed, its mean and variance are not identified without further constraints. One can either fix the mean and variance (e.g., zero mean and unit variance), or one can fix two of the thresholds (e.g., at zero and unity). The latter is not possible with dichotomous items, because they are associated with just a single threshold.

Various estimation methods have been proposed for the factor analysis of (observed) discrete responses with (unobserved) underlying continuous scores. Here we discuss the weighted least squares method, the multivariate maximum likelihood method, and the bivariate maximum likelihood method.

### 4.1.1 Weighted Least Squares Method

The weighted least squares method is a two–step method. In the first step, the polychoric correlations between the observed variables are estimated. In the second step, the parameters of the structural equation model are estimated on the basis of the polychoric correlations. The general WLS fit function for discrete data, based on Browne (1984) who described the WLS fit for continuous data, is given by

$$F_{WLS} = (\hat{q} - g)'W^{-1}(\hat{q} - g), \tag{4.2}$$

where $\hat{q}$ is a vector with the non–redundant elements of the $k \times k$ matrix of polychoric correlations and $g$ is a vector with the corresponding elements of the $k \times k$ matrix of model–implied correlations. The weight matrix $W$ is a positive definite matrix of order $v \times v$, with $v = k(k+1)/2$. It contains consistent estimates of the asymptotic variances and covariances of the polychoric correlations (e.g., Jöreskog, 1990, 1994). Other authors (e.g., Muthén, 1984; Muthén et al., 1997) also included the observed and model implied threshold values in the $\hat{q}$ and $g$ vectors, and the associated $W$ matrix that contains the asymptotic covariances of $\hat{q}$ (Muthén, 1989b). As the weight matrix can only be accurately estimated with large sample sizes (e.g., Rigdon & Ferguson Jr, 1991; Muthén & Kaplan, 1992; Dolan, 1994), it is practically unfeasible to use the WLS function with the full weight matrix. An alternative is to use the WLS function with diagonal matrix $W_D$, containing only the diagonal elements of $W$ to estimate the parameter estimates. However, for inference, one needs the full weight matrix as implemented in the so–called robust WLS. The robust variant of WLS with mean–and–variance corrected chi-square and standard errors (WLSMV; Muthén et al., 1997; Asparouhov & Muthén, 2010b) has been advocated because of good performance in simulation studies (e.g., Beauducel & Herzberg, 2006; Barendse et al., 2014).

### 4.1.2 Multivariate ML Estimation Method

In the multivariate maximum likelihood estimation method the likelihood of the complete response patterns is maximized. The method is also known as full information maximum likelihood (FIML; Lee, Poon & Bentler, 1990).

Let $\rho$ denote the vector containing the correlations between all pairs of continuous item score variables $X_i^*$ and $X_j^*$ with $i, j = 1 \ldots k$ and $i \neq j$. The expected proportion $\pi$ of response vector $x$, given correlations $\rho$ and thresholds $\tau$, is given by

$$\pi_{x_1,x_2,\ldots,x_k}(\rho,\tau) = Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k|\rho,\tau)$$
$$= \int_{\tau_{x_1-1}}^{\tau_{x_1}} \int_{\tau_{x_2-1}}^{\tau_{x_2}} \ldots \int_{\tau_{x_k-1}}^{\tau_{x_k}} f(x_1^*, x_2^*, \ldots, x_k^*|\rho,\tau)dx_1^*dx_2^*\ldots dx_k^*, \quad (4.3)$$

where $f$ denotes the $k$-dimensional normal density. Let index $r$ refer to a complete item response pattern $(x_1, x_2, \ldots, x_k)$, and let $p_r$ denote the observed proportion of respondents with response pattern $r$ in the sample. The log–likelihood of response pattern $r$ is given by

$$\ln L(\rho,\tau) = \sum_{r=1}^{m^k} p_r \ln[\pi_r(\rho,\tau)] + constant, \quad (4.4)$$

which is maximized to obtain the estimates for the parameters $\rho$ and $\tau$. As maximizing this log–likelihood requires numerical evaluation of high–dimensional integration over $x^*$ (Equation 4.3) in order to obtain the probability function of a response vector, Jöreskog & Moustaki (2001) already concluded that FIML is only feasible with a small numbers of variables (e.g., four or less). This seriously limits the application of FIML in practice.

### 4.1.3 Bivariate ML Estimation Method

In the bivariate maximum likelihood estimation method, high numerical integration is avoided by considering bivariate information only. In this method, the sum of the likelihoods of all possible bivariate response patterns is maximized, instead of the likelihood of the full multivariate response pattern.

