*Behavior Research Methods*
*2005, 37 (4), 672-676*

# Confirmatory factor analysis using Microsoft Excel

JEREMY N. V. MILES
*University of York, York, England*

This article presents a method for using Microsoft (MS) Excel for confirmatory factor analysis (CFA). CFA is often seen as an impenetrable technique, and thus, when it is taught, there is frequently little explanation of the mechanisms or underlying calculations. The aim of this article is to demonstrate that this is not the case; it is relatively straightforward to produce a spreadsheet in MS Excel that can carry out simple CFA. It is possible, with few or no programming skills, to effectively program a CFA analysis and, thus, to gain insight into the workings of the procedure.

Microsoft (MS) Excel is a widely used spreadsheet package that can carry out many types of analysis. Excel can also be used for statistical analysis, either by employing the built-in Analysis ToolPak or by creating the appropriate spreadsheet to do the calculations required. We do not use Excel for everyday statistical analysis and would strongly recommend against it (see, e.g., Knusel, 1998; Simon, 2000). However, many people might be surprised to discover that MS Excel can be used to do simple (and more complex) confirmatory factor analysis (CFA). In this short article, we will present a method that allows the reader to do CFA in Excel—not, we would like to emphasize, because we think that this is the most useful tool. If one really needs to do CFA and has no suitable program, there is free software out there that can be used—for example, Mx (Neale, Boker, Xie, & Maes, 2004) or sem, a part of the free R package, (Fox, 2004). Excel may have a useful educational purpose, either for teaching or for people who wish to deepen their understanding of this statistical method.

## Confirmatory Factor Analysis

CFA is a technique based on a framework of structural equation modeling (SEM). It is contrasted with exploratory factor analysis (EFA). EFA is a data-driven process; the data are used to derive a model in an exploratory fashion. When CFA is used, the model first is proposed and then is applied to the data. The question is asked, is it feasible that these data could have been generated by this process?

The data are usually given in the form of a covariance matrix, shown in Table 1, and the model can be specified in terms of a path diagram, such as that shown in Figure 1. The data are taken from 358 respondents to the first six items on the Interval General Health Questionnaire (I-GHQ; Miller & Surtees, 1991; Surtees & Miller, 1990).

The model shown in Figure 1 proposes a single latent variable, which is able to account for the covariance between the items. The aim of a CFA is to find values for the unknowns in the model (the arrows) such that the difference between the sample covariance matrix and the covariance matrix implied by the model is minimized (see, e.g., Loehlin, 2004).

The implied matrix is calculated using matrix algebra. If the matrix of factor loadings is **L**, the matrix of errors is **E**, and the matrix of variances and covariances of the latent variables is **F**, the implied covariance matrix ($\boldsymbol{\Sigma}$) is given as Equation 1 (in this case, with one latent variable only, the matrix **F** will be a scalar):[1]

$$\boldsymbol{\Sigma} = \mathbf{LFL}' + \mathbf{E}. \quad (1)$$

Since the variance of the latent variables is frequently fixed to 1 for identification purposes, the matrix **F** becomes an identity matrix, and this drops out of the equation; it is simplified as Equation 2:

$$\boldsymbol{\Sigma} = \mathbf{LL}' + \mathbf{E}. \quad (2)$$

Estimates are found for the (unconstrained) elements of **L**, **F**, and **E**, which minimize the discrepancy between the implied covariance matrix and the sample covariance matrix. This is most frequently done using the maximum likelihood estimation, where the discrepancy function $D$ is given as in Equation 3:[2]

$$D_{\text{ml}} = \log|\boldsymbol{\Sigma}| + \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) - \log|\mathbf{S}| - p, \quad (3)$$

where **S** represents the sample covariance matrix, $\boldsymbol{\Sigma}$ the implied matrix, and $p$ is the number of indicator variables in the model. (Note that if the two covariance matrices are equal, the values $\log|\mathbf{S}|$ and $\log|\boldsymbol{\Sigma}|$ will be equal; in addition, $\mathbf{S}\boldsymbol{\Sigma}^{-1}$ will be an identity matrix, and hence, the trace of this matrix will be equal to the number of variables; hence, $D$ will equal zero, meaning that there is no discrepancy.)

