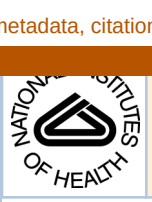



Published in final edited form as:
*Sociol Methods Res*. 2010 October 7; 39(2): 127–156. doi:10.1177/0049124110366238.

# MODEL IDENTIFICATION AND COMPUTER ALGEBRA

**Kenneth A. Bollen** and
Department of Sociology, H. W. Odum Institute for Research in Social Science, Carolina Population Center, University of North Carolina at Chapel Hill

**Shawn Bauldry**
Department of Sociology, University of North Carolina at Chapel Hill

## Abstract

Multiequation models that contain observed or latent variables are common in the social sciences. To determine whether unique parameter values exist for such models, one needs to assess model identification. In practice analysts rely on empirical checks that evaluate the singularity of the information matrix evaluated at sample estimates of parameters. The discrepancy between estimates and population values, the limitations of numerical assessments of ranks, and the difference between local and global identification make this practice less than perfect. In this paper we outline how to use computer algebra systems (CAS) to determine the local and global identification of multiequation models with or without latent variables. We demonstrate a symbolic CAS approach to local identification and develop a CAS approach to obtain explicit algebraic solutions for each of the model parameters. We illustrate the procedures with several examples, including a new proof of the identification of a model for handling missing data using auxiliary variables. We present an identification procedure for Structural Equation Models that makes use of CAS and that is a useful complement to current methods.

## 1 INTRODUCTION

Multiequation models with observed and latent variables are common in the social sciences. Economists routinely estimate simultaneous equation models that track the complex relations between endogenous and exogenous observed variables (e.g., Greene, 2008). Factor analysis is a multiequation system relating observed variables to one or more latent variables. Finally, general structural equation models (SEMs) combine features of simultaneous equations and factor analysis and are a well-known tool in the social and behavioral sciences.

Though the parameters might differ, these models have in common the identification problem. The question of identification is whether it is possible to find unique values of the model parameters. It is an issue not so much at the sample level as one that is best answered at the population level.[i] That is, if we had population moments of the observed variables (typically the means and covariance matrix), then identification asks whether this would be enough to uniquely solve for all of the model parameters.

Direct correspondence to: Kenneth A. Bollen, CB 3210 Hamilton, Department of Sociology, University of North Carolina, Chapel Hill, NC 27599-3210 (bollen@unc.edu).

An earlier version of this paper was presented at the American Sociological Association meeting held in August 2008 in Boston, MA.

[i]There is a separate issue of "empirical identification" (Kenny, 1979) that occurs when a model is identified in the population, but the sample values are close to values that would lead to an underidentified model. We do not focus on empirical identification.

In special cases rules of identification have solved the problem. For instance, in simultaneous equation models with no measurement error, the recursive rule holds when the disturbances of all equations are uncorrelated and there is no feedback in the system (e.g., Bollen, 1989, pp.95–98). Or if *all* simultaneous equation disturbances are correlated without restrictions, we have the rank and order conditions to which we can refer (e.g., Fisher, 1966; Bollen, 1989, pp. 98–103). Similarly, there are rules of identification for some factor analysis models and latent variable SEMs (e.g., Bollen, 1989, pp. 238-47, 326-32; Davis, 1993; O'Brien, 1994; Reilly, 1995; Rigdon, 1995). Unfortunately, many researchers' models are not covered by these identification rules. This has sometimes led to underidentified models being presented and later corrected (Burt, Wiley, Minor, and Murray 1978; Burt, Fischer, and Christman 1979).

In practice researchers rarely mention rules of identification and instead rely on SEM software that checks the singularity of the information matrix to determine identification. A singular information matrix is a symptom of an underidentified model (Rothenberg, 1971). Sample estimates replace population parameters in the information matrix for the singularity check in SEM software. Though this works well in most situations, it has limitations. One is that it is an empirical test based on the sample estimates rather than the population values of parameters. To the degree that the sample estimates differ from the population ones it is possible to reach different conclusions on singularity (McDonald and Krane, 1979; Bollen, 1989, pages 249-51) and hence identification. A second related limitation is the uncertainty in numerically determining the rank of the estimated information matrix (Bentler and Weeks, 1980). This problem would be most serious when there is near singularity of the information matrix. The near singularity could be due to a property of the population information matrix or to estimates that lead to near singularity even if the population information matrix is not problematic. The near singularity might challenge the limits of the numerical methods used to determine whether the rank condition of the information matrix is met. As Schoenberg (1981, page 2) states: "A numerical inverse may exist when the theoretical inverse does not, and vice versa; in other words, the computer may generate an inverse from a matrix that is theoretically indefinite, or may fail to compute an inverse when in fact it is theoretically positive definite." SEM software that relies on empirical checks of the singularity of the information matrix provide warning messages about potential identification problems. LISREL provides an example:

> W_A_R_N_I_N_G: [*parameter name*] may not be identified.
>
> Standard Errors, T-Values, Modification Indices, and Standardized Residuals cannot be computed.

Given the limitations of the empirical checks described above the use of the word *may* in this warning is well-advised. Other SEM software use similar cautionary wording. Though most times these messages will signify a real problem with identification, this will not always be true. Having another approach to determining identification status could lessen the ambiguity in determining whether this is a real problem or a false alarm.

A third limitation of reliance on the information matrix singularity check is that it is a check of *local* identification rather than *global* identification. To illustrate, consider Figure 1. It contains a model from Bollen and Hoyle (1990) with equal coefficients for the feedback relations between two latent variables. We analyzed the same data and the two runs differ only in their starting values. All of the parameters are identical except for the feedback coefficient and the error variances for the two latent variables. There are no warnings from the SEM software about model underidentification. Indeed, a researcher probably would not know that there was a second set of parameter values that were consistent with the data without having stumbled upon it and might incorrectly interpret the parameters from one

model rather than the other. In one set of coefficients the feedback loop is "explosive" and the system is unstable whereas in the other the relation converges to an equilibrium. We suspect that some readers will be surprised by this example, but it illustrates the difference between local and global identification. As we will show later, this model is locally, but not globally identified. As such the information matrix is nonsingular and checking the information matrix will not detect a problem even with the population parameter values.

Algebraic solutions are perhaps the surest way to establish model identification (e.g., Long, 1983, page 44). The multiequation models each lead to expressions that represent the means and covariances of the observed variables as functions of the parameters in the model. If we can show that each model parameter (e.g., coefficient, covariance of disturbances) is solvable as a *unique* function of some elements of the means or covariance matrix of the observed variables, then we can establish model identification.[ii] Solving for these model parameters by hand is sometimes possible, but with modestly-sized models such as that in Figure 1 or larger models, it often is difficult and prone to human errors. Furthermore, Duncan (1975, pages 84–86) provides an example where the algebraic approach seems to be misleading in suggesting a unique solution when there is none.