For two items $i$ and $j$, the proportion of respondents with scores $x_i$, $x_j$ is given by

$$\pi_{x_i,x_j}(\rho_{ij},\tau_i,\tau_j) = \int_{\tau_{x_i-1}}^{\tau_{x_i}} \int_{\tau_{x_j-1}}^{\tau_{x_j}} f(x_i^*, x_j^*|\rho_{ij},\tau_i,\tau_j)dx_i^*dx_j^*, \quad (4.5)$$

for $\tau_i = (\tau_{1_i}, \tau_{2_i}, \ldots, \tau_{m-1})$ and $\tau_j = (\tau_{1_j}, \tau_{2_j}, \ldots, \tau_{m-1})$. With $p_{x_i,x_j}$ denoting the sample proportion of responses $x_i$ and $x_j$, where $\pi_r$ and $p_r$ in Equation 4.4 can be substituted by $\pi_{x_i,x_j}$ and $p_{x_i,x_j}$, and instead of maximizing the multivariate log-likelihood, the sum of all possible bivariate log-likelihoods can be maximized. This is equivalent to minimizing

$$F(\rho_{ij}, \tau_i, \tau_j) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{x_i=1}^{m} \sum_{x_j=1}^{m} p_{x_i,x_j} \ln[p_{x_i,x_j} / \pi_{x_i,x_j}(\rho_{ij}, \tau_i, \tau_j)], \qquad (4.6)$$

to obtain the likelihood estimates of the parameters $\rho_{ij}$ and $\tau_i, \tau_j$.

Jöreskog & Moustaki (2001) referred to this method as the underlying bivariate normal method. They originally suggested to use both the univariate and bivariate distributions, but in a simulation study Katsikatsou, Moustaki, Yang-Wallentin & Jöreskog (2012) concluded that the univariate distributions have no additional value in parameter estimation. The estimation method that only relies on bivariate likelihoods is referred to as the pairwise maximum likelihood (PML) method.

The PML estimation method has the advantage over FIML that it is computationally feasible, but it has the disadvantage that it only uses the bivariate distributions of the observed variables, and thus does not utilize all available information.

As an overall measure of fit, Jöreskog & Moustaki (2001) proposed to use the average of all bivariate likelihood tests, but this statistic can not be used as a goodness–of–fit test as its distribution is unknown. Maydeu-Olivares (2006) and Maydeu-Olivares & Joe (2006) introduced a family of goodness–of–fit statistics for testing composite null hypotheses in multidimensional contingency tables. As the PML method has been recognized as a special case of the maximum composite likelihood method (Varin, 2008; Varin, Reid & Firth, 2011), the method can be used to obtain a residual based goodness–of–fit test (Maydeu-Olivares, 2006; Maydeu-Olivares & Joe, 2006) and standard errors for PML estimates (Xi, 2011). In a simulation study, Xi (2011) found the composite likelihood goodness–of–fit test and standard errors of the PML estimates to be appropriate, when compared to a full information expectation maximization algorithm. However, the composite likelihood method has not yet been implemented in a readily available computer program.

Below we propose goodness-of-fit tests that are based on the ordinary likelihood ratios of either full response patterns or pairwise response patterns. These fit statistics through PML can be obtained with computer programs such as Mx (Neale, Boker, Xie & Maes, 2002), OpenMx (Boker, Neale, Maes, Wilde, Spiegel, Brick, Spies, Estabrook, Kenny, Bates & others, 2011), and Lavaan (Rosseel, 2012). The purpose of the present paper is to investigate the value of these fit statistics.

## 4.2 Model Fit Statistics

In the PML method, model parameters are estimated by maximizing the sum of the likelihoods of all bivariate responses patterns, for all pairs of items. As the distribution of this sum is not known, we propose three other measures of fit that are based on likelihood ratios: $C_F$, $C_M$, and $C_P$. The $C_F$ compares the model–implied proportions of response patterns with observed proportions of full response patterns (signified by subscript $F$). The $C_M$ fit statistic compares the model–implied proportions of response patterns with the expected proportions under the assumption of multivariate normality (signified by subscript $M$). The $C_P$ fit statistic compares the model–implied proportions of pairs of item responses (signified by subscript $P$) to the observed proportions of pairs of item responses.

