

A Note on Normal Theory Power Calculation in SEM With Data Missing Completely at Random

Conor Dolan, Sophie van der Sluis, and Raoul Grasman
University of Amsterdam

We consider power calculation in structural equation modeling with data missing completely at random (MCAR). Muthén and Muthén (2002) recently demonstrated how power calculations with data MCAR can be carried out by means of a Monte Carlo study. Here we show that the method of Satorra and Saris (1985), which is based on the nonnull distribution of the (normal theory) log-likelihood ratio test, can also be used. Compared to a Monte Carlo study, this method is computationally less intensive. We discuss 2 ways to calculate power when data are MCAR, one based on multigroup analysis and summary statistics, the other based on transformed raw data. The latter method is quite simple to carry out. Four examples are presented. This article is limited to data MCAR. Generally MCAR is a strong assumption. We demonstrate that results of power analyses based on the MCAR assumption are not informative if the data are actually missing at random.

In structural equation modeling (SEM), power calculation based on the normal theory likelihood ratio test (LRT) has been developed and discussed by Satorra and Saris (1985; Saris & Satorra, 1993). In this method, power is calculated by integrating the nonnull distribution of the LRT. This method is quite easy to carry out, but is based on the various assumptions associated with the normal theory LRT, such as large samples and multivariate normality (Azzellini, 1996; Bollen, 1989) and small to moderate misspecification (Curran, Bollen, Paxton, Kirby, & Chen, 2002). An alternative approach to power calculation is based on the empirical, rather than the theoretical distribution of the test statistic (Lei & Dunbar, 2004; Muthén & Muthén, 2002; Yuan & Hayashi, 2003; Yung & Bentler, 1996). The empirical distribution is obtained by means of a Monte Carlo study. This method is

flexible as it furnishes the null and nonnull empirical distribution of any test statistic, or goodness of fit index, and it does not depend on distributional assumptions concerning the data. In addition, this method allows one to assess the effect of missing data. However, the Monte Carlo approach is computationally demanding and requires a dedicated program, such as *Mplus* (Muthén & Muthén, 1998).

The aim of this article is to show that power calculations with data missing completely at random (MCAR; Little & Rubin, 1989; Schafer & Graham, 2002) may be conducted using a method proposed by Satorra and Saris (1985). Power based on the nonnull distribution of the LRT may be calculated in two ways. First, one may adopt a multigroup setup using population covariance matrices, which remain sufficient statistics when data are MCAR. Second, one may use raw data, which are transformed to fit the population covariance matrix exactly (Bollen & Stine, 1993). The latter method requires the simulation, transformation, and analysis of a single data set, and thus is not a Monte Carlo approach, like that suggested by Muthén and Muthén (2002). We emphasize that the assumption of MCAR is highly restrictive. The results concerning power obtained given this assumption are not informative if the data are missing at random (MAR). This is demonstrated later. The assumption of MCAR is a good place to start in the absence of any hypothesis concerning the nature of the missingness.

The article first discusses briefly the Satorra and Saris (1985) method. Subsequently the definition of MCAR (Little & Rubin, 1989; Schafer & Graham, 2002) is presented. Next we discuss how power can be evaluated with data MCAR using the method of Satorra and Saris (1985) and present the results of four illustrative power calculations. Finally, some results to demonstrate that MCAR and MAR may differ greatly in their effect on the power to detect a nonzero correlation coefficient are presented. The article concludes with a brief discussion.

POWER CALCULATION BASED ON THE NONCENTRAL χ^2 DISTRIBUTION

Let the p -dimensional random variable x be distributed $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in the population. The population covariance matrix and mean vector are a function of a q -dimensional parameter vector $\boldsymbol{\theta}$: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite. The function is known as a LISREL model (e.g., Jöreskog & Sörbom, 1988, 1993). The q parameters in $\boldsymbol{\theta}$ are the unknown (to be estimated) parameters. We consider the true model $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}_0)$, and an alternative, false model $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}_A)$ and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}_A)$, where $\boldsymbol{\theta}_A$ is the q_A -dimensional vector of unknown parameters in the false model, and $\boldsymbol{\theta}_0$ is the q_0 -dimensional vector of unknown parameters in the true model ($q_0 > q_A$). The false model is nested in the true model (Bollen, 1989; Satorra & Saris, 1985); that is, the parameters in $\boldsymbol{\theta}_A$ represent a subset of the parameters in

θ_0 . To calculate power, we adopt the method of Satorra and Saris (1985; see also Saris & Satorra, 1993). This method is based on the normal theory log-likelihood ratio test statistic T , which is calculated as follows (Bollen, 1989; Lawley & Maxwell, 1971):

$$T = N^*[\log|\Sigma(\theta)| + \text{trace}(\Sigma(\theta)^{-1}\mathbf{S} - \log|\mathbf{S}| - p + (\mathbf{m} - \mu(\theta))'\Sigma(\theta)^{-1}(\mathbf{m} - \mu(\theta))] \quad (1)$$

where $\Sigma(\theta)$ and $\mu(\theta)$ represent the hypothesized structure, and \mathbf{S} and \mathbf{m} are the sample covariance matrix and mean vector based on N cases.¹ Under the assumption that $\Sigma = \Sigma(\theta)$ and $\mu = \mu(\theta)$ represent the true model, N is large, and the data are independently and identically (multivariate normal) distributed, the statistic T is χ^2 distributed with df_0 degrees of freedom (Azzelini, 1996; Bollen, 1989). We denote this $T \sim \chi^2(df_0)$. If $\Sigma = \Sigma(\theta)$ and $\mu = \mu(\theta)$ do not represent the true model, T is noncentral χ^2 distributed with df_A degrees of freedom and noncentrality parameter λ ; that is, $T \sim \chi^2(df_A, \lambda)$, where $\lambda > 0$.

In power calculations, the aim is to determine the probability $\text{prob}[\chi^2(df_A, \lambda) > c_\alpha]$, given $\Sigma(\theta_A)$ and $\mu(\theta_A)$, $\Sigma(\theta_0)$ and $\mu(\theta_0)$, N , and α . The critical value c_α is the value for which $\text{prob}[\chi^2(df_0) > c_\alpha] = \alpha$. $\text{Prob}[\chi^2(df_0) > c_\alpha]$ or α is called the Type 1 error probability. $\text{Prob}[\chi^2(df_A, \lambda) < c_\alpha]$, denoted β , is called the Type 2 error probability. The focus of this article is on the power of the test; that is, $\text{prob}[\chi^2(df_A, \lambda) > c_\alpha]$ or $1 - \beta$. This is the probability that the incorrect model $\Sigma(\theta_A)$ and $\mu(\theta_A)$ is rejected in favor of the correct model $\Sigma(\theta_0)$ and $\mu(\theta_0)$.