The purpose of our paper is to recommend a complementary approach to identification that makes use of computer algebra systems (CASs) and that can avoid some of the limitations we described. One role of CAS focuses on local identification. As we stated above the most common identification check relies on the information matrix evaulated at the sample estimates. The problems with this are traceable to the discrepancy between sample estimates and population values and the issues of numerical accuracy. CASs address both of these problems by performing the checks using symbolic algebra rather than sample estimates. In addition, we use the Wald Rank Rule (Wald, 1950; Bollen, 1989, pages 248-49) to evaluate rank. This CAS approach to local identification is well worked out in Bekker, Merckens, and Wansbeek (1994) who developed software to check local identification based on the Wald Rank Rule (the rank of a Jacobian matrix). This is closely related to the check on the singularity of the information matrix, except the evaluation occurs on the Jacobian matrix which involves first order derivatives rather than second order derivatives. More importantly, the assessment occurs not with empirical estimates but at the symbolic level. Though their book is out of print, we will discuss how to use CAS software to examine the Jacobian matrix and to check local identification using symbolic terms rather than empirical estimates.

A second role for CASs that is less explored is to solve for the model parameters in terms of the means, variances, and covariances of the observed variables to establish identification. To the best of our knowledge, no one has provided a general presentation on the use of computer algebra programs to solve for the model parameters in terms of the means and covariance matrix elements of the observed variables, though this possibility is known.[iii] There are complications with the CAS approach to finding unique solutions that have not been discussed in the literature that we address as well. In addition, we will look at the Duncan's (1975, pages 84–86) example that is sometimes raised as an objection to an algebraic approach to model identification.

[ii]With rare exceptions, the means and covariances of the observed variables are identified and showing that each parameter is a unique function of one or more of them establishes identification.

[iii]One of the authors attempted this in the early 1980s using Macsyma, a programming language from Bell Labs, but encountered computing difficulties that prevented successful implementation. A reviewer pointed out that Rigdon (1995, page 371) mentions that he used computer algebra to establish rules of identification for a class of block-recursive models. The computer algebra program or results are not provided.

More specifically, the purposes of our paper are: (1) to review the differences between local and global identification, (2) to review CAS approaches to local identification, (3) to develop a procedure for using CASs to assess global identification, and (4) to illustrate our procedure with four examples.

In the next sections, we present the general SEM and give the implied means and implied covariance matrix that correspond to the means and covariance matrix of the observed variables. After this is a section on local and global identification. Then come sections on CAS methods for local identification and CAS for global identification that finds algebraic solutions using symbolic manipulations. A concluding section follows with an appendix containing the programs for the examples in the text.

## 1.1 MODEL SPECIFICATION

Structural equation models (SEMs) include many models that are typical in the social and behavioral sciences (e.g., multiple regression, ANOVA, simultaneous equation models, confirmatory factor analysis, latent growth curve models, etc.). Here we present the general SEM so that our treatment is inclusive of a wide variety of identification issues that 6 occur in practice. We use a modified version of Jöreskog and Sörbom's (1993) LISREL notation for SEMs[iv], the most common notation in the field. It is convenient to distinguish between the *latent variable model* and the *measurement model*. The latent variable model is

$$\boldsymbol{\eta}_i=\boldsymbol{\alpha}_\eta+\mathbf{B}\boldsymbol{\eta}_i+\boldsymbol{\Gamma}\boldsymbol{\xi}_i+\boldsymbol{\zeta}_i \tag{1}$$

where $\boldsymbol{\eta}_i$ is a vector of latent endogenous variables for unit $i$, $\boldsymbol{\alpha}_{\boldsymbol{\eta}}$ is a vector of intercept terms for the equations, $\mathbf{B}$ is the matrix of coefficients giving the effects of the latent endogenous variables on each other, $\boldsymbol{\xi}_i$ is the vector of latent exogenous variables, $\boldsymbol{\Gamma}$ is the coefficient matrix giving the effects of the latent exogenous variables on the latent endogenous variables, and $\boldsymbol{\zeta}_i$ is the vector of disturbances. The $i$ subscript indexes the $i$th case in the sample. We assume that $E(\boldsymbol{\zeta}_i)=\mathbf{0}$, $COV(\boldsymbol{\xi}_i', \boldsymbol{\zeta}_i)=\mathbf{0}$, and that $(\mathbf{I}-\mathbf{B})$ is invertible. Exogenous variables are variables that are not explained within the model and that are uncorrelated with all disturbances in the system. Endogenous variables are ones that are directly influenced by at least one other variable besides its disturbance. Two covariance matrices are part of the latent variable model: $\boldsymbol{\Phi}$ the covariance matrix of the exogenous latent variables ($\boldsymbol{\xi}$) and $\boldsymbol{\Psi}$ the covariance matrix of the equation disturbances ($\boldsymbol{\zeta}$). The mean of $\boldsymbol{\xi}$ is $\boldsymbol{\mu}_{\boldsymbol{\xi}}$.

The measurement model shows the relation between the latent to the observed responses (indicators). It has two equations:

$$\mathbf{y}_i=\boldsymbol{\alpha}_y+\boldsymbol{\Lambda}_y\boldsymbol{\eta}_i+\boldsymbol{\varepsilon}_i \tag{2}$$

$$\mathbf{x}_i=\boldsymbol{\alpha}_x+\boldsymbol{\Lambda}_x\boldsymbol{\xi}_i+\boldsymbol{\delta}_i \tag{3}$$

where $\mathbf{y}_i$ and $\mathbf{x}_i$ are vectors of the observed indicators of $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$, respectively, $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_{\mathbf{x}}$ are intercept vectors, $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ are matrices of factor loadings or regression coefficients giving the impact of the latent $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ on $\mathbf{y}_i$ and $\mathbf{x}_i$, respectively, and $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\delta}_i$ are the unique

[iv]The intercept notation is a slight modification of the LISREL notation to keep them all as αs.

factors of $\mathbf{y}_i$ and $\mathbf{x}_i$. We assume that the unique factors ($\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\delta}_i$) have expected values of zero, covariance matrices of $\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\Theta}_{\delta}$, respectively, and are uncorrelated with each other and with $\zeta_i$ and $\xi_i$.

## 2 IMPLIED MOMENT MATRICES

Given a model, we can write the population means, variances, and covariances of the observed variables as a function of the parameters. We write this as

$$\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\theta}) \tag{4}$$

$$\sum=\sum(\boldsymbol{\theta}) \tag{5}$$

where $\boldsymbol{\mu}$ is the population mean vector of the observed variables, $\boldsymbol{\theta}$ is the vector that contains the parameters in the model, and $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the population implied mean vector that is a function of the model parameters. Similarly, $\boldsymbol{\Sigma}$ is the population covariance matrix of the observed variables and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the implied covariance structure that is a function of the parameters. The implied moments and their relation to the model parameters forms the basis of the algebraic approach to identification.