The first PML measure of fit compares the log–likelihood of the model–implied proportions of multivariate response patterns given by Equation 4.4 with the observed sample proportions of response patterns. Multiplied by two times the sample size, we obtain

$$C_F = 2N \sum_r p_r \ln[p_r/\hat{\pi}_r], \tag{4.7}$$

which asymptotically has a chi–square distribution with degrees of freedom equal to the difference between the number of possible response patterns and the number of model parameters to be estimated minus one,

$$df_F = m^k - n - 1, \tag{4.8}$$

where $n$ is the number of parameters to be estimated.

As the number of possible response patterns $m^k$ is usually much larger than sample size $N$, most response patterns will not be observed at all, yielding many empty cells in the multivariate $m^k$ table, thereby causing bias in the $C_F$ statistic. Therefore, as a possible solution, we consider a second measure of fit, $C_M$, that compares the log–likelihood of the model-implied proportions of the model–of–interest with the model–implied proportions of the model that only assumes an underlying multivariate normal distribution (without any further restrictions):

$$C_M = C_{F1} - C_{F0}, \tag{4.9}$$

where $C_{F1}$ is $C_F$ for Model 1, the model of interest, and $C_{F0}$ is $C_F$ for Model 0, the model that assumes underlying multivariate normality and that has all polychoric correlations $\rho$ and all thresholds $\tau$ as its parameters. Statistic $C_M$ has a chi–square distribution with degrees of freedom equal to the differences in the numbers of parameters of Models 0 and 1,

$$df_M = k(k-1)/2 + k(m-1) - n_1, \tag{4.10}$$

where $k(k-1)/2$ is the number of polychoric correlations, $k(m-1)$ is the number of thresholds, and $n_1$ is the number of parameters of the model of interest. If the bias caused by empty cells in the $m^k$ table cancels out, then $C_M$ may outperform $C_F$.

With the third measure of fit, $C_P$, we only consider pairs of responses, and compare observed and model–implied proportions of pairs of responses. For items $i$ and $j$ (Agresti, 2002),

$$C_P = 2N \sum_{x_i=1}^{m} \sum_{x_j=1}^{m} p_{x_i,x_j} \ln[p_{x_i,x_j}/\hat{\pi}_{x_i,x_j}], \tag{4.11}$$

which has an asymptotic chi–square distribution with degrees of freedom equal to the information (which is $(m^2-1)$) minus the number of parameters (i.e., $2(m-1)$ thresholds and 1 correlation),

$$df_P = m^2 - 2(m-1) - 2. \tag{4.12}$$

As there are $k(k-1)/2$ possible pairs of items, this $C_P$ should be applied with a Bonferroni adjusted level of significance $\alpha^*$, with

$$\alpha^* = \frac{2\alpha}{k(k-1)}, \tag{4.13}$$

to keep the family–wise error rate at $\alpha$. The hypothesis of overall goodness-of-fit is tested at $\alpha$ and rejected as soon as $C_P$ is significant at $\alpha^*$ for at least one pair of items. Notice that with dichotomous items, $m = 2$, $df_P = 0$, so that the hypothesis of an underlying bivariate normal distribution cannot be tested. So, statistic $C_P$ can only be applied when there are more than two response options.

In the present paper we conduct a simulation study to investigate the value of statistics $C_F$, $C_M$, and $C_P$ to evaluate the fit of models that are estimated through PML. We also compare the performance of these statistics with the chi-square measure of overall goodness–of–fit that is associated with robust WLS estimation, which we will refer to as $C_W$. This $C_W$ statistic requires corrections to the standard errors and test statistics (Muthén et al., 1997; Satorra & Bentler, 2001; Asparouhov & Muthén, 2010b). Here we will use the mean–and–variance corrected chi–square statistic (Asparouhov & Muthén, 2010b).

## 4.3 Method

To evaluate fit statistics $C_F$, $C_M$, $C_P$, and $C_W$, we conduct a simulation study in which we vary sample size (200, 500, and 1,000) and the number of response options (2, 3, and 4), yielding nine conditions. With 1,000 replications, we obtain 9,000 datasets that are analyzed using the PML and robust WLS estimation methods.

### 4.3.1 Data Generation

We partly replicate the simulation study conducted by Katsikatsou et al. (2012). They generated item scores on six items according to a two factor model with factor loadings

$$\Lambda = \begin{bmatrix} 0.9 & 0 \\ 0.8 & 0 \\ 0.7 & 0 \\ 0.5 & 0.6 \\ 0 & 0.7 \\ 0 & 0.8 \end{bmatrix}, \tag{4.14}$$

common factor variances and covariances

$$\Phi = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \tag{4.15}$$

and residual variances

$$\Theta = \mathtt{I} - \text{diag}(\Lambda\Phi\Lambda'). \tag{4.16}$$

Continuous item scores are drawn from a multivariate normal distribution with variances and covariances

$$\Sigma = \Lambda\Phi\Lambda' + \Theta, \tag{4.17}$$

and zero means.