This article focuses on the difference between the false model $\Sigma(\theta_A)$ and $\mu(\theta_A)$ and the true model $\Sigma(\theta_0)$ and $\mu(\theta_0)$. This difference concerns a small number of parameters (i.e., $df_0 - df_A$). As explained by Satorra and Saris (1985), the parameter λ approximately equals

$$\lambda \approx N^*[\log|\Sigma(\theta_A)| + \text{trace}(\Sigma(\theta_A)^{-1}\Sigma(\theta_0)) - \log|\Sigma(\theta_0)| - p + \{\mu(\theta_0) - \mu(\theta_A)\}'\Sigma(\theta_A)^{-1}\{\mu(\theta_0) - \mu(\theta_A)\}] \quad (2)$$

The nonnull distribution of the test statistic is then $\chi^2(df_0 - df_A, \lambda)$. For instance, suppose $\Sigma(\theta_0)$ is 2×2 , and we expect a correlation of .25. We want to know the power to reject the alternative hypothesis that the correlation is zero, given $N = 123$ and $\alpha = .05$. The true model has $df_0 = 5$ parameters (2 mean, 2 standard deviations, 1 correlation) and the false model has $df_A = 4$ parameters (2 means, 2 standard deviations). We focus on the log-likelihood ratio test with $df_0 - df_A = 1$ df (given $\alpha = .05$, $c_\alpha = 3.8414$). As pointed out by Satorra and Saris (1985), the parameter λ is obtained by fitting the false model to the true covariance matrix (using any SEM program). In this case $\lambda \approx 7.938$. We can use a variety of programs to integrate the $\chi^2(1, 7.938)$ distribution (see Appendix A). We find that the power ($1 - \beta$) to reject the incorrect model, equals about .80.

¹Usually T is calculated using $N - 1$ rather than N in Equation 1 (see also Equations 2 and 6). Here N is convenient to retain comparability with power calculations based on raw data (see later).

MISSINGNESS MECHANISM: MCAR

To calculate power, one requires specific information (i.e., N, λ, α). The calculation of power given missing data requires additional information concerning the expected percentages of missing data, and the missingness mechanism. Here we consider the simplest mechanism, MCAR (Little & Rubin, 1989). Let \mathbf{X} denote the data matrix ($N \times p$), and \mathbf{R} denote an ($N \times p$) indicator matrix, and let \mathbf{x}_i ($1 \times p$) denote the i th row in \mathbf{X} , and \mathbf{r}_i ($1 \times p$) the i th row in \mathbf{R} . A given element of \mathbf{R} assumes the value 1 to indicate that the corresponding element in \mathbf{X} is observed, and 0 to indicate that it is missing. Suppose that the elements of \mathbf{R} assume the value 0 according to some fixed probability, for example, $\text{prob}[\mathbf{r}_{ij} = 0] = \tau_j$, where \mathbf{r}_{ij} denotes the element in the i th row and j th column of \mathbf{R} . Let $\boldsymbol{\tau}$ denote the vector of probabilities of missingness, $\boldsymbol{\tau} = [\tau_1, \tau_2, \dots, \tau_{1-p}, \tau_p]$. If the probabilities in $\boldsymbol{\tau}$ are independent of the data, the missingness mechanism is called MCAR (Little & Rubin, 1989; Schafer & Graham, 2002). We thus have

$$\text{prob}[\mathbf{r}_{ij} = 0|\mathbf{X}] = \text{prob}[\mathbf{r}_{ij} = 0] \tag{3}$$

In data simulation, we can generate MCAR by drawing the p -dimensional data vector \mathbf{x}_i for a given case i , and assigning a missing value code to each component j of \mathbf{x}_i with probability τ_j . The probability of observing \mathbf{r}_i , a given configuration of k missing components of \mathbf{x}_i , is thus:

$$\text{prob}[\mathbf{r}_i | \boldsymbol{\tau}] = \prod_{j=1}^p \tau_j^{1-r_{ij}} (1 - \tau_j)^{r_{ij}} \tag{4}$$

For instance, given $p = 3$, $\boldsymbol{\tau} = [.2, .2, .2]$ and $\mathbf{r}_i = [0, 1, 1]$, $\text{prob}[\mathbf{r}_i|\boldsymbol{\tau}]$ equals $.2^1 \cdot .8^0 \cdot .2^0 \cdot .8^1 \cdot .2^0 \cdot .8^1 = .2 \cdot .8^2 = .128$. We thus expect to observe this pattern of missings $N_i = N \cdot .128$ times in a given data set. Note that the probability of losing a complete case due to missingness is $\prod \tau_j$. Given $\boldsymbol{\tau} = [.2, .2, .2]$, this probability is $.2^3 = .008$.

In fitting structural equation models, the presence of data MCAR may be accommodated in several ways (Arbuckle, 1996; Gold & Bentler, 2000). Here we consider raw data likelihood estimation (Finkbeiner, 1979; Lee, 1986), which is currently available in several SEM programs, including Mx (Neale, Boker, Xie, & Maes, 1999),² Amos (Arbuckle, 1995), Mplus (Muthén & Muthén, 1998), and LISREL (Jöreskog & Sörbom, 1993, 1999). The incomplete data log-likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^N \{ -1/2 [p_i \cdot \log(2) + p_i \cdot \log(\pi) + \log|\mathbf{W}_i \boldsymbol{\Sigma}(\boldsymbol{\theta}) \mathbf{W}_i'| + (\mathbf{W}_i \mathbf{x}_i - \mathbf{W}_i \boldsymbol{\mu}(\boldsymbol{\theta}))' [\mathbf{W}_i \boldsymbol{\Sigma}(\boldsymbol{\theta}) \mathbf{W}_i']^{-1} (\mathbf{W}_i \mathbf{x}_i - \mathbf{W}_i \boldsymbol{\mu}(\boldsymbol{\theta}))] \} \tag{5}$$

that is, the natural logarithm of the multivariate normal distribution (Bollen, 1989). The matrix \mathbf{W}_i in Equation 5 equals $\text{diag}(\mathbf{r}_i)$, where the zero rows are de-