To derive the implied moments, it is useful to work with the reduced-form of the models, where each endogenous variable is a function of only exogenous variables, coefficients, and disturbances. The reduced-form for the latent variable model is

$$\boldsymbol{\eta}_i=(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\alpha}_\eta+\boldsymbol{\Gamma}\boldsymbol{\xi}_i+\boldsymbol{\zeta}_i) \tag{6}$$

The measurement model for the $\mathbf{x}_i$ already is in reduced-form. The reduced-form equation for $\mathbf{y}_i$ is

$$\mathbf{y}_i=\boldsymbol{\alpha}_y+\boldsymbol{\Lambda}_y(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\alpha}_\eta+\boldsymbol{\Gamma}\boldsymbol{\xi}_i+\boldsymbol{\zeta}_i) \tag{7}$$

More specifically, in the model in equations (1), (2), and (3) the population implied mean vector is

$$\boldsymbol{\mu}(\boldsymbol{\theta})=\begin{bmatrix} \boldsymbol{\alpha}_y+\boldsymbol{\Lambda}_y(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\alpha}_\eta+\boldsymbol{\Gamma}\boldsymbol{\mu}_\xi) \\ \boldsymbol{\alpha}_x+\boldsymbol{\Lambda}_x\boldsymbol{\mu}_\xi \end{bmatrix} \tag{8}$$

The top part of the right-hand side vector is the implied mean vector for $\mathbf{y}$ and the lower part is the implied mean vector of $\mathbf{x}$.

The implied covariance matrix is fairly complex, so we partition the matrix to correspond to the implied covariance matrix of $\mathbf{y}$, of $\mathbf{x}$, and of $\mathbf{y}$ and $\mathbf{x}$:

$$\sum(\boldsymbol{\theta})=\begin{bmatrix} \sum_{\mathbf{yy}}(\boldsymbol{\theta}) & \sum_{\mathbf{yx}}(\boldsymbol{\theta}) \\ \sum_{\mathbf{xy}}(\boldsymbol{\theta}) & \sum_{\mathbf{xx}}(\boldsymbol{\theta}) \end{bmatrix} \tag{9}$$

These parts of the implied covariance matrix are

$$\sum_{\mathbf{xx}}(\boldsymbol{\theta})=\boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Lambda}_x'+\boldsymbol{\Theta}_\delta \tag{10}$$

$$\sum_{\mathbf{xy}}(\boldsymbol{\theta})=\boldsymbol{\Lambda}_x\boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Lambda}_y' \tag{11}$$

$$\sum_{\mathbf{yy}}(\boldsymbol{\theta})=\boldsymbol{\Lambda}_y(\mathbf{I}-\mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'+\boldsymbol{\Psi})(\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Lambda}_y'+\boldsymbol{\Theta}_\varepsilon \tag{12}$$

These equations show the variance and covariance of the observed variables as functions of the model parameters. The equations are sufficiently general to capture most SEMs with continuous variables.[v] We just substitute the specific matrices for a particular model into these expressions.

## 3 LOCAL AND GLOBAL IDENTIFICATION

*Global identification* of the parameters ($\boldsymbol{\theta}$) in a SEM holds if there are no vectors of values for the parameters, say $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, such that $\boldsymbol{\mu}(\boldsymbol{\theta}_a)=\boldsymbol{\mu}(\boldsymbol{\theta}_b)$ or $\boldsymbol{\Sigma}(\boldsymbol{\theta}_a)=\boldsymbol{\Sigma}(\boldsymbol{\theta}_b)$ unless $\boldsymbol{\theta}_a=\boldsymbol{\theta}_b$.[vi] This means that once we define the parameter space, we will not be able to find two different vectors of parameter values that lead to the same values for the implied means or for the implied covariance matrix. If we were able to find two or more sets of values, then the model is globally underidentified.

In contrast, *local identification* is a weaker concept of uniqueness. The parameter vector $\boldsymbol{\theta}$ is locally identified at a point $\boldsymbol{\theta}_a$, if in the neighborhood of $\boldsymbol{\theta}_a$ there is no vector $\boldsymbol{\theta}_b$ for which $\boldsymbol{\mu}(\boldsymbol{\theta}_a)=\boldsymbol{\mu}(\boldsymbol{\theta}_b)$ or $\boldsymbol{\Sigma}(\boldsymbol{\theta}_a)=\boldsymbol{\Sigma}(\boldsymbol{\theta}_b)$ unless $\boldsymbol{\theta}_a=\boldsymbol{\theta}_b$ (Bollen, 1989, p.248). A key difference between global and local identification is that local identification is examined at a particular set of values, $\boldsymbol{\theta}_a$, and it only searches for alternative values ($\boldsymbol{\theta}_b$) in the neighborhood of $\boldsymbol{\theta}_a$. Global identification searches more broadly, not just the neighborhood of one set of values, to ensure that there are no two sets of different parameter values that lead to the same implied moment matrices. Global identification implies local identification, but a locally identified model need not imply a globally identified model.

We can illustrate this difference with a simple factor analysis model with one factor, three indicators, and uncorrelated unique factors. To simplify the illustration we assume all variables are deviated from their means so that we can ignore the means and intercepts. The matrices for this model are:

[v]In some special models it can be more convenient to use alternative forms of this model to capture unusual structures. One such model is called the "all y" model where $\mathbf{y}$ contains all observed variables and $\boldsymbol{\eta}$ has all of the latent variables. See, for example, Bollen (1989, 395–400).

[vi]For a more formal statement of the conditions for global and local identification see Bekker et al. (1994, pgs. 16–19) or Davidson and MacKinnon (1993, pg. 143).

$$\Lambda_{\mathbf{x}}=\begin{bmatrix}\lambda_{11}\\ \lambda_{21}\\ \lambda_{31}\end{bmatrix} \quad diag\,\boldsymbol{\Theta}_{\boldsymbol{\delta}}=[\Theta_{\delta 11}\Theta_{\delta 22}\Theta_{\delta 33}] \quad \boldsymbol{\Phi}=[\Phi_{11}] \tag{13}$$

We have not yet scaled the latent factor. If we choose to scale the latent variable by setting its variance to one ($\Phi_{11}$=1), then each factor loading has two possible solutions. For example, $\lambda_{11}=(\sigma_{21}\sigma_{31}/\sigma_{32})^{1/2}$ means that we can have the positive or negative square root solution. The same is true for the remaining factor loadings. As is easy to verify, using all the positive square roots for the factor loadings or all of the negative roots of the factor loadings will lead to the identical implied covariance matrix. Thus we have two different values of the factor loadings that result in the same implied covariance matrix and the model is *not* globally identified. However, the model is locally identified. Interestingly, if a researcher scales the latent variable by setting $\lambda_{11}$ to one and estimating $\Phi_{11}$, this model is locally and globally identified. The difference in local and global identification is rather benign in this case, but it could generate misleading results in more complex models.

This one factor, three indicator example also gives us an opportunity to point out that some values of a parameter might create an underidentified model. For instance, if any of the factor loadings or if $\varphi_{11}$ were zero, then the model would be underidentified despite having three indicators. These “degenerate” solutions are a possibility, but a remote one for most applications. When we describe a model as identified, we are assuming that such degenerate values are not present.

Sample estimates that do not equal but are close to the problematic values can create what are called empirical identification issues. A factor loading near zero in the previous example could create empirical identification problems. The computer algebra methods that we discuss could be helpful in better understanding the sources of empirical underidentification, but our focus is on the population values of parameters not on sample estimates.