For each sample size (200, 500, and 1,000), we generate 1,000 datasets of continuous scores. These scores are categorized into two categories (threshold 0, yielding expected proportions 0.50 and 0.50), three categories (thresholds –0.6 and 0.6, yielding expected proportions of 0.27, 0.45, and 0.27), and four categories (thresholds –1.2, 0, and 1.2, yielding expected proportions 0.11, 0.39, 0.39, and 0.11; in line with Katsikatsou et al., 2012).

### 4.3.2 Analysis

We fit three models to each of the 9,000 datasets: a baseline model, a one–factor model, and a two–factor model. The baseline model includes all polychoric correlations and thresholds. If the baseline model does not fit then we must reject the hypothesis of an underlying multivariate normal distribution. The one–factor model has a free $6 \times 1$ matrix $\Lambda$ and a free diagonal $6 \times 6$ matrix $\Theta$. The $1 \times 1$ matrix $\Phi$ is fixed at unity. The two-factor model corresponds to the data generation model and has a $6 \times 2$ matrix $\Lambda$ with a pattern of free factor loadings that corresponds with $\Lambda$ above, a $2 \times 2$ symmetric matrix $\Phi$ with diagonal elements fixed at unity and a free off–diagonal element, and a $6 \times 6$ diagonal $\Theta$ matrix equal to $\mathtt{I} - \text{diag}(\Lambda\Phi\Lambda)$.

We use two estimation methods: PML and robust WLS. Model fit will be evaluated with measures $C_F$, $C_M$, and $C_P$ after PML estimation and with measure $C_W$ after robust WLS estimation. The computer program Mx (Neale et al., 2002) is used for PML estimation, and the computer program Mplus (Muthén & Muthén, 2010) for robust WLS estimation. The computer program R is used to calculate fit measures (using the "mvtnorm" package; R version 2.12.0; R Development Core Team, 2010).

The performance of the four measures of fit will be evaluated by calculating the proportions of model rejection in each of the conditions. The baseline model and the two–factor model should fit. When testing at a 5% level of significance, these two models should be rejected in 5% of all cases. The one–factor model should not fit and should always be rejected.

## 4.4 Results

Before presenting the results of the different methods for the evaluation of model fit, we briefly comment on the accuracy and efficiency of parameter estimation through PML. The accuracy is evaluated by calculating the absolute differences between the parameter estimates and the population values. The standard deviations indicate the efficiency of the parameter estimates.

Katsikatsou et al. (2012) already reported on the accuracy and efficiency of the parameter estimates in the case of four point response options and our results are very similar. Across all conditions, the average absolute difference of the factor loadings is 0.001 and the average standard deviation is 0.047. The average absolute difference of the correlation between the latent variables across all conditions is 0.001 and the average standard deviation across all conditions is 0.065. Noteworthy, the accuracy of the PML correlation parameter estimates across all conditions is slightly better than the robust WLS accuracy, with average absolute differences of 0.001 for PML and 0.003 for robust WLS, but the efficiency is the same. All other parameter results of the PML and robust WLS were quite similar. Our results are consistent with the results of Katsikatsou et al. (2012).

### 4.4.1 The $C_F$ Fit Statistic

Table 4.1 gives the results of fit evaluation with the $C_F$ statistic for the baseline model, the one–factor model, and the two–factor model. For each condition, the mean and standard deviation of the fit statistic, and the rejection rate at the 5% level of significance are calculated across 1,000 replications. The 95% confidence interval gives an idea of the variability of the rejection rate.

The fit of the baseline model is a test of the assumption of underlying multivariate normality, so we would expect rejection rates that equal the level of significance (5%). However, the overall rejection rates in the conditions with two–point response scales are too high (13.5%, 15.4%, 8.8%). With three–point and four-point response scales, the mean $C_F$ statistic is consistently too low to reject models, and the associated rejection rates are zero. The same is true for the two-factor model that should fit the data but is rejected too often in the conditions with two–point response scales (12.0%, 16.0%, 9.6%) and never rejected in the other conditions. The one–factor model is not correct and should be rejected, which is the case in conditions with two–point response scales but not in conditions with three–point scales and four–point scales.