²The Mx program is freely available. See www.vcu.edu/mx/.

leted, and p_i equals the number of observed variables in case i ; that is, $p_i = \sum r_{ij}$ (summation over j). For instance, the matrix \mathbf{W}_i equals the $p \times p$ identity matrix, if all elements in the vector \mathbf{r}_i equal one (i.e., \mathbf{x}_i contains no missing data). If the k th component of \mathbf{r}_i is zero, $r_{ik} = 0$ (i.e., k component of \mathbf{x}_i is missing), the matrix $\text{diag}(\mathbf{r}_i)$ is $p \times p$, and \mathbf{W}_i is formed by removing the row with the zero on the k th diagonal element. \mathbf{W}_i is then a $p_i \times p$ matrix. For instance, if $p = 4$ and $\mathbf{r}_i = [1, 0, 1, 0]$, then $p_i = 2$ and

$$\begin{array}{cccc} \text{diag}(\mathbf{r}_i) = & 1 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 \\ & 0 & 0 & 1 & 0 \\ & 0 & 0 & 0 & 0 \end{array} \quad \mathbf{W}_i = \begin{array}{cccc} 1 & 0 & 0 & 0 \\ & 0 & 0 & 1 & 0 \end{array}$$

Let $L(\theta_S|\mathbf{X})$ and $L(\theta|\mathbf{X})$ denote the log-likelihoods for the saturated model and for the hypothesized model $\Sigma(\theta)$ in Equation 1, respectively. Bollen (1989) showed that $T = -2\log[L(\theta|\mathbf{X})/L(\theta_S|\mathbf{X})]$ equals the expression in Equation 1.

POWER CALCULATION WITH DATA MCAR

We consider two ways to calculate power when data are MCAR, based on Equation 1 and Equation 5. We first discuss the method based on Equation 1. As indicated by Jöreskog and Sörbom (1988, p. 259; Muthén, Kaplan, & Hollis, 1987), in the case of MCAR, a multigroup approach may be taken, provided that the subsample associated with each unique pattern of missingness exceeds the number of observed variables. For instance, given p variables with the probability of missingness greater than zero for each variable—that is, $\tau_j > 0$ ($j = 1, p$)—we expect $M = 2^p - 1$ unique patterns of missingness, \mathbf{r}_i ($i = 1 \dots M$). We discard the pattern consisting of p missing values. Note that the number of groups decreases if a given τ_j is zero. If k components of τ equal zero, the number of groups equals $M = 2^{p-k}$.

Associated with each pattern of missingness, there is an expected sample size $N_i = N^* \text{prob}[\mathbf{r}_i|\boldsymbol{\tau}]$, and there are p_i observed variables. As long as $N_i > p_i$, one can in practice adopt a multigroup approach to estimation based on Equation 1, rather than the raw data-likelihood estimation based on Equation 5. The requirement $N_i > p_i$ is necessary to ensure that the sample covariance matrix in each of the M groups is positive definite. This multigroup approach, however impractical in real data analysis (often $N_i \leq p_i$), is suitable to calculate power when data are expected to be MCAR. Here the matrices $\mathbf{W}_i \Sigma(\theta_0) \mathbf{W}_i'$ and $\mathbf{W}_i \Sigma(\theta_A) \mathbf{W}_i'$ are positive definite by definition:

$$\begin{aligned} \lambda \approx & \sum_i^M \{N_i^* (\log|[\mathbf{W}_i \Sigma(\theta_A) \mathbf{W}_i']| + \text{trace}([\mathbf{W}_i \Sigma(\theta_A) \mathbf{W}_i']^{-1} [\mathbf{W}_i \Sigma(\theta_0) \mathbf{W}_i']) \\ & - \log|[\mathbf{W}_i \Sigma(\theta_0) \mathbf{W}_i']| - p_i + \\ & [\mathbf{W}_i \boldsymbol{\mu}(\theta_0) - \mathbf{W}_i \boldsymbol{\mu}(\theta_A)]' [\mathbf{W}_i \Sigma(\theta_A) \mathbf{W}_i']^{-1} [\mathbf{W}_i \boldsymbol{\mu}(\theta_0) - \mathbf{W}_i \boldsymbol{\mu}(\theta_A)]\} \end{aligned} \tag{6}$$

This method has two complications. First N_i is not necessarily an integer. A simple solution is to round N_i to the nearest integer value. The second complication is that the number of groups in the analysis tends to become large as p increase ($M = 2^p - 1$). For instance, given $p = 6$, we have 63 groups in the analysis.

The second method consists of simulating data \mathbf{X} for each of the M groups, and transforming the data such that within each group i ($i = 1 \dots M$) the data fit $\mathbf{W}_i \Sigma(\theta_0) \mathbf{W}_i'$ exactly. Bollen and Stine (1993) discussed a transformation, which can be used to this end. In the complete data case, the following transformation of \mathbf{X} ensures that the covariance matrix and mean vector of \mathbf{X}^* equals $\Sigma(\theta_0)$ and $\mu(\theta_0)$, exactly:

$$\mathbf{X}^* = [\mathbf{X} - \mathbf{J} \otimes \mathbf{m}'] \mathbf{S}^{-1/2} \Sigma(\theta_0)^{1/2} + [\mathbf{J} \otimes \mu(\theta_0)'] \tag{7}$$