Two methods to check local identification in SEM are known. The best known method is checking the singularity of the information matrix evaluated at the sample estimates of the parameters. Keesling (1972) and Wiley (1973) recommended this local identification check and it has been implemented in all of the major SEM packages. The information matrix is minus the expected value of the second-order partial derivatives of the maximum likelihood estimator with respect to $\boldsymbol{\theta}$. The $\boldsymbol{\theta}$ parameter is locally identified at $\boldsymbol{\theta}_a$, if and only if the inverse of the information matrix exists (Rothenberg, 1971). The asymptotic covariance matrix of the parameter estimators is routinely computed from the inverse of the estimated information matrix in SEM software. Because of this, it is easy to check whether this condition of local identification holds at the parameter estimates, $\boldsymbol{\theta}$. In fact, this information matrix check is automatic in SEM software.

A second check on local identification is based on Wald (1950). To explain the Wald Rank Rule, we define $\boldsymbol{\sigma}(\boldsymbol{\theta})$ to be a vector of the nonredundant elements of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\mu}(\boldsymbol{\theta})$. The parameter vector $\boldsymbol{\theta}$ is locally identified at a point $\boldsymbol{\theta}_a$, if and only if the rank of $\left(\frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$ evaluated at $\boldsymbol{\theta}_a$ is equal to the number of parameters in $\boldsymbol{\theta}$. None of the SEM software packages implements this check using $\boldsymbol{\theta}$ as a point of evaluation in a way analogous to the information matrix check. McDonald and Krane (1979) provide some simulation evidence that favors the accuracy of this approach compared to the information matrix check.

When either the singularity of information matrix or the Wald Rank Rule are applied using the sample estimates ($\boldsymbol{\theta}$), there are the limitations mentioned in the introduction: singularity

or rank might differ at sample estimates vs. population parameter, numerical accuracy could affect the results, and both are local identification checks not global identification. Computer algebra systems (CASs) can address the first two issues when symbols replace the estimated values. In fact, Bekker, Merckens, and Wansbeek (1994) propose using a CAS to assess local identification based on the Wald Rank Rule. They use the Jacobian matrix of the covariance matrix and the matrix of restrictions on parameters and examine the rank of the resulting matrix to check for local identification. More specifically, they define $\boldsymbol{\sigma}(\boldsymbol{\theta})$ to equal the partitioned column vector of $\boldsymbol{\mu}(\boldsymbol{\theta})$ and vech $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\rho}(\boldsymbol{\theta})$ is a vector of constraints on $\boldsymbol{\theta}$ imposed by a model such that $\boldsymbol{\rho}(\boldsymbol{\theta}^o)=\mathbf{0}$. They define the Jacobian matrix as formed by

$$J(\boldsymbol{\theta})=\begin{bmatrix} \partial\boldsymbol{\sigma}(\boldsymbol{\theta})/\partial\theta' \\ \partial\boldsymbol{\rho}(\boldsymbol{\theta})/\partial\theta' \end{bmatrix} \tag{14}$$

and show that if $\boldsymbol{\theta}^o$ is a regular point of $J(\boldsymbol{\theta})$, then a necessary and sufficient condition for a locally identified $\boldsymbol{\theta}^o$ is that $J(\boldsymbol{\theta})$ has rank equal to the dimension of $\boldsymbol{\theta}$ (Bekker, et al., 1994, pp. 27–28). The advantage of their CAS approach is that the evaluation occurs with the symbols rather than the estimates so that the potential confounding due to sample estimates in checking the estimated information matrix is removed.

The Bekker, et al. (1994) approach was the first systematic application to CAS checks on local identification in SEM and we see it as a valuable supplement to the usual check on the singularity of the estimated information matrix. Though the Bekker, et al. (1994) book and software are out-of-print, CAS software are available to write code to provide this check as we will illustrate.

In the next section we explain another application of CAS to directly examine the identification of parameters using the implied moment matrices.

## 4 Algebraic Solution

In principal one can assess the global identification of a model through a manual check by substituting the parameters into the implied mean and covariance matrices and attempting to algebraically solve each parameter in terms of observed means, covariances and variances and establishing that the solutions are unique. In practice, the algebra is often difficult. Given the advances in CASs, such as Maple and Mathematica, in the ability to solve nonlinear systems of equations, it is now possible with a minimum of programming to find explicit algebraic solutions, if they exist, for many SEMs (see Geddes, Czapor, and Labahn (1992) for a discussion of the algorithms Maple employs to solve nonlinear systems of equations).

### 4.1 General Steps

Before using a CAS the analyst must first determine the number of equations and the number of parameters given the number of observed variables and the proposed model. Let $p$ be the number of $y$ observed variables and $q$ the number of $x$ observed variables. In many models, the means of the observed variables and the intercepts in the latent variable and measurement models are ignored. Attention focuses on the covariance matrix and other model parameters. To simplify our discussion, we concentrate on the implied covariance matrix and the model parameters that are part of it, though it is straightforward to add the means and intercepts to these procedures.

The covariance matrix of observed variables, Σ, will have $(\frac{1}{2})(p+q)(p+q+1)$ nonredundant elements. [vii] Let $t$ equal the number of unknown parameters in the parameter vector θ. The three possible cases that can arise are

$$\begin{aligned} Case1 &: t>(\tfrac{1}{2})(p+q)(p+q+1) \\ Case2 &: t=(\tfrac{1}{2})(p+q)(p+q+1) \\ Case3 &: t<(\tfrac{1}{2})(p+q)(p+q+1). \end{aligned}$$

In case 1 where there are a greater number of unknown parameters in θ than there are nonredundant elements of Σ, by the *t*-rule (Bollen, 1989, p.328) one can see immediately that the proposed model is underidentified and no further work is required. viii In case 2, the number of unknown parameters in θ is equal to the number of nonredundant elements of Σ, which leads to an equal number of equations and unknowns in the resulting system of equations. Finally in case three where the number of unknown parameters in θ is less than the number of nonredundant elements of Σ the resulting system of equations will have more equations than unknown parameters (i.e., in mathematical terms it is an overdetermined system), raising the possibility that more than one way to solve for a parameter exists. Due to how CASs operate, cases 2 and 3 require somewhat different approaches and will be discussed separately.

**4.1.1 Case 2—**The first step in assessing identification with a CAS is to enter the system of equations implied by the observed covariance matrix and the model under consideration. We have found this is most easily accomplished by defining partitions of the implied covariance matrix (equations 12–14) as functions and then substituting in the model parameters (see Appendix for a sample program). When there are an equal number of equations and unknowns in the resulting system, the analyst can immediately proceed to obtaining an algebraic solution, if it exists.

In solving the system it is possible to find three different types of results: (1) a unique algebraic solution is found for each parameter, (2) an algebraic solution is found that is not unique for one or more parameters (e.g., the solution for a parameter may involve a square root, thus admitting a positive and a negative solution), (3) no algebraic solution is found. [ix] If the first result is obtained, then the model is globally identified. If the second type of result is obtained, then the model is locally identified. In this case, one can examine which parameters admit multiple solutions and potentially assess the significance of this for the specific model under consideration. Finally, if the third type of result is obtained, then the model is not identified.