We attribute the bad results in conditions with three–point scales and four–point scales to the large numbers of empty cells in the multivariate contingency tables. In the cases of three–point response scales and four–point response scales the numbers of possible response patterns are 729 and 4,096, whereas the total numbers of observations are only 200, 500, or 1,000, rendering the $C_F$ statistic unsuitable.

### 4.4.2 The $C_M$ Fit Statistic

Table 4.2 gives the results of fit evaluation with the $C_M$ statistic for the one–factor model and the two–factor model. The one–factor model is almost always rejected, except in the condition with sample size 200 and two–point response scales (with 0.987 rejection rate). The rejection rates for the two–factor model should be about equal to the level of significance (5%), but vary from 6.8% to 9.6%.

Overall, we consider the $C_M$ results satisfactory. Apparently, the sparseness of data and (almost) empty cells that invalidate the use of the $C_F$ statistic does not seem to affect the $C_M$ statistic much.

### 4.4.3 The $C_P$ Fit Statistic

The $C_P$ results are given in Table 4.3. As explained above, the $C_P$ statistic cannot be used with two-point response scales. For all other conditions Table 4.3 gives the means, standard deviations, and rejection rates of the highest $C_P$ among the 15 bivariate tests that are conducted with each dataset. To guard against inflation of the family–wise error rate, the level of significance is adjusted to 5% / 15 = 0.33%.

The rejection rates for the baseline model vary between 4.4% and 5.5% and for the two–factor model between 3.5% and 6.0%, which is reasonably close to the significance level of 5%. The one–factor model is almost always rejected in conditions with sample sizes of 500 and 1,000. However, in the small sample conditions rejection rates are only 67.0% and 53.9%.

Table 4.1: $C_F$ Rejection Rates

| Conditions | | $C_F$ | | | | 95% CI | |
|---|---|---|---|---|---|---|---|
| N | Scale | M($C_F$) | SD($C_F$) | *df* | RR ($C_F$) | $Q_{2.5}$ | $Q_{97.5}$ |
| *Baseline model* | | | | | | | |
| 200 | 2-point | 48.079 | 8.584 | 42 | 0.135 | 0.114 | 0.156 |
| | 3-point | 292.774 | 21.491 | 701 | 0.000 | - | - |
| | 4-point | 495.966 | 29.052 | 4,062 | 0.000 | - | - |
| 500 | 2-point | 47.921 | 10.052 | 42 | 0.154 | 0.132 | 0.176 |
| | 3-point | 396.944 | 25.581 | 701 | 0.000 | - | - |
| | 4-point* | 757.367 | 37.709 | 4,062 | 0.000 | - | - |
| 1,000 | 2-point | 44.524 | 9.815 | 42 | 0.088 | 0.070 | 0.106 |
| | 3-point | 470.109 | 27.018 | 701 | 0.000 | - | - |
| | 4-point | 982.595 | 41.825 | 4,062 | 0.000 | - | - |
| *One-factor model* | | | | | | | |
| 200 | 2-point | 90.359 | 15.587 | 51 | 0.919 | 0.902 | 0.936 |
| | 3-point | 363.119 | 25.975 | 710 | 0.000 | - | - |
| | 4-point | 580.380 | 33.606 | 4,071 | 0.000 | - | - |
| 500 | 2-point | 140.112 | 23.840 | 51 | 1.000 | - | - |
| | 3-point | 561.390 | 34.551 | 710 | 0.000 | - | - |
| | 4-point* | 957.356 | 46.486 | 4,071 | 0.000 | - | - |
| 1,000 | 2-point | 221.741 | 31.113 | 51 | 1.000 | - | - |
| | 3-point | 790.216 | 45.708 | 710 | 0.633 | 0.603 | 0.663 |
| | 4-point | 1,374.973 | 59.976 | 4,071 | 0.000 | - | |
| *Two-factor model* | | | | | | | |
| 200 | 2-point | 55.817 | 9.096 | 49 | 0.120 | 0.100 | 0.140 |
| | 3-point | 300.436 | 21.524 | 708 | 0.000 | - | - |
| | 4-point | 503.774 | 29.109 | 4,069 | 0.000 | - | - |
| 500 | 2-point | 55.420 | 10.588 | 49 | 0.160 | 0.137 | 0.183 |
| | 3-point | 404.481 | 25.652 | 708 | 0.000 | - | - |
| | 4-point* | 765.212 | 37.649 | 4,069 | 0.000 | - | - |
| 1,000 | 2-point | 52.086 | 10.773 | 49 | 0.096 | 0.078 | 0.114 |
| | 3-point | 477.726 | 27.308 | 708 | 0.000 | - | - |
| | 4-point | 990.351 | 41.933 | 4,069 | 0.000 | - | - |