where \mathbf{S} and \mathbf{m} are the observed covariance matrix and mean vector (as defined in Equation 1), respectively, $\mathbf{S}^{-1/2} \mathbf{S}^{-1/2} = \mathbf{S}^{-1}$, $\Sigma(\theta_0)^{1/2} \Sigma(\theta_0)^{1/2} = \Sigma(\theta_0)$, \mathbf{J} is the N -dimensional column vector containing 1s, and \otimes is the Kronecker product (Schott, 1997). The operation $[\mathbf{X} - \mathbf{J} \otimes \mathbf{m}']$ centers the data, so that the means are zero. The matrices $\mathbf{S}^{-1/2}$ and $\Sigma(\theta_0)^{1/2}$ may be calculated by means of a Cholesky or triangular decomposition (Schott, 1997). This method involves the following steps. For each group i ($i = 1 \dots M$), we simulate N_i complete cases, transform the data according to Equation 7, and recode to missing the values expected to be missing in group i . For instance, given $N = 100$, $p = 3$, and $\tau = [.2, .2, .2]$, we have $\mathbf{r}_2 = [1, 1, 0]$ in the second group, and a sample size N_2 of $100 * .2 * (1 - .2)^2 = 12.8 \approx 13$. To obtain the data in this group we simulate a data set \mathbf{X}_2 ($N_2 = 13 \times p = 39$), transform the data according to Equation 7, and recode all values of the third variable to missing. We do this in each of the M groups (each with its own N_i and unique \mathbf{r}_i) and subsequently pool the M data sets. The pooled data will display exactly the expected missingness, and the covariance matrix and means vector of the pooled data will equal $\Sigma(\theta_0)$ and $\mu(\theta_0)$, respectively. With these data in place, we can then obtain the parameter λ by fitting the true and the false models by maximizing Equation 5 ($\lambda \approx -2 \log [(L(\theta_A, \mathbf{X}^*) / L(\theta_0, \mathbf{X}^*))]$). Clearly this method requires that $N_i > p_i$, so that \mathbf{S}_i , the sample covariance matrix in group i , is positive definite. The condition $N_i > p_i$ is not trivial. For instance, given $p = 4$ and constant MCAR probability $\tau_j = \tau$, we have 16 groups of size ranging from $N^*(1 - \tau)^4$ to $N^* \tau^*(1 - \tau)^3$. Given $\tau = .10$ and $N = 1,000$, the smallest group is expected to be $1,000 * .9^4 = 630.9 \approx 631$. Of course, one may simply choose N to be large to ensure that $N_i > p_i$, or simply discard the groups where $N_i \leq p_i$. Because these groups will be small in terms of sample size, they are unlikely to contribute much to power. Appendix B contains an R script (e.g., Dalggaard, 2002; R Development Core Team, 2004) to generate data in this manner. Next four small examples of power calculations based on the normal theory LRT with data MCAR are presented.

EXAMPLE 1: CORRELATION COEFFICIENT

Let us reconsider the example just discussed; that is, the power to reject the hypothesis that the correlation coefficient equals zero. To reject this hypothesis, given its true

value of .25 and $\alpha = .05$ ($c_\alpha = 3.8414$), we require $N = 123$ to attain power of about .80. Suppose one expects data to be MCAR with a probability of $\tau_1 = \tau_2 = .10$. Adopting the first procedure discussed previously (based on summary statistics), we have three configurations of missing data $\mathbf{r}_1 = [1, 1]$ (no missing), $\mathbf{r}_2 = [1, 0]$, and $\mathbf{r}_3 = [0, 1]$. The expected sample sizes for each pattern of missingness, given $N = 123$, and rounded to integer values, are $N_1 \approx 100$ ($.9 \cdot .9 \cdot 123 = 99.6$), and $N_2 = N_3 \approx 11$ ($.9 \cdot .1 \cdot 123 = 11.07$). We expect to lose one or two cases due to complete missingness ($.1 \cdot .1 \cdot 123 = 1.23$). We specify and fit the three-group model in Mx (Neale et al., 1999). We find that given MCAR with $\tau = .10$, the power to reject the alternative hypothesis decreases to .715 ($\lambda \approx 6.453$). The loss in power is thus about .085.

In this specific example, we find that Group 2 (missing variable 1) and 3 (missing variable 2) do not contribute to λ ; that is, the power. This is to be expected, as $\mathbf{W}_2 \Sigma(\theta_0) \mathbf{W}_2'$ and $\mathbf{W}_3 \Sigma(\theta_0) \mathbf{W}_3'$ are 1×1 covariance matrices, which do not contain the parameter of interest (i.e., the covariance). As such, these groups cannot contribute to the power to detect this parameter. In this specific example, listwise deletion, which amounts to retaining only the group with a full data matrix \mathbf{X} , or cases that do not include missing values, would have resulted in the same power. It is certainly not generally true that cases that include missing data are uninformative. The extent to which a case with missing values can contribute to power depends on the exact model, and the parameters of interest (i.e., the pattern of missingness).

We repeat this calculation using the second method based on raw data transformation. Fitting the raw data using Equation 5, we again find $\lambda \approx -2 \log [(L(\theta_A, \mathbf{X})/L(\theta_0, \mathbf{X}))] = 6.453$.

EXAMPLE 2: A SIMPLE STRUCTURAL REGRESSION MODEL

For the second example, we consider the following model:

$$\eta_1 = \phi_1 \cdot \xi + \zeta_1$$

$$\eta_2 = \phi_3 \cdot \eta_1 + \phi_2 \cdot \xi + \zeta_2$$

where $\sigma^2(\xi) = 1.0$, $\sigma^2(\zeta_1) = .84$, and $\sigma^2(\zeta_2) = .654$, $\phi_1 = \phi_2 = .4$, and $\phi_3 = .3$. The associated covariance matrix, $\Sigma(\theta_0)$, equals

$$\begin{matrix} & \eta_1 & \eta_2 & \xi \\ \eta_1 & \left[\begin{array}{ccc} 1 & & \\ .46 & 1 & \\ .40 & .52 & 1 \end{array} \right] & & \end{matrix}$$

A path diagram is shown in Figure 1 (top).

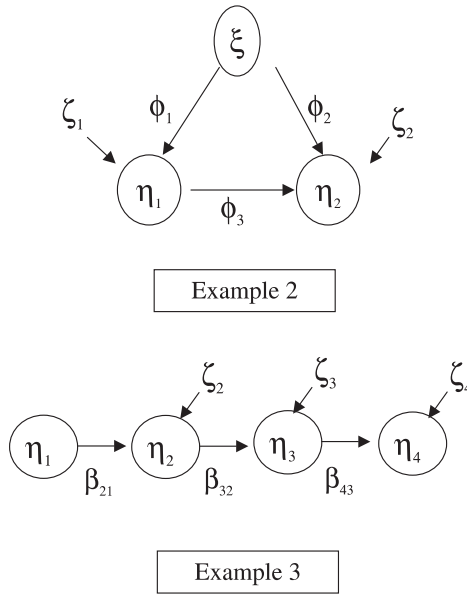


FIGURE 1 Top model used in Example 2; bottom model used in Example 3.