It is important to keep in mind that there are situations where even though a parameter is equal to a specific function of the variances and covariances of the observed variables, the variances or covariances might take values resulting in an undefined solution. For instance, if a covariance is the denominator of a solution and that covariance of observed variables is zero in the population, then an undefined solution would result and that parameter and hence

[vii] If the vector of means is used in addition to the covariance matrix, then there are $(\frac{1}{2})(p+q)(p+q+3)$ nonredundant elements.
[viii] It may be possible to impose inequality constraints on the model such that it can be identified, but this rarely occurs in practice.
[ix] It is conceivable that there is a solution, but the CAS program has failed to find it. In all of the examples that we have tried, we never encountered that problem though it remains a possibility.

the model would not be identified. The CAS approach makes it easier to see these points of undefined solutions by giving the explicit functions that a parameter equals.

**4.1.2 Case 3—**In most situations there will be more equations than parameters (in fact, it is this difference that allows for tests of model fit). When this is the case, we are not aware of a direct method to instruct a CAS to solve the system. From a purely mathematical viewpoint, the system in fact does not have a single solution as it is overdetermined. This means that it might be possible to have a parameter written as two or more different functions of the variances or covariances of the observed variables. But if the model is valid, then these solutions should result in the same numerical value for the parameter when the population variances or covariances values are substituted in. So being overdetermined does not mean that different model parameter values result. It just means that it might be possible to arrive at the same parameter value in multiple ways.

An overdetermined system does raise problems for a CAS. A practical way to approach this is to try to turn Case 3 into a Case 2 situation. More specifically, the analyst should find a subset of equations containing all of the unknown parameters and equal to the number of unknown parameters. Then, Case 3 reduces to Case 2 and the subset of equations can be solved for an explicit algebraic solution, if it exists. As with Case 2, it is possible to obtain three different types of results, but one is now faced with the added complication that this is for a subset of the model equations and it is possible to obtain different types of results with other subsets of equations. In a companion piece (Bollen and Bauldry, forthcoming), we provide a proof that if a unique algebraic solution is found with one subset of equations, this is sufficient to establish global identification. It is not the case, however, that if either a non-unique algebraic solution is found or no solution is found with a given subset of equations then the model is respectively locally identified or not identified. In practice, this means that one needs to solve different subsets of equations until either a unique algebraic solution is found with a given subset, a multiple solution results in generating the same implied moment matrices, or all possible subsets of equations are exhausted.

The number of subsets of equations equals the number of unknown parameters taken t at a time or the binomial expression:

$$(\frac{1}{2})(p+q)(p+q+1)t, \tag{15}$$

This potentially could result in a large number of subsets of equations to evaluate. However, there are factors that reduce this number. Fortunately, we can eliminate any subset of equations that does not include all model parameters. For many models this substantially reduces the number of systems of subsets of equations to be solved.

There is one other way in which we can reduce the number of equations. [x] We know that the variances of unique factors or errors only appear in the variances of the observed variables to which they correspond. This means that only if the other parameters in the implied variance equations are identified will we be able to identified the error variances. This implies that we can reduce the number of equations that we simultaneously consider by eliminating all the variances of observed variables and eliminate the variances of the unique factors as a way of reducing the total number of equations to solve. If we are able to solve for all the other parameters, we know that we will be able to solve for the unique factor variance. So if a researcher encounters any difficulty in solving for all parameters, it is

[x]We thank a reviewer for suggesting this.

possible to reduce the size of the problem by eliminating these variances and solving for the other parameters.

In practice we have found unique algebraic solutions within the first few subsets of equations we checked for models that are globally identified. Finally, we know of one additional shortcut. If a given subset of equations returns a non-unique algebraic solution, then one can check whether the solution violates the definition of identification numerically and, if so, conclude that the model is only locally identified. We propose the following algorithm for conducting a numerical check:

1. Choose a set of numerical values for all of the parameters.
2. Calculate the values of the moment matrices implied by the values of the parameters.
3. Using the same subset of equations that generated a non-unique algebraic solution, solve for the numeric values of the parameters. Given that the original algebraic solution was not unique, this will result in at least two sets of values for the parameters. By definition, one of the sets will match the values chosen in step 1.
4. Substitute any set of parameter values that does not match those chosen in step 1 into $\Sigma=\Sigma(\theta)$ and $\mu=\mu(\theta)$. Compare the results with step 2.

By the definition of identification, if two sets of parameter values generate the same covariance matrix, then the model is not globally identified. Therefore, if in step 4 one finds the same covariance matrix as in step 2, then the model is locally identified. If, however, in step 4 one does not obtain the same covariance matrix as in step 2, then nothing can be concluded and additional subsets of equations need to be assessed. We have not encountered a situation where it was necessary to examine all possible solutions, though this does not guarantee it will never occur.

## 5 Identification Procedure

In the previous sections we have presented CAS approaches to local and global identification. This section briefly describes how these alternative methods can be part of an effective identification strategy in SEM applications. We put this procedure in a flow chart in Figure 2. As a starting point, we recommend that a researcher apply any of the available rules of identification. If a model conforms to the conditions of an established *sufficient* rule of identification, then the model is identified. If it fails a *necessary* condition, then the model is not identified. Unfortunately, a number of models are not covered by existing rules. We then recommend the CAS approach to Wald's Rank Rule (or the Jacobian) to provide information on whether local identification holds. Passing this CAS check is a good safeguard to check local identification. A rank less than the number of parameters indicates the failure of identification. Satisfying this CAS check on the rank provides assurance of local identification. Failing it means the model is underidentified. Finally, a researcher can apply CAS to find unique solutions for the parameters in terms of the means, variances, and covariances as we have demonstrated.

We illustrate this procedure with several examples in the next section.

## 6 Examples of Identification Procedure

In this section we work through the proposed identification procedure for four models. We strategically chose models that illustrate a variety of situations an analyst might encounter. Our first example, a model Duncan (1975) explored, demonstrates that our approach to identification does not fall into the potential pitfalls of algebraically solving models by hand.

We chose our second example, a model Bollen (1989) originally examined, to illustrate our approach to overidentified models. For our third example, we examine the identification status of a recently proposed model, Enders (2008) auxiliary variable model for missing data. Finally, in our last example we return to Bollen and Hoyle (1990) from the introduction and illustrate how our approach determines that this model is locally, but not globally identified.

### 6.1 Duncan's (1975) Model

Duncan (1975, pages 84–86) provides a cautionary note on the use of algebraic means to establish model identification. He presents an underidentified model where algebraic solutions appear to be available for all parameters (see Figure 3). This is an important example to examine since it suggests that the CAS approach to algebraic solutions might fail in that it would suggest that some underidentified models were identified. [xi]

As a system of equations, we can express this model as:

$$\begin{aligned} y_1 &= \beta_{12}y_2+\gamma_{11}x_1+\zeta_1 \\ y_2 &= \beta_{21}y_1+\gamma_{21}x_1+\zeta_2 \\ y_3 &= \beta_{31}y_1+\beta_{32}y_2+\gamma_{32}x_2+\zeta_3. \end{aligned} \tag{16}$$

This is a simultaneous equation model with a feedback relationship between $y_1$ and $y_2$ as well as a correlation between the disturbances of $y_1$ and $y_2$. Duncan demonstrates that through a simple process of substitution one can obtain algebraic solutions for all of the γs and βs. He notes, however, that a close inspection of the solutions reveals that $\beta_{21}=1/\beta_{12}$ and $\gamma_{21}=-\gamma_{11}/\beta_{12}$ and therefore that the model is not in fact identified despite the apparent set of algebraic solutions. In other words, the solutions are not *unique* as required for identification.