Note. Means (M) and standard deviations (SD) of the fit statistic, rejection rates (RR) at a 5% level of significance, and 95% confidence intervals (CI) of the rejection rates are calculated across the 1,000 simulated datasets; * Results are given for a second drawing of 1000 as the first drawing, by chance, produced unexpected $C_P$ RR. Results for the first drawing are; M($C_F$) = 756.36, SD($C_F$) = 36.475, RR = 0.000 in a baseline model, M($C_F$) = 956.484, SD($C_F$) = 48.437, RR = 0.000 in a one–factor model, and M($C_F$) = 764.031, SD($C_F$) = 36.752, RR = 0.000 a two–factor model.

Table 4.2: $C_M$ Rejection Rates

| Conditions | | $C_M$ | | | | 95% CI | |
|---|---|---|---|---|---|---|---|
| N | Scale | M($C_M$) | SD($C_M$) | *df* | RR ($C_M$) | $Q_{2.5}$ | $Q_{97.5}$ |
| *One-factor model* | | | | | | | |
| 200 | 2-point | 42.280 | 13.885 | 9 | 0.987 | 0.980 | 0.994 |
| | 3-point | 70.344 | 18.109 | 9 | 1.000 | - | - |
| | 4-point | 84.415 | 21.178 | 9 | 1.000 | - | - |
| 500 | 2-point | 92.191 | 20.965 | 9 | 1.000 | - | - |
| | 3-point | 164.446 | 28.851 | 9 | 1.000 | - | - |
| | 4-point* | 199.989 | 31.577 | 9 | 1.000 | - | - |
| 1,000 | 2-point | 177.217 | 28.776 | 9 | 1.000 | - | - |
| | 3-point | 320.107 | 40.680 | 9 | 1.000 | - | - |
| | 4-point | 392.379 | 43.946 | 9 | 1.000 | - | - |
| *Two-factor model* | | | | | | | |
| 200 | 2-point | 7.738 | 4.260 | 7 | 0.091 | 0.073 | 0.109 |
| | 3-point | 7.661 | 4.099 | 7 | 0.072 | 0.056 | 0.088 |
| | 4-point | 7.808 | 4.173 | 7 | 0.083 | 0.066 | 0.100 |
| 500 | 2-point | 7.499 | 3.870 | 7 | 0.068 | 0.052 | 0.084 |
| | 3-point | 7.537 | 4.170 | 7 | 0.081 | 0.064 | 0.098 |
| | 4-point* | 7.845 | 4.387 | 7 | 0.096 | 0.078 | 0.114 |
| 1,000 | 2-point | 7.561 | 4.143 | 7 | 0.085 | 0.068 | 0.102 |
| | 3-point | 7.617 | 4.164 | 7 | 0.082 | 0.065 | 0.099 |
| | 4-point | 7.756 | 3.958 | 7 | 0.080 | 0.063 | 0.097 |

Note. Means (M) and standard deviations (SD) of the fit statistic, rejection rates (RR) at a 5% level of significance, and 95% confidence intervals (CI) of the rejection rates are calculated across the 1,000 simulated datasets; * Results are given for a second drawing of 1000 as the first drawing, by chance, produced unexpected $C_P$ RR. Results for the first drawing are; M($C_M$) = 200.116, SD($C_M$) = 32.076, RR = 1.000 in a one–factor model, and M($C_M$) = 7.663, SD($C_M$) = 4.276, RR = 0.073, CI = 0.057– 0.089 in a two–factor model.