In this model, 34.6% of the variance in η_2 is explained by regression on η_1 and ξ . We are interested in the power to reject the hypothesis $\phi_3 = 0$, given $\alpha = .01$ ($c_\alpha = 6.634$). When $\phi_3 = 0$, the explained variance of η_2 decreases to 27.0%. The $\phi_3 = .3$ corresponds to an effect size of about 7.6% in terms of explained variance of η_2 . Power calculations reveal that we require 107 complete cases to attain a power of .80 ($df = 1, \alpha = .01, \lambda \approx 11.595$).

We assume that data are MCAR with a probability of $\tau = [.2, .2, .2]$. Using the multigroup approach based on sufficient statistics, we have $M = 2^3 - 1 = 7$ groups. Expected sample sizes are shown in Table 1. The power given data MCAR, and given $N = 107$ ($df = 1, \alpha = .01, \lambda \approx 7.133$), is .54. Table 1 contains the contributions of each group to power. The groups with one observed variable (Groups 4, 6, and 7) hardly contribute at all to power, as is to be expected. Of the groups with two observed variables, the group in which only x is missing ξ contributes most to power. This is comprehensible, as ϕ_3 is important in the covariance between η_1 and η_2 . If we were to carry out listwise deletion, which amounts to retaining only Group 1, the power would equal about .45. The incomplete data groups thus contribute about .09 to the power. In conclusion, we require $N = 107$ given $\tau = 0$ (no missing data) and $\alpha = .01$, to attain a power of .80. Given MCAR with $\tau = [.2, .2, .2]$, we would require about $N = 185$ to attain a power of .80.

TABLE 1
 MCAR Illustration 2 ($N = 107, \lambda \approx 7.133, \tau = [.1, .1, .1]$)

Group i	r_i	Size ($\tau = .2$)	N_i	λ_i	% Contribution to λ
1	1,1,1	$N^* \tau^{0*} (1 - \tau)^3$	55	6.06	84.97%
2	1,1,0	$N^* \tau^{1*} (1 - \tau)^2$	14	1.06	14.86%
3	1,0,1	$N^* \tau^{1*} (1 - \tau)^2$	14	.008	0.11%
4	1,0,0	$N^* \tau^{2*} (1 - \tau)^1$	3	< 1.E-6	~0%
5	0,1,1	$N^* \tau^{*} (1 - \tau)^2$	14	.004	0.06%
6	0,1,0	$N^* \tau^{2*} (1 - \tau)^1$	3	< 1.E-6	~0%
7	0,0,1	$N^* \tau^{2*} (1 - \tau)^1$	3	.0002	~0%

Note. Number of groups equals $M = 2^3 - 1 = 7$, r_i indicates the missingness in group i ($i = 1 \dots M$). The order of the variables is: η_1, η_2, ξ . For instance, $r_2 = [0,1,0]$ indicates that η_2 and ξ are missing.

EXAMPLE 3: THE SIMPLEX MODEL

Our third example is based on the simplex model (Jöreskog, 1971). This model is an example of an important class of time series models (autoregression-moving average [ARMA] models; e.g., Harvey, 1993):

$$\begin{aligned} \eta_1 &= \xi_1 \\ \eta_2 &= \beta_{21} \eta_1 + \xi_2 \\ \eta_3 &= \beta_{32} \eta_2 + \xi_3 \\ \eta_4 &= \beta_{43} \eta_3 + \xi_4 \end{aligned}$$

We specify $\beta_{21} = .6, \beta_{32} = .7, \beta_{43} = .8, \sigma^2(\zeta_1) = 100, \sigma^2(\zeta_2) = \sigma^2(\zeta_3) = \sigma^2(\zeta_4) = 64$. This model gives rise to the following covariance matrix $\Sigma(\theta_0)$:

$$\begin{bmatrix} 100.0 & & & & \\ 60.0 & 100.0 & & & \\ 42.0 & 70.0 & 113.0 & & \\ 33.6 & 56.0 & 90.4 & 136.32 & \end{bmatrix}$$

The path diagram of this model is shown in Figure 1 (bottom). We are interested in the power to reject the hypothesis $\beta_{21} = \beta_{32} = \beta_{43}$, given $\alpha = .05$. The associated log-likelihood ratio test has 2 df ($c_\alpha = 5.991$). Power calculations reveal that we require about 294 complete cases to attain a power of .80 ($\lambda \approx 9.647, \alpha = .05, df = 2$).

We suppose data are MCAR with probabilities of $\tau = [.2, .2, .2, .2]$. We expect $M = 2^4 - 1 = 15$ unique patterns of missingness (see Table 2 for expected sample sizes). Both the multigroup approach and the raw data transformation approach in-

TABLE 2
MCAR Illustration 3 ($N = 294$, $\lambda \approx 7.07$, $\tau = [.2, .2, .2, .2]$)

Group i	r_i	Size ($\tau = .2$)	N_i	λ_i	% Contribution to λ
1	1,1,1,1	$N^*\tau^{0*}(1 - \tau)^4$	120	3.99	56.5%
2	1,1,1,0	$N^*\tau^{1*}(1 - \tau)^3$	30	0.48	6.8%
3	1,1,0,1	$N^*\tau^{1*}(1 - \tau)^3$	30	0.70	9.9%
4	1,1,0,0	$N^*\tau^{2*}(1 - \tau)^2$	8	0.11	1.5%
5	1,0,1,1	$N^*\tau^{1*}(1 - \tau)^3$	30	0.67	9.5%
6	1,0,1,0	$N^*\tau^{2*}(1 - \tau)^2$	8	0.04	0.6%
7	1,0,0,1	$N^*\tau^{2*}(1 - \tau)^2$	8	0.04	0.6%
8	1,0,0,0	$N^*\tau^{3*}(1 - \tau)^1$	2	<1.E-4	<0.01%
9	0,1,1,1	$N^*\tau^{1*}(1 - \tau)^3$	30	0.72	10.2%
10	0,1,1,0	$N^*\tau^{2*}(1 - \tau)^2$	8	0.05	0.7%
11	0,1,0,1	$N^*\tau^{2*}(1 - \tau)^2$	8	0.10	1.4%
12	0,1,0,0	$N^*\tau^{3*}(1 - \tau)^1$	2	0.014	0.2%
13	0,0,1,1	$N^*\tau^{2*}(1 - \tau)^2$	8	0.13	1.8%
14	0,0,1,0	$N^*\tau^{3*}(1 - \tau)^1$	2	0.003	0.04%
15	0,0,0,1	$N^*\tau^{3*}(1 - \tau)^1$	2	0.007	0.10%

Note. Number of groups equals $M = 2^4 - 1 = 15$ ($i = 1 \dots M$), r_i denotes missingness in group i (e.g., in Group 10, the first and fourth variables are missing).

dicating that the power drops from .80 to about .625 given τ ($\lambda \approx 6.55$, $\alpha = .05$, $df = 2$). Listwise deletion would result in an expected sample size of $294^*.8^4 \approx 120$, and a power of about .41 ($\lambda \approx 3.95$, $\alpha = .05$, $df = 2$). So the additional 14 groups contributed substantially to the power (about .24). Table 2 shows the contributions of each group to the power. Given the MCAR probabilities $\tau = [.2, .2, .2, .2]$, we would require about $N = 398$ cases to ensure power of .80.