Duncan's result raises the possibility that one might encounter similar problems using a CAS, so it is instructive to see how our identification procedure fares with this model (see flow chart in Figure 2). First, Duncan (1975) and Rigdon (1995) note that this is a block-recursive model (Fisher, 1966) and that it is not identified according to this rule. Thus, the problem with this model would be caught in the first step of our identification procedure, but for illustrative purposes we continue with our identification procedure.

The second step of our identification procedure assesses Wald's Rank Rule with a CAS. We find that the rank of the Jacobian is 12 whereas we have 14 parameters in the model, and therefore we see that the model is not locally identified. Having determined that the model is not locally identified there is no reason to proceed to attempt to find algebraic solutions for the parameters. Nonetheless, we do so to see if the CAS would find the potentially misleading algebraic results. It does not. When we examine all 15 subsets of 14 equations for a solution, the CAS returns a null result (i.e., it correctly determines that there is no solution). With this example, we see that our procedure for assessing identification catches the fact that this model is not identified at every step.

### 6.2 A Model of Subjective Class

For our second example we draw on a model of subjective class that Bollen (1989, pages 172–175) uses to explore the consequences of allowing for measurement error (see Figure 4). The model is nonrecursive with actual income ($x_1$) influencing subjective income ($\eta_1$) and actual occupational prestige ($x_2$) impacting subjective occupational prestige ($\eta_2$). There

[xi] We thank a reviewer who suggested this possibility.

is a feedback relation between subjective income and subjective occupation with both of these variables affecting subjective overall status ($\eta_3$). With the variables in deviation form this model is

$$\begin{aligned}
\eta_1 &= \beta_{12}\eta_2+\gamma_{11}\xi_1+\zeta_1\\
\eta_2 &= \beta_{21}\eta_1+\gamma_{22}\xi_2+\zeta_2\\
\eta_3 &= \beta_{31}\eta_1+\beta_{32}\eta_2+\zeta_3\\
y_1 &= \eta_1+\varepsilon_1\\
y_2 &= \eta_2+\varepsilon_2\\
y_3 &= \eta_3\\
x_1 &= \xi_1\\
x_2 &= \xi_2.
\end{aligned}$$

None of the published rules of identification cover this model, [xii] so our next step is to check whether the model passes Wald's Rank Rule. We find that the rank of the Jacobian is 14 and we have 14 unknown parameters (in addition to the coefficients in the system of equations above, we also have $\varphi_{11}$, $\varphi_{21}$, $\varphi_{22}$, $\theta_{\varepsilon 1}$, $\theta_{\varepsilon 2}$, $\psi_{11}$, $\psi_{22}$, and $\psi_{33}$). This indicates that the model is locally identified. We now turn to an assessment of whether the model is globally identified.

Ignoring the means and intercepts, we have 15 nonredundant elements in $\Sigma$ and 14 unknown parameters. With more nonredundant elements in $\Sigma$ than unknown parameters we have an example of our third case and need to work with subsets of equations. For this model we have 1514=15 subsets of equations to consider, though not all of the subsets contain each parameter in at least one equation. To simplify notation we array the observed variables $y_1$ to $y_3$ followed by $x_1$ and $x_2$. This allows us to reference specific elements of the observed covariance matrix by subscripts (e.g., $COV(x1,y2)=\sigma_{42}$). To select a subset of equations we first make sure to include all of the variances ($\sigma_{11}$, $\sigma_{22}$, $\sigma_{33}$, $\sigma_{44}$, $\sigma_{55}$). Second, we include the covariance between $x_1$ and $x_2$ ($\sigma_{54}$) in order to include $\varphi_{21}$. Third, we arbitrarily select eight additional covariances ($\sigma_{21}$, $\sigma_{31}$, $\sigma_{41}$, $\sigma_{51}$, $\sigma_{42}$, $\sigma_{52}$, $\sigma_{43}$, $\sigma_{53}$). Solving this subset of equations we obtain a unique algebraic solution for all of the parameters. As we mention above, the fact that we are able to obtain a unique algebraic solution for all of the parameters with one subset of equations is sufficient to establish that this model is globally identified.

To illustrate a few of the solutions, we find

$$\begin{aligned}
\gamma_{11} &= \frac{\sigma_{42}\sigma_{51}-\sigma_{41}\sigma_{52}}{\sigma_{42}\sigma_{54}-\sigma_{44}\sigma_{52}},\\
\beta_{21} &= \frac{\sigma_{52}\sigma_{54}-\sigma_{42}\sigma_{55}}{\sigma_{51}\sigma_{54}-\sigma_{41}\sigma_{55}},\\
\beta_{31} &= \frac{\sigma_{42}\sigma_{53}-\sigma_{43}\sigma_{52}}{\sigma_{42}\sigma_{51}-\sigma_{41}\sigma_{52}}.
\end{aligned}$$

These equations provide insight into values of variances or covariances that could underidentify a parameter or create estimation difficulties. For instance, if $\sigma_{42}\sigma_{54}=\sigma_{44}\sigma_{52}$, then the solution for $\gamma_{11}$ is undefined due to a zero in the denominator. Even if this denominator is not exactly zero in the population, there would likely be estimation problems if it is near zero. Similar observations hold for the other denominators. Without these explicit solutions for the parameters, these potential pitfalls would not be obvious.

[xii] Bollen and Bauldry (forthcoming) provide a new rule of identification that would cover this example. Note also that if we allowed a disturbance in the $y_3$ equation, we could not separately identify the measurement error variance in $y_3$ and the disturbance variance for $\eta_3$. A more realistic perspective on this model would treat the error variance of $\zeta_3$ consisting of both the equation disturbance ($\zeta_3$) and the measurement error in $y_3$.

### 6.3 Auxiliary Variables and Missing Data Model

A direct maximum likelihood approach to missing data is widely available in SEM software (Arbuckle 1996). In this approach a researcher includes the substantive variables of the model as well as auxiliary variables that predict the missingness in the data such that the data are Missing at Random (MAR) when both sets of variables are in the model. Several authors have discussed the importance of including the auxiliary variables to satisfy MAR. Enders (2008), for example, presents the model in Figure 5 Panel A to illustrate a SEM with auxiliary variables for the direct ML approach to missing data. Researchers have not examined the identification of such models. This model is not covered by rules of identification so it makes for an interesting and timely case to examine.