Thanks to Martin Bland and Brendan Bunting for their comments on earlier drafts of the manuscript. Correspondence concerning this article should be addressed to J. N. V. Miles, Department of Health Sciences (Area 4), University of York, York YO10 5DD, England (e-mail: jnvm1@york.ac.uk).

**Table 1**
**Covariances (Above Diagonal) and Correlations (Below Diagonal) of Six Items From the I-GHQ**

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Been feeling unhappy and depressed? | .804 | .399 | .500 | .367 | .451 | .510 |
| 2. Been having restless and disturbed nights? | .487 | .833 | .433 | .283 | .372 | .377 |
| 3. Found everything getting "on top" of you? | .621 | .529 | .805 | .339 | .551 | .543 |
| 4. Been thinking of yourself as a worthless person? | .478 | .362 | .441 | .733 | .332 | .341 |
| 5. Felt constantly under strain? | .570 | .461 | .695 | .438 | .780 | .556 |
| 6. Been feeling nervous and strung up all the time? | .626 | .455 | .667 | .438 | .693 | .825 |

The maximum likelihood approach carries out an iterative search to find the values for the unknowns in the model that minimize the discrepancy function ($D$).

If the $D$ statistic is multiplied by the sample size − 1, this is distributed as a $\chi^2$ statistic. The degrees of freedom of the model are calculated using the number of elements in the covariance matrix analyzed—the number of parameters estimated in the model. This $\chi^2$ test is the test that the population residuals in the model are equal to zero (Bollen, 1989).

## CFA in Excel

From the extremely brief description of CFA above, it should be clear that in order to carry out CFA, a program needs the following capabilities.

1. Matrix algebra. A program must be able to invert and multiply matrices. Excel can do this using the minverse( ) and mmult( ) spreadsheet commands.

2. An iterative solver. Excel contains the Solver, which is able to search iteratively for a solution.

3. Statistical distribution functions for $\chi^2$. Excel contains the chidist( ) function for calculating probability values associated with $\chi^2$ at different degrees of freedom.

A further feature of Excel that proves useful is the ability to name ranges of cells. This enables one to define a matrix and refer to it by name.

## An Example

Using the data from the covariance matrix above, this section will show how to carry out a CFA in Excel and to calculate parameter estimates and fit statistics.

A range of cells can be defined and given a name in Excel. This range of cells can then be treated as a matrix or a vector, or individual cells can be referenced within the range, using the index( ) command.

The first matrix to be defined is **S**, the sample covariance matrix. This is shown in Figure 2; note that the name of the range of cells (**S**) appears in the name box on the top left.

In addition, we must define **L** (the factor loading matrix). This is a vector of six cells; it is useful to provide starting values that you believe may be in the appropriate range. It is also necessary to create a range of cells **L′**, this can be used with the index( ) command. The section of the spreadsheet is shown in Figure 3.

The matrix **L** is shown on the left-hand side. On the right-hand side is the matrix **L**-transposed (referred to as **transl**). Below the matrix **transl** is shown the Excel command (for illustrative purposes) to create **transl**. The index command contains two arguments: the matrix, and the element number.

The next matrix to be created is the matrix of unique variances and covariances, **E**. In the model shown, there

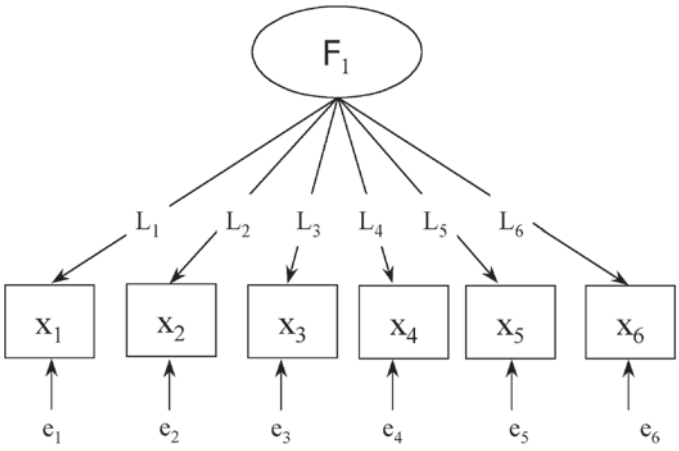
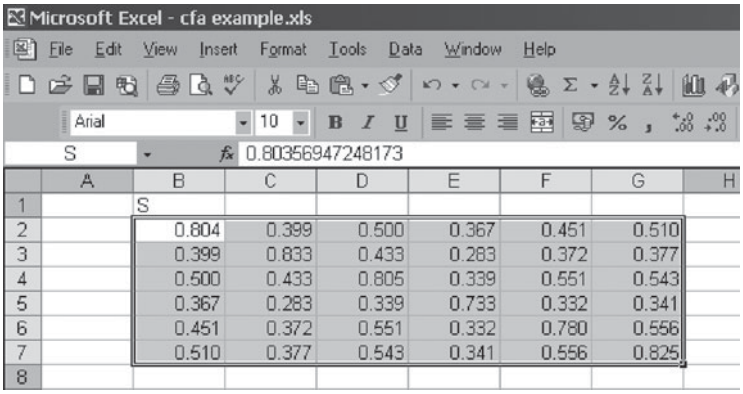