Table 4.3: $C_P$ Rejection Rates

| Conditions | | $C_P$ | | | | 95% CI | |
|---|---|---|---|---|---|---|---|
| N | Scale | M($C_P$) | SD($C_P$) | *df* | RR ($C_P$) | $Q_{2.5}$ | $Q_{97.5}$ |
| *Baseline model* | | | | | | | |
| 200 | 3-point | 8.653 | 2.783 | 3 | 0.050 | 0.036 | 0.064 |
| | 4-point | 16.044 | 3.656 | 8 | 0.046 | 0.033 | 0.059 |
| 500 | 3-point | 8.393 | 2.893 | 3 | 0.055 | 0.041 | 0.069 |
| | 4-point* | 15.953 | 3.740 | 8 | 0.050 | 0.036 | 0.064 |
| 1,000 | 3-point | 8.419 | 2.846 | 3 | 0.049 | 0.036 | 0.062 |
| | 4-point | 16.141 | 3.640 | 8 | 0.044 | 0.031 | 0.057 |
| *One-factor model* | | | | | | | |
| 200 | 3-point | 16.577 | 5.307 | 3 | 0.670 | 0.641 | 0.699 |
| | 4-point | 24.338 | 6.159 | 8 | 0.539 | 0.508 | 0.570 |
| 500 | 3-point | 31.554 | 9.173 | 3 | 0.996 | 0.992 | 1.000 |
| | 4-point* | 42.918 | 10.273 | 8 | 0.995 | 0.991 | 0.999 |
| 1,000 | 3-point | 58.083 | 12.343 | 3 | 1.000 | - | - |
| | 4-point | 76.562 | 14.507 | 8 | 1.000 | - | - |
| *Two-factor model* | | | | | | | |
| 200 | 3-point | 8.918 | 2.789 | 3 | 0.057 | 0.043 | 0.071 |
| | 4-point | 16.306 | 3.675 | 8 | 0.052 | 0.038 | 0.066 |
| 500 | 3-point | 8.626 | 2.908 | 3 | 0.060 | 0.045 | 0.075 |
| | 4-point* | 16.190 | 3.744 | 8 | 0.049 | 0.036 | 0.062 |
| 1,000 | 3-point | 8.672 | 2.844 | 3 | 0.054 | 0.040 | 0.068 |
| | 4-point | 16.409 | 3.659 | 8 | 0.054 | 0.040 | 0.068 |

Note. Means (M) and standard deviations (SD) of the fit statistic, rejection rates (RR) at a 5% level of significance, and 95% confidence intervals (CI) of the rejection rates are calculated across the 1,000 simulated datasets; * Results are given for a second drawing of 1000 as the first drawing, by chance, produced unexpected results (i.e., M($C_P$) = 15.765, SD($C_P$) = 3.535, RR = 0.032, CI = 0.021–0.043 in a baseline model, M($C_P$) = 42.815, SD($C_P$) = 9.991, RR = 0.997, CI = 0.994–1.000 in a one–factor model, and M($C_P$) = 16.008, SD($C_P$) = 3.542, RR = 0.035, CI = 0.024–0.046 in a two–factor model).

### 4.4.4 The $C_W$ Fit Statistic

For the purpose of comparison, Table 4.4 gives the $C_W$ results after analysing all data sets with the robust WLS method of estimation. The one-factor model is almost always rejected. The rejection rates for the two-factor model vary between 3.9% and 6.4%.

The $C_W$ results with the two-factor model are somewhat better (closer to 5% rejection rates) than the $C_M$ results. The $C_W$ results are about similar to the $C_P$ results, except for the rejection rates of the one-factor model in small sample size conditions, in which the $C_W$ statistic seems to have more power.

## 4.5 Discussion

We proposed three statistics for goodness of overall fit of models that are estimated through the pairwise maximum likelihood (PML) method. With the $C_F$ statistic we test the difference between the model–implied proportions of multivariate response patterns and the observed proportions of multivariate response patterns. With the $C_M$ statistic we test the difference between model–implied proportions of multivariate response patterns and the proportions of response patterns that are implied by the assumption of underlying multivariate normally distributed continuous variables. With the $C_P$ statistic we test the difference between model–implied proportions of bivariate response patterns and observed proportions of bivariate response patterns.

The $C_F$ statistic appeared unsuitable for the evaluation of model fit. The performance of the $C_M$ statistic was good, although the rejection rates for the two factor model were consistently a little too high (varying between 6.8% and 9.6% instead of 5%). The $C_P$ statistic showed the best results with rejection rates close to the expected values (around 5% for models that should fit, and close to 100% for models that should not fit), except for relatively small sample sizes of 200 with which the rejection rates for the wrong one–factor model were substantially too low. For all fit statistics, we only reported results of testing at the 5%