In the auxiliary model Enders (2008) proposes, he depicts the auxiliary variable as correlated with the error terms of all of the observed variables. In order to translate this into a standard form for the implied moment matrices we make a few adjustments (see Figure 5, Panel B). First, we treat the auxiliary variable as a latent exogenous variable that is perferctly measured by the observed auxiliary variable. Second, we treat all of the disturbances of the observed variables as latent exogenous variables. These two adjustments allow us to represent the correlations among the auxiliary variable and the disturbances as covariances among latent exogenous variables. Third, we create three endogenous latent variables that are perfectly measured by the endogenous observed variables. This allows for a consistent treatment of the disturbances of the endogenous variables. These adjustments have no effect on the overall parameterization of the model. With these adjustments, we can write the system of equations for this model as:

$$\begin{aligned}
\eta_1 &= \eta_4+\xi_5 \\
\eta_2 &= \alpha_{\eta_2}+\beta_{24}\eta_4+\xi_6 \\
\eta_3 &= \alpha_{\eta_3}+\beta_{34}\eta_4+\xi_7 \\
\eta_4 &= \alpha_{\eta_4}+\gamma_{48}\xi_8+\zeta_4 \\
x_1=\xi_2+\xi_8 & \qquad y_1=\eta_1 \\
x_2=\alpha_{x_2}+\xi_3+\lambda_{28}\xi_8 & \qquad y_2=\eta_2 \\
x_3=\alpha_{x_3}+\xi_4+\lambda_{38}\xi_8 & \qquad y_3=\eta_3 \\
x_4=\xi_1. &
\end{aligned}$$

Checking Wald's Rank Rule we find that the Jacobian has a rank of 27 and we have 27 parameters to estimate in this model (in addition to the parameters in the system above, we have 14 variances and covariances among the latent exogenous variables, the variance of $\zeta_4$, and the means of $\xi_1$ and $\xi_8$ to identify). This establishes that the model is locally identified, so we turn to determining whether there is an algebraic solution for all of the parameters.

In this model we have 7 observed variables and therefore 28 nonredundant variances and covariances and 7 means with which to work. We only have 27 parameters to identify, so we again have an example of our third case. We adopt the same approach as in our last example and array the variables $y_1$ to $y_3$ and then $x_1$ to $x_4$. Selecting a subset of equations to solve is not too difficult for this model because most of the equations involve only a few parameters. We started by selecting all of the variances and all of the means ($\sigma_{11}$, $\sigma_{22}$, $\sigma_{33}$, $\sigma_{44}$, $\sigma_{55}$, $\sigma_{66}$, $\sigma_{77}$, $\mu_{y1}$, $\mu_{y2}$, $\mu_{y3}$, $\mu_{x1}$, $\mu_{x2}$, $\mu_{x3}$, $\mu_{x4}$). Then, by inspection, it is clear we need to include the covariances between the auxiliary variable and all other variables ($\sigma_{71}$, $\sigma_{72}$, $\sigma_{73}$, $\sigma_{74}$, $\sigma_{75}$, $\sigma_{76}$). Finally, we choose the final seven equations such that any remaining parameters are included at least once ($\sigma_{21}$, $\sigma_{41}$, $\sigma_{51}$, $\sigma_{61}$, $\sigma_{42}$, $\sigma_{43}$, $\sigma_{54}$). Solving this subset of equations we find a unique algebraic solution for all of the parameters and thus that the model is globally identified. For instance, we find that

$$\begin{aligned} \lambda_{28} &= \frac{\sigma_{51}}{\sigma_{41}}, \\ \beta_{34} &= \frac{\sigma_{43}}{\sigma_{41}}, \\ \gamma_{48} &= \frac{\sigma_{51}}{\sigma_{54}}, \\ \alpha_{\eta_4} &= \mu_{y_1} - \mu_{x_1}\frac{\sigma_{51}}{\sigma_{54}}. \end{aligned} \tag{17}$$

The first two equations show the critical role that $\sigma_{41}$ plays in that if it is zero, then these solutions are not defined; if it is near zero, it is likely to cause estimation difficulties.

This example not only illustrates our procedure but provides the first proof that the auxiliary variable approach to missing data such as in Figure 4 is identified.

### 6.4 Bollen and Hoyle (1990) Model

We opened our paper with a model taken from Bollen and Hoyle (1990, p. 499, Figure 2 part (3)) where SEM software revealed no identification problem yet depending on starting values a researcher could end up with two different sets of values for a subset of the parameters. The model is repeated in Figure 6 and the equations for the model are below where the two latent variables are morale and sense of belonging with each of these measured with three indicators.

$$\begin{aligned} \eta_1 &= \beta\eta_2+\zeta_1 \\ \eta_2 &= \beta\eta_1+\zeta_2 \\ y_1 &= \eta_1+\varepsilon_1 \\ y_2 &= \lambda_{21}\eta_1+\varepsilon_2 \\ y_3 &= \lambda_{31}\eta_1+\varepsilon_3 \\ y_4 &= \eta_2+\varepsilon_4 \\ y_5 &= \lambda_{52}\eta_2+\varepsilon_5 \\ y_6 &= \lambda_{62}\eta_2+\varepsilon_6 \end{aligned} \tag{18}$$

This model is not covered by rules of identification. As we mentioned in the introduction, this model passes the check on the singularity of the information matrix implemented in SEM software. As a further test, we check whether this model passes Wald's Rank Rule. We find that the Jacobian has rank 13 and there are 13 parameters to estimate (in addition to the parameters in the system of equations we have $\psi_{11}$, $\psi_{22}$, and $\theta_{\varepsilon 1}$–$\theta_{\varepsilon 6}$), so the model is clearly locally identified.

In this model we have 6 observed variables and therefore $(\frac{1}{2})(6)(7)=21$ nonredundant elements in $\Sigma$. For the proposed model we have 13 unknown parameters to identify. Since we have quite a few more equations than unknown parameters, this corresponds to case 3. To select a subset of equations we start with all of the variances ($\sigma_{11}$, $\sigma_{22}$, $\sigma_{33}$, $\sigma_{44}$, $\sigma_{55}$, $\sigma_{66}$) and then select seven additional covariances that includes all of the parameters in at least one equation ($\sigma_{21}$, $\sigma_{32}$, $\sigma_{43}$, $\sigma_{54}$, $\sigma_{65}$, $\sigma_{31}$, $\sigma_{64}$). Solving this subset we obtain an algebraic solution, but the solution is not unique for $\beta$, $\psi_{11}$, and $\psi_{22}$. For instance, for $\beta$ we find

$$\beta=\frac{\sigma_{32}\sigma_{54}\sigma_{64}+\sigma_{21}\sigma_{31}\sigma_{65} \pm \sqrt{4\sigma_{21}^2\sigma_{43}^2\sigma_{65}^2+(-\sigma_{32}\sigma_{54}\sigma_{64}-\sigma_{21}\sigma_{31}\sigma_{65})^2}}{2\sigma_{21}\sigma_{43}\sigma_{65}}. \tag{19}$$

The fact that one subset of equations leads to a non-unique algebraic solution is not sufficient to determine the identification status of the model. In principle, we could continue to solve other subsets of equations to see if we can find one which leads to a unique algebraic solution. There are, however, quite a few subsets to consider, so instead we adopt a numeric check that can demonstrate that a model is not globally identified.