Finally to obtain an indication of the effects of MCAR in this model, we considered MCAR probabilities from $\tau_j = 0$ to $\tau_j = .5$ in steps of .05. Figure 2 depicts the power ($\alpha = .05$, $df = 2$) given increasing τ_j . Figure 2 also depicts the power given listwise deletion.

EXAMPLE 4: THE COMMON FACTOR MODEL (MUTHÉN & MUTHÉN, 2002)

We replicated the power calculation of Muthén and Muthén (2002, Table 1), who considered a two common factor model, where both common factors have five indicators (see Muthén & Muthén, 2002, for the path diagram). The reliability of each indicator is .64, and the correlation between the common factors equals .25. Muthén and Muthén (2002) considered the complete normal data case, and established that a sample size of $N = 150$ furnishes power of .81, to reject the hypothesis

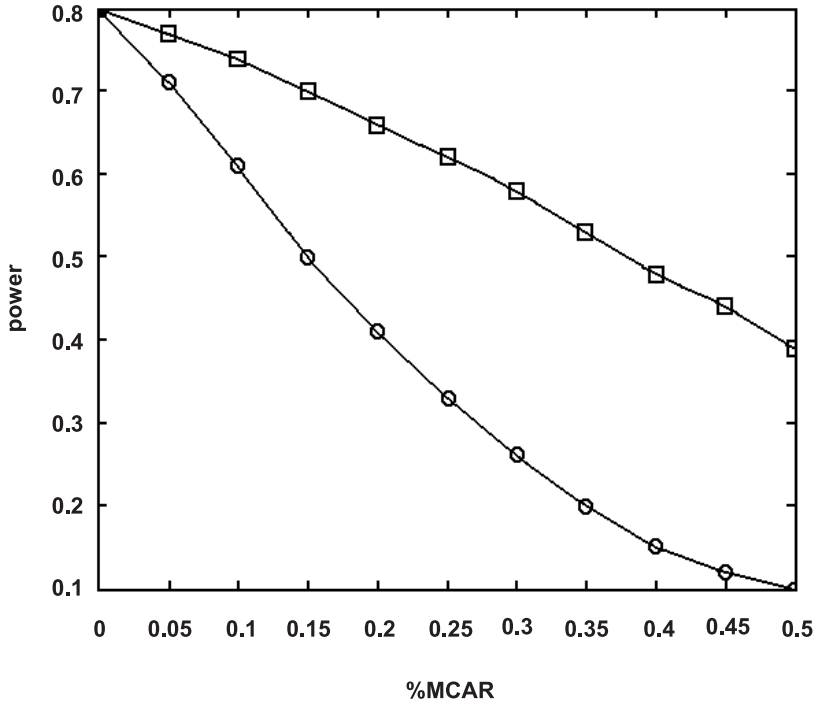


FIGURE 2 Power calculation in illustration 3 ($df = 2$, $\alpha = .05$). Squares: power based on all available data (raw data likelihood, see Equation 5). Circles: Power given listwise deletion. Y-axis: Power; X-axis: Percentage MCAR (%MCAR). When %MCAR is zero, $N = 294$, and the power equals .80. MCAR probability is equal over the four variables, and increases from 0 to .5.

that the common factor correlation equals zero, given $\alpha = .05$ ($c_\alpha = 3.8414$). Results obtained by means of standard power calculations using the method of Satorra and Saris (1985) indicated that $N = 150$ furnishes a power of .796. Muthén and Muthén (2002) also considered the situation in which 50% of the observations of each indicator of the second factor are MCAR. We thus have $\tau = [0, 0, 0, 0, 0, .5, .5, .5, .5, .5]$. Muthén and Muthén found that $N = 175$ is required to ensure a power of .81 ($\alpha = .05$, $c_\alpha = 3.8414$). Using the method of data transformation, we find that the power afforded by $N = 175$ equals .796. We checked this result by analyzing simulated data using our own Fortran program. With 3,099 replications each comprising $N = 175$ cases, we found that the observed log-likelihood ratio exceeded $c_\alpha = 3.8414$ in 80.1% of the cases. We consider the differences in the probabilities .801, .810, and .796, to be trivial. Note that the multigroup approach would now have to accommodate $M = 2^5 - 1 = 31$ unique patterns of missingness, that is, groups. This shows that the multigroup approach quickly becomes unmanageable.

MCAR VERSUS MAR: AN ILLUSTRATION

In the absence of any clear hypothesis concerning the nature of the missing data mechanism, the assumption of MCAR is a good point of departure. However, it is important to realize that results of power calculation based on the assumption of MCAR are not informative if the data are actually MAR. As earlier, let \mathbf{X} denote the data matrix (including observed and missing values), and let \mathbf{x}_i denote the data vector of the i th case. Let \mathbf{x}_{0i} denote the vector of observed components of \mathbf{x}_i ; thus \mathbf{x}_{0i} is a subset of \mathbf{x}_i . Whereas MCAR implies that $\text{prob}[\mathbf{r}_{ij} = 0 | \mathbf{x}_i] = \text{prob}[\mathbf{r}_{ij} = 0]$, MAR implies that $\text{prob}[\mathbf{r}_{ij} = 0 | \mathbf{X}] = \text{prob}[\mathbf{r}_{ij} = 0 | \mathbf{x}_{0i}]$ (Schafer & Graham, 2002). Schafer and Graham (2002) explained this as follows: “MAR means that a participant’s probabilities of response may be related only to his or her own observed set of observed items, a set that may change from one participant to another” (p. 152).