**Figure 1. A confirmatory factor analysis model with six items and one factor.**




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | S | | | | | | |
| 2 | | 0.804 | 0.399 | 0.500 | 0.367 | 0.451 | 0.510 | |
| 3 | | 0.399 | 0.833 | 0.433 | 0.283 | 0.372 | 0.377 | |
| 4 | | 0.500 | 0.433 | 0.805 | 0.339 | 0.551 | 0.543 | |
| 5 | | 0.367 | 0.283 | 0.339 | 0.733 | 0.332 | 0.341 | |
| 6 | | 0.451 | 0.372 | 0.551 | 0.332 | 0.780 | 0.556 | |
| 7 | | 0.510 | 0.377 | 0.543 | 0.341 | 0.556 | 0.825 | |
| 8 | | | | | | | | |

**Figure 2. Defining the matrix S.**

| L | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.700 | | Transl | | | | | |
| 0.700 | | 0.700 | 0.700 | 0.700 | 0.700 | 0.700 | 0.700 |
| 0.700 | | = index(L, 1) | | | | | |
| 0.700 | | | | | | | |
| 0.700 | | | | | | | |
| 0.700 | | | | | | | |

**Figure 3. L and the transpose of L.**

are no covariances between the unique variances; however, it is still worth creating the full matrix. Again, starting values are required. The matrix is shown in Figure 4.

The final matrix to create is $\boldsymbol{\Sigma}$, the implied covariance matrix, which is given in Equation 4:

$$\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{E}. \quad (4)$$

Note that we have removed **F** from the equation, since the latent variable is given a variance equal to 1, in order to identify the model. Ignoring the **F** matrix is feasible as long as there is only one factor or the factors are uncorrelated (in which case, the **F** matrix is an identity matrix. The $\boldsymbol{\Sigma}$ matrix is shown in Figure 5. (The numbering on the top and left of the matrix makes referring to elements of matrices simpler.)

The formula within each cell is as follows: =INDEX (MMULT(L,transl),$B26,C$25) + INDEX(E,$B26,C$25). The function mmult( ) is used to multiply two matrices—in this case, **L** and **transl**. $B26 and C$25 are used to reference the elements of the matrix. These take the row and column number from the edge of the matrix. The value is then added to the appropriate element of the **E** matrix, again using the referencing elements of the matrix from the edges.

By comparing the matrix **S** with $\boldsymbol{\Sigma}$, using Equation 3, we can calculate the discrepancy function. The equation is relatively straightforward, although the $\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1})$ requires that a large number of cells be referenced. It is easier (but much more computationally expensive) to write the spreadsheet to calculate each element of the diagonal of the matrix separately and sum them. The formula is written as =INDEX(MMULT(S, MINVERSE(sigma)), 1, 1) and is continued for each diagonal element of $\mathbf{S}\boldsymbol{\Sigma}^{-1}$ (2, 2, etc.). The more complex but computationally less expensive approach is to create the matrix $\mathbf{S}\boldsymbol{\Sigma}^{-1}$ and then extract the elements. Using the first method requires that the computer invert and multiply six matrices on each iteration; using the second approach requires that the operation be carried out only once.