Table 4.4: $C_W$ Rejection Rates

| Conditions | | $C_W$ | | | | 95% CI | |
|---|---|---|---|---|---|---|---|
| N | Scale | M($C_W$) | SD($C_W$) | *df* | RR ($C_W$) | $Q_{2.5}$ | $Q_{97.5}$ |
| *Baseline model* | | | | | | | |
| 200 | 2-point | 46.014 | 14.775 | 9 | 0.994 | 0.989 | 0.999 |
| | 3-point | 73.982 | 19.193 | 9 | 1.000 | - | - |
| | 4-point | 88.832 | 22.784 | 9 | 1.000 | - | - |
| 500 | 2-point | 102.986 | 23.243 | 9 | 1.000 | - | - |
| | 3-point | 177.773 | 30.827 | 9 | 1.000 | - | - |
| | 4-point* | 213.134 | 34.428 | 9 | 1.000 | - | - |
| 1,000 | 2-point | 201.828 | 33.761 | 9 | 1.000 | - | - |
| | 3-point | 348.523 | 43.996 | 9 | 1.000 | - | - |
| | 4-point | 421.149 | 47.477 | 9 | 1.000 | - | - |
| *One-factor model* | | | | | | | |
| 200 | 2-point | 7.032 | 3.780 | 7 | 0.044 | 0.031 | 0.057 |
| | 3-point | 6.933 | 3.557 | 7 | 0.041 | 0.029 | 0.053 |
| | 4-point | 6.992 | 3.686 | 7 | 0.060 | 0.045 | 0.075 |
| 500 | 2-point | 7.028 | 3.557 | 7 | 0.053 | 0.039 | 0.067 |
| | 3-point | 6.804 | 3.604 | 7 | 0.039 | 0.027 | 0.051 |
| | 4-point* | 7.186 | 3.530 | 7 | 0.044 | 0.031 | 0.057 |
| 1,000 | 2-point | 7.114 | 3.967 | 7 | 0.064 | 0.049 | 0.079 |
| | 3-point | 7.056 | 3.803 | 7 | 0.054 | 0.040 | 0.068 |
| | 4-point | 7.012 | 3.548 | 7 | 0.046 | 0.033 | 0.059 |

Note. Means (M) and standard deviations (SD) of the fit statistic, rejection rates (RR) at a 5% level of significance, and 95% confidence intervals (CI) of the rejection rates are calculated across the 1,000 simulated datasets; * Results are given for a second drawing of 1000 as the first drawing, by chance produced unexpected $C_P$ RR. Results for the first drawing are; M($C_W$) = 213.317, SD($C_W$) = 34.507, RR = 1.000 in a one–factor model, and M($C_W$) = 6.902, SD($C_W$) = 3.643, RR = 0.048, CI = 0.035–0.061 in a two–factor model.

level of significance, as the results at the 1% level of significance were very similar.

The performance of the PML fit statistics is only partly dependent on sample size. The $C_F$ statistic is not suitable with any sample size, as we observe the negative consequences of very large contingency tables affected by sparseness of data, as discussed earlier by Agresti & Yang (1987), Bartholomew & Leung (2002), Reiser & VandenBerg (1994), Reiser & Lin (1999), and Jöreskog & Moustaki (2001). Jöreskog & Moustaki (2001) tried to overcome the problem by adjusting the degrees of freedom. The $C_M$ statistic seems not that much affected by sparseness of data. The $C_P$ statistic uses bivariate tables only, but its power for rejecting the one–factor model is mediocre when the sample size is small. Still, the $C_P$ rejection rates for the correct models are not affected by small sample size. In our simulation study we also varied the number of response options, but this manipulation did not affect the results of the $C_M$ and $C_P$ fit statistics much.

We compared the results of the PML fit statistics results with results of robust weighted least squares (WLS) with the adjusted chi–square statistic $C_W$. The performance of $C_W$ was very similar to the performance of $C_P$, and in small sample conditions $C_W$ outperformed $C_P$ in rejecting the one–factor model. Still, robust WLS estimation is very different from PML estimation. Robust WLS is a multiple–step method that relies on the estimated polychoric correlations. The model–implied correlations are then fitted to fixed polychoric correlations, so there is no direct relation between the model–implied correlations and the observed discrete responses. That is why we really expected PML to behave better than robust WLS. However, in the present simulation study of six variables measuring two common factors, robust WLS did at least as well as PML.

We still do not know how robust WLS and PML compare in larger datasets, with more variables, and more complex models. There may be conditions in which PML show advantages over robust WLS, as we would expect from a theoretical perspective. We think that the PML method is a convenient alternative to FIML. The PML method seems a promising method that can be used to estimate all structural equation models, such as exploratory factor analysis models, multigroup models and longitudinal models (Moustaki, 2003; Vasdekis, Cagnone & Moustaki, 2012). We used Mx for PML, but PML including our proposed fit statistics has recently been implemented in OpenMX (Boker et al., 2011) and Lavaan (Rosseel, 2012) as well. When in future PML applications will become more common, the $C_M$ and $C_P$ statistics seem useful for the evaluation of overall goodness of fit.