We demonstrate the difference in the effect of MAR and MCAR in a small simulation study. We consider two standard normally distributed variables x_1 and x_2 that are correlated .25. Given $N = 180$ and $\alpha = .01$ ($c_\alpha = 6.634$), the power to reject the hypothesis of a zero correlation is .80. The power given 10% MCAR in x_2 equals .75. Following the procedure suggested by Muthén and Muthén (2002), but using our own programs, we simulated $N = 180$ cases and introduced missingness according to Scenarios 3 to 9 as described in Table 3. In these scenarios x_2 is MAR, because the probability of missingness on x_2 depends on the values of x_1 . However, the actual number of missing cases in x_2 is always 10%, as in the MCAR Scenario

TABLE 3
Power of Likelihood Ratio Test to Detect a Correlation of .25 Between
Bivariate Standard Normal Variables x_1 And x_2 , given $\alpha = .01$, $df = 1$, and
NCP λ , in 9 Scenarios

Scenario	NCP λ	Power
1. No missing $N = 182$	11.7	.80
2. 10% of x_2 MCAR (effective $N = 164$)	10.5	.75
Variable x_2 MAR if		
3. $x_1 > 1.2815$	7.46	.56
4. $.8416 < x_1 < 1.2815$	10.2	.74
5. $.5244 < x_1 < .8416$	11.1	.77
6. $.2534 < x_1 < .5244$	11.5	.79
7. $0.0 < x_1 < .2534$	11.6	.80
8. $x_1 > 1.644$ or $x_1 < -1.644$	6.59	.50
9. $-.1256 < x_1 < .1256$	11.7	.80

Note. Power calculation given $N = 182$ (Scenario 1), given $N = 182$, with 10% of the cases missing x_2 completely at random (MCAR; Scenario 2), and given $N = 182$, with 10% missing x_2 at random (MAR; Scenarios 3–9). Note that the thresholds in Scenarios 3–9 are chosen such that 10% of x_2 is always missing. The NCP λ and power in Scenarios 3–9 are based on the analysis of simulated data (1,000 replications).

2. We carried out 1,000 replications, fitted the true model (correlation estimated) and the false model (correlation zero), and obtained empirical estimates of the power and of the noncentrality λ . The results are shown in Table 3.

It is evident from these results that the power to reject the zero-correlation hypothesis varies from .50 to .80, depending on the details of the MAR mechanism. Clearly the power of .75, given 10% MCAR in x_2 , is totally uninformative, if in reality the data are MAR. It is interesting to note that the effect of the missingness in x_2 on the power depends greatly on the values of x_1 that are associated with the missingness in x_2 . For instance, if missingness in x_2 is associated with intermediate values of x_1 , the power is hardly affected (e.g., in Scenario 3 the power is .80). However, if missingness in x_2 is associated with extreme values of x_1 , as is illustrated in Scenario 8, the power is greatly affected (.50).

CONCLUSION

The aim of this article was to consider power calculations based on the normal theory noncentral χ^2 distribution, as developed and discussed by Satorra and Saris (1985), when data are MCAR. In principle, when data are MCAR such calculations do not pose a problem. We suggested two ways in which one can carry out such power calculations: one based on summary statistics in a multigroup analysis (Jöreskog & Sörbom, 1988, p. 259), the other based on transformed data (Bollen & Stine, 1993). In practice, the latter method proved to be easier to carry out, as model specification and the analysis of raw data in SEM programs, such as LISREL (Jöreskog & Sörbom, 1999) or Mx (Neale et al., 1999), are quite straightforward. In addition, the actual generation and transformation of the data is not particularly difficult or time consuming (see Appendix B).

If one is interested only in the comparison of the complete data case and the listwise deletion case, one only has to alter the sample size from the total N to the N expected under listwise deletion. As demonstrated, however, incomplete cases, which are discarded in listwise deletion, may contribute significantly to power (see Figure 2). The size of the contribution will vary with the exact model, as can be inferred from Examples 1 and 3. Assessment of this contribution is relatively simple using the present methods.

The present use of the null, $\chi^2(df_0)$, and nonnull distributions, $\chi^2(df_A, \lambda)$, is based on many assumptions concerning the true and false models, the degree of misspecification, and the distribution of the data (Azzelini, 1996; Satorra & Saris, 1985). Importantly, the assumption of a true model $\{\Sigma(\theta_0) \text{ and } \mu(\theta_0)\}$ may be hard to justify in practice. MacCallum, Browne, and Sugawara (1996) discussed power calculations based on a test of close fit using the root mean squared error of approximation (RMSEA; Browne & Cudeck, 1993), rather than on a test of exact fit (i.e.,

$\chi^2(df_0)$). The methods of power calculation given data MCAR reported here may be used with RMSEA. Yuan and Hayashi (2003, section 2.4) explained how a population covariance matrix, which includes a controllable degree of approximation error, may be constructed. Once this matrix has been constructed to satisfy a given degree of approximation error, it may be used in raw data simulation and transformation (see Appendix B). Given the restrictions associated with the method of Satorra and Saris (1985), bootstrapping procedures remain an important alternative approach (Curran et al., 2002; Lei & Dunbar, 2004; Yuan & Hayashi, 2003).

We have limited our attention to the missing mechanism MCAR. Muthén and Muthén (2002) also considered the MAR mechanism, which is less restrictive than MCAR (Little & Rubin, 1989), and considered to be often more plausible (Schafer & Graham, 2002). Evaluation of power given data MAR may be carried out readily using the Monte Carlo method (Muthén & Muthén, 2002). Power evaluation by means of the method of Satorra and Saris (1985), given data MAR, is not possible using summary statistics (means vectors and covariance matrices), because these are not sufficient. A data transformation method may be possible, but is likely to be more complicated (again due to the absence of sufficient statistics). This is a subject for further study. Our limited comparison of the effects of MCAR and MAR on power (see Table 3) demonstrates that the evaluation of power given MCAR is not informative, if the data are in fact MAR.

REFERENCES

- Arbuckle, J. L. (1995). *Amos for Windows: Analysis of moment structures. Version 3.5*. Chicago: SmallWaters.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Azzelini, A. (1996). *Statistical inference based on the likelihood*. London: Chapman & Hall.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Stine, R. (1993). Bootstrapping goodness of fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavior Research*, *37*, 1–36.
- Dalgaard, P. (2002). *Introductory statistics with R*. New York: Springer.
- Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*, 409–420.
- Gold, M. S., & Bentler, P. M. (2000). Treatment of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, *7*, 319–355.
- Harvey, A. C. (1993). *Time series models* (2nd ed.). New York: Harvester Wheatsheaf.