The formula for *D* (the discrepancy function) is given by =LN(MDETERM(sigma))+SUM(trace)−LN(MDETERM(S))−6. The Solver is used to find the values for **L** and **E** that minimize the value of *D*. (Note that the Solver is an optional add-in in Excel; click Tools, Add-Ins, and ensure that Solver Add-In is ticked.) When the Solver is chosen from the Tools menu, the dialog box shown as Figure 6 appears. The value of the cell *D* is set to a minimum (take care here, because the default is maximum). The elements to change are **L** and the diagonals of **E** (which have been named e_diag). A number of options are available, but it is not necessary to change these.

| E | | | | | |
|---|---|---|---|---|---|
| 0.3 | 0 | 0 | 0 | 0 | 0 |
| 0.0 | 0.3 | 0 | 0 | 0 | 0 |
| 0.0 | 0 | 0.3 | 0 | 0 | 0 |
| 0.0 | 0 | 0 | 0.3 | 0 | 0 |
| 0.0 | 0 | 0 | 0 | 0.3 | 0 |
| 0.0 | 0 | 0 | 0 | 0 | 0.3 |

**Figure 4. Matrix of unique variances and covariances for E.**

| Sigma | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0.79 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| 2 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| 3 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| 4 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| 5 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |
| 6 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 | 0.49 |

**Figure 5. Sigma, the implied model covariance matrix.**

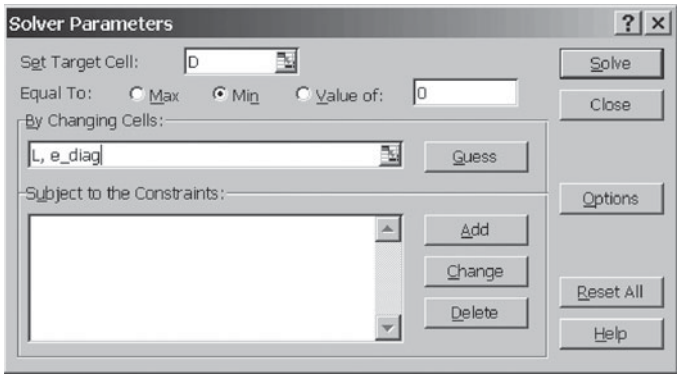



**Figure 6. Solver dialog box.**

Multiplying $D$ by $N-1$ gives the value for $\chi^2$. This is distributed with *df* equal to the number of sample moments—the number of elements estimated. It is easier to calculate the *df* manually; the number of elements in the covariance matrix is given by Equation 5,

$$\frac{p(p+1)}{2}, \tag{5}$$

where $p$ is the number of variables. The number of sample moments—that is, variances and covariances—is given by $k(k+1)/2$, where $k$ is the number of variables. We are estimating six elements of **L** and six of **E**; hence, there is a total of 12 parameters ($21 - 12 = 9$ *df*). The probability value associated with the $\chi^2$ test is found using the chidist function. The formula is given as =CHIDIST(chisquared, df).

## Estimation and Comparison

Estimating the model took 14 iterations and (I would estimate) less than 1 sec; however, it could be argued that the starting values were similar to the final values. For different starting values, it is often necessary to add constraints to the **E** matrix. This is done using the Solver, and all values were constrained to be above 0.001. If, for example, all elements of the **L** matrix are given starting values of 0 and all diagonal elements of **E** are given starting values of 0.1, the model will not converge without constraints on the values of **E**, and 21 iterations are required.

## Model Results

**Model fit**. The model fit statistics from Amos 5.0 (Arbuckle, 2002) and Mplus 2.14 (Muthén & Muthén, 2002) are shown in Table 2. The values are not in complete agreement but are very close.

**Parameter estimates**. The parameter estimates of the **L** and **E** matrices, as estimated by Amos, Mplus, and Excel are shown in Table 3. Although the estimates are not in complete agreement, they are equivalent to two decimal places and are close to the estimates to three decimal places.