- Jöreskog, K. G. (1971). Estimation and testing of simplex models. *The British Journal of Mathematical and Statistical Psychology*, 23, 121–145.
- Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications*. Chicago: SPSS.
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.
- Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8 [Computer Software]*. Chicago: Scientific Software International.
- Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London: Butterworth.
- Lee, S.-Y. (1986). Estimation for structural equation models with missing data. *Psychometrika*, 51, 93–99.
- Lei, P.-W., & Dunbar, S. B. (2004). Effects of score discreteness and estimating alternative model parameters on power estimation methods in structural equation modeling. *Structural Equation Modeling*, 11, 20–44.
- Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research*, 18, 292–326.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.
- Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431–462.
- Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide*. Los Angeles: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9, 599–620.
- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical modeling* (5th ed.). Richmond: Virginia Commonwealth University, Department of Psychiatry.
- R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.
- Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50, 83–90.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schott, J. R. (1997). *Matrix analysis for statistics*. New York: Wiley.
- Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology*, 56, 93–110.
- Yung, Y.-F., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195–226). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

APPENDIX A

We use the following R-script to calculate power given N , T , df , and α :

```
#start script
alpha=0.05      # user specified: type I error prob.
df=1            # user specified: degrees of freedom
```

```

T=44.037          # user specified: the test statistic T, the
                  log-likelihood ratio
N=992            # user specified: sample size used to calculate T
Ntrue=175        # user specified: sample size of interest
ca=qchisq(alpha,df,ncp=0,lower.tail=F) # critical value given alpha
lambda=(T/N)*Ntru # noncentrality parameter lambda
power=pchisq(ca,df,ncp=lambda,lower.tail=F)
print(power)
#end script

```

R is a statistical computer program that includes a programming language and many statistical and graphical procedures (e.g., see Dalgaard, 2002). It may be downloaded from www.r-project.org. The program Mx (Neale et al., 1999), which is also freely available, has handy facilities to fit alternative models, and calculate power using the method of Satorra and Saris (1985).

APPENDIX B

R-script (e.g., see Dalgaard, 2002) to generate data with data MCAR, which fit the population covariance matrix and mean vector exactly. The user may wish to alter the script. The underlined parts are user specified. This script was used in Example 3. This script can be obtained at <http://users.fmg.uva.nl/cdolan/>

```

#start script
rm(list=ls(all=TRUE))
library(norm)
ntot=5000          #sample size
nv<-4             #number of variables
sigma=matrix(c(
100.0,60,42.0,33.60,
60.0,100,70.0,56.00,
42.0,70,113.0,90.40,
33.6,56,90.4,136.32
),nv,nv)         #population cov matrix (Sigma[null])
mxm=rep(0,nv)    #means (mu[null])
tau<-c(.45,.45,.45,.45) #mcar probabilities
#
ngroups<-2^nv
xdata=matrix(NA,ntot*2,nv)
#- - - - -
o1<-matrix(c(1,0),2,1,byrow=T)
o2<-matrix(c(1,1),2,1,byrow=T)
#- - - - -
# 4 variables 2^4 patterns of missingness
c1<-o1%x%o2%x%o2%x%o2
c2<-o2%x%o1%x%o2%x%o2
c3<-o2%x%o2%x%o1%x%o2

```

```

c4<-o2%x%o2%x%o2%x%o1
mispat<-cbind(c1,c2,c3,c4)
##- - - - -
## NOTE FOR 2 variables ... patterns of missingness, see
illustration 1
# c1<-o1%x%o2
# c2<-o2%x%o1
# mispat<-cbind(c1,c2)
##- - - - -
## FOR 3 variable ... patterns of missingness, see illustration 2
# c1<-o1%x%o2%x%o2
# c2<-o2%x%o1%x%o2
# c3<-o2%x%o2%x%o1
# mispat<-cbind(c1,c2,c3)
##- - - - -
grsp=rep(1,ngroups)      # probability of missing configuration
grsn=rep(0,ngroups)      # sample size
grrn=rep(0,ngroups)      # rounded sample size
csigma=chol(sigma)
for (igr in 1:ngroups)
  {
    patm=mispat[igr,]
    patnom=1-mispat[igr,]
    for (i in 1:nv)
      {
        grsp[igr]=grsp[igr]*(tau[i]^(1-patm[i]))*
          ((1-tau[i])^patm[i])
      }
    grsn[igr]=ntot*grsp[igr]
    grrn[igr]=round(grsn[igr])
  }
ntotal=0
mgroup1=0
mgroup2=0
for (igr in 1:(ngroups-1))
  {
    if (grrn[igr] > 0.0)
      {
        mgroup1=mgroup1+1
        if (grrn[igr]>nv)
          {
            mgroup2=mgroup2+1
            np=grrn[igr]
            ntotal=ntotal+np
            ivec=matrix(rep(1,np),np,1)
            ydata=(matrix(rnorm(np*nv),np,nv,byrow=T))%*%csigma
            xmy=apply(ydata,2,mean)
            sy=(cov(ydata)*(np-1))/np
            ydata=(ydata-t(xmy%x%t(ivec))%*%t(chol
              (solve(sy))%*%csigma+t(xmx%x%t(ivec)))
            ydata[1:np,mispat[igr,1:nv]==0]=NA
          }
      }
  }

```

```

        xdata[((ntotal-np+1):ntotal),]=ydata[, ]
        rm("ydata")
    }
}
}
#check: calculate means and cov. matrix
sdata <- prelim.norm(xdata[1:ntotal,1:nv])
result <- em.norm(sdata,crit=0.00000000001)
result <- getparam.norm(sdata,result,corr=F)
osigma<-result$sigma
omu<-result$mu
print(osigma)
print(omu)
print(ntot)
#check groups sizes
ntau0=sum(tau==0)
if (ntau0==0) print(c(ngroups,ngroups-1,mgroup1,mgroup2))
if (ntau0>0) print(c(ngroups,2^(nv-ntau0),mgroup1,mgroup2))
# recode missing code NA to numeric code, say, -999
xdata[is.na(xdata)]<-(-999)
# write results
write(t(xdata[1:ntotal,1:nv]),file="dat.eg3",ncolumn=nv)
#end script

```