## Concluding Remarks

This article has demonstrated a simple spreadsheet for CFA. The spreadsheet has little functionality and would not be recommended as a useful tool; it does provide a useful pedagogical exercise. CFA is frequently considered an impenetrable technique, but this approach shows that it can be programmed in an Excel spreadsheet, the whole

**Table 2**
**Estimates of Model Fit From Excel, Amos, and Mplus**

| | Excel | Amos | Mplus |
|---|---|---|---|
| $\chi^2$ | 21.707 | 21.707 | 21.768 |
| $p$ | 0.0098 | 0.0096 | 0.0096 |

**Table 3**
**Estimates of L and E Matrices From Excel, Amos, and Mplus**

| **L** Matrix | | | **E** Matrix | | |
|---|---|---|---|---|---|
| Excel | Amos | Mplus | Excel | Amos | Mplus |
| 0.673 | 0.672 | 0.672 | 0.351 | 0.350 | 0.350 |
| 0.549 | 0.548 | 0.548 | 0.531 | 0.530 | 0.530 |
| 0.749 | 0.748 | 0.748 | 0.244 | 0.244 | 0.244 |
| 0.475 | 0.475 | 0.475 | 0.507 | 0.505 | 0.505 |
| 0.720 | 0.719 | 0.719 | 0.262 | 0.262 | 0.262 |
| 0.742 | 0.741 | 0.741 | 0.274 | 0.273 | 0.273 |

of which can be viewed on a monitor, without scrolling; the spreadsheet described in this article is eight columns wide and 39 rows long.

The spreadsheet has deliberately not added a great deal of functionality, which it would be possible to include. A range of fit indices would be straightforward to include; the null model would need to be calculated in a separate model for calculation of the incremental indices. (Calculation of the GFI and AGFI, which require further matrix algebra, would also be possible.) Standard errors and confidence intervals of parameter estimates are also not included; however, again, these could be added, with the price of additional complexity of the spreadsheet. Alternative estimation procedures, such as generalized least squares or unweighted least squares, could be added by altering the specification of the discrepancy function.

By adding constraints to the Solver estimation, it would be possible to make equality constraints—for example, fixing the factor loadings to be equal (a tau-equivalent model) or fixing both the loadings and the errors (a parallel model). A multiple group model could be estimated, and parameters constrained across groups. Means and intercepts could also be added to the model.

One use may be in simulation studies or, at least, learning about simulation studies. Excel contains functions for the generation of random data, and it is possible to use Excel to generate random data to fit a known model, apply transformation to those data, and then fit a confirmatory factor analysis model.

Excel also contains a programming language, VBA, that can automate many of the commands, or one can use buttons on the spreadsheet to run some commands. If this were intended to be more than an educational exercise, it could be done; however, although this increases user friendliness, it decreases the transparency of the model.

The spreadsheet can be downloaded from www-users .york.ac.uk/~jnvm1/excelcfa/ (note this is www-dash-users, not dot).

## REFERENCES


Arbuckle, J. (2002). Amos 5.0 [Computer software]. Chicago: SmallWaters.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Fox, J. (2004). sem [Computer program]. Available at socserv.mcmaster .ca/jfox/Misc/sem/.

Knusel, L. (1998). On the accuracy of statistical distributions in MS Excel 97. *Computational Statistics & Data Analysis*, **26**, 375-377.

Loehlin, J. C. (2004). *Latent variable models*. Mahwah, NJ: Erlbaum.

Miller, P. M., & Surtees, P. G. (1991). Psychological symptoms and their course in first year medical students as assessed by the Interval General Health Questionnaire (I-GHQ). *British Journal of Psychiatry*, **159**, 199-207.

Muthén, B., & Muthén, L. (2002). Mplus 2.14 [Computer software]. Los Angeles: Authors.

Neale, M. C., Boker, M., Xie, G., & Maes, H. H. (2004). *Mx: Statistical modelling*. Richmond: Virginia Institute for Psychiatric and Behavioral Genetics.

Simon, G. (2000). ExcelBAD04.doc [Data file]. Available at www.jiscmail .ac.uk/cgi-bin/wa.exe?A2=ind0012&L=assume&D=0&P=830.

Surtees, P. G., & Miller, P. M. (1990). The Interval General Health Questionnaire. *British Journal of Psychiatry*, **157**, 679-686.


## NOTES

1. Note that in the commonly used LISREL notation, **L** refers to the lambda (**Λ**) matrix, **E** to the theta (**Θ**) matrix, **F** to the phi (**Φ**) matrix, and the variables in the **F** matrix are referred to as xi (**ξ**).

2. Note that the discrepancy function is often labeled as *F*; however, to avoid confusing this with the factor matrix, *D* is used here.

(Manuscript received July 6, 2004;
revision accepted for publication January 6, 2005.)