Following the algorithm outlined above, we choose the following numeric values for the parameters:

$$\mathbf{B}=\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix} \quad \mathbf{\Psi}=\begin{bmatrix} 1.0 & 0 \\ 0 & 2.0 \end{bmatrix}$$

$$\mathbf{\Lambda}_y=\begin{bmatrix} 1.0 & 0 \\ 1.2 & 0 \\ 0.8 & 0 \\ 0 & 1.0 \\ 0 & 0.8 \\ 0 & 1.2 \end{bmatrix} \quad \mathbf{\Theta}_\varepsilon=\begin{bmatrix} 1.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 \end{bmatrix}. \tag{20}$$

These values generate the following implied covariance matrix (rounded to two digits):

$$\begin{array}{l} 3.20\&4.34\&\&\&\& \qquad 3.67 \\ 2.13\&2.56\&2.21\&\&\& \\ 2.67\&3.20\&2.13\&4.50\&\& \\ 2.13\&2.56\&1.71\&3.20\&3.56\& \\ 3.20\&3.84\&2.56\&4.80\&3.84\&6.26. \end{array} \tag{21}$$

Using this numeric covariance matrix and solving for all of the parameters we obtain, as expected, two sets of parameter values. One set corresponds to the parameter values we chose. In the other set, the λs and θs are identical (again, as we would expect given the unique algebraic solution for these parameters), but $\beta=2.00$, $\psi_{11}=8.05$, and $\psi_{22}=4.02$. Substituting this second set of parameter values into $\Sigma=\Sigma(\theta)$, we obtain the same numerical covariance matrix (within rounding error). This demonstrates that this model is not globally identified. Furthermore the solution for β reveals that $\sigma_{21}$, $\sigma_{43}$, and $\sigma_{65}$ play an important role in that if any of these covariances are zero, then the denominator would be zero and the solution undefined.

## 7 CONCLUSIONS

Identification remains an issue in SEMs. Fortunately, most SEM software has empirical checks on local identification of a model at the estimated values of the parameters. But as we explained in this paper, this check works well most of the time, but can fail or create ambiguous situations due to the use of sample estimates rather than population values, the limits of numerical assessments of rank, and the fact that it is a check of local rather than global identification. It is for these reasons that the empirical checks of identification use warning messages suggesting that a parameter *may* not be identified. The tentativeness of these checks and the possibility that a model might be locally but not globally identified suggests that complementary procedures could help.

In this paper we suggested an identification procedure that makes use of identification rules when available and local and global CAS approaches to model identification. The starting and possible ending point are rules of model identification. If a model satisfies the conditions for a sufficient rule of model identification, then researchers should use it and need go no further. Similarly, if a model fails a necessary condition, then researchers know

the model is not identified. In the frequent situation where neither is true, we recommend a local identification check using CAS implementation of the Wald Rank Rule (Jacobian). An advantage of this local identification check over the singular information matrix check is that the CAS approach is done with symbols rather than sample estimates. Bekker, et al. (1994) systematically explored this CAS approach. Even though their book and software are out of print, researchers can use contemporary CAS software to carry out this local identification check using the Wald Rank Rule.

One contribution of our paper is the development of a CAS approach to find unique solutions for each parameter as a way of addressing global identification. The most straightforward situation is when the number of variances, covariances, and means exactly equals the number of model parameters so that we can use CAS to solve for each parameter in terms of these moments. In overdetermined system where there are more means, variances, and covariances than there are parameters in the model, we presented an algorithm to follow that enables checks on model identification. In the examples that we tried, we found this algorithm to be successful. Furthermore, having the solutions for different parameters in terms of the means, variances, and covariances of the observed variables can provide insight into what determines the magnitude of the parameter estimates and what might cause problems. Substituting the sample means, variances, and covariances of the observed variables in for their population counterparts and using different solutions for the same parameters would reveal similarities or discrepancies in different ways of estimating the same parameter. In addition, the CAS approach could show which parameters are underidentified and possibly provide insight into how to identify these parameters.

Despite these advantages, there are limitations to keep in mind. First, if a model is underidentified, this means that at least one parameter is not identified, not that *all* parameters are. Our check on model identification might reveal an underidentified model even when some parameters are identified. The last example illustrates this where the model as a whole is not globally identified even though most of the parameters are. A second limitation is that the user must choose a subset of equations equal to the number of parameters. In overidentified models there might be many subsets of equations to try. Fortunately, we can reduce these possibilities in that each subset of equations needs to contain all of the parameters of a model. However, a search of more than one subset of equations might still be needed.

A third limitation occurs when the model is large and complex with numerous parameters. We see two ways to simplify the CAS programming in such cases. In the common situation when the concerns about identification are concentrated in one part of the model (e.g., a part with feedback relations), researchers can concentrate their identification efforts on that sector of the model. This would include a subset of the observed variables and model parameters, reducing the size of the problem. They could select the parameters of particular interest and then choose variances, covariances, and means of observed variables that include these parameters. Researchers might be able to use CAS to establish identification of this subset of problematic parameters and if successful combine it with knowledge of the identification of the more straightforward sectors of the model.

A second way to approach large models is to combine the CAS approach with established rules of identification. The prime example of this is the Two Step Rule of identification (Bollen, 1989, pages 328–31). This Rule reformulates a latent variable structural equation model first into a measurement model replacing coefficients linking latent variables with associations and examines the identification of the measurement model. If the measurement model is identified, then the second step of the Two Step Rule is to return to the latent variable model and to treat it as if it were a simultaneous equation model of observed

variables. If this part of the model is identified, then the whole model is identified. If the rules of identification apply to the measurement model and the simultaneous equation steps, then there is no need for CAS, But CAS could play a role in identifying either the first step of identifying the measurement model or in the second step of identifying the simultaneous equation model when either is not covered by identification rules. An advantage of this approach is that a complex model is made simpler when broken into two parts as in the Two Step Rule.

In conclusion, just as there is no rule of identification that covers all models, the CAS approach will not solve all identification problems in SEMs. Rather, CAS holds much promise in advancing the identification of SEMs and is a useful complement to the current practice of relying solely on the estimated information matrix singularity check. By passing these CAS checks, a researcher can have greater confidence in the identification of a model.

## Acknowledgments

The authors thank anonymous reviewers for helpful comments on this manuscript and the financial support provided by NSF SES 0617276 and from NIDA DA013148-05A2.

## References

Arbuckle, James L. Full Information Estimation in the Presence of Incomplete Data. In: Marcoulides, GA.; Schumacker, RE., editors. Advanced Structural Equation Modeling: Issues and Techniques. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 1996. p. 243-277.

Bekker, Paul A.; Merckens, Arjen; Wansbeek, Tom J. Identification, Equivalent Models, and Computer Algebra. San Diego, CA: Academic Press; 1994.

Bentler, Peter M.; Weeks, DG. Multivariate Analysis with Latent Variables. In: Krishnaiah, PR.; Kanal, L., editors. Handbook of Statistics. Vol. 2. Amsterdam: North-Holland; 1980. p. 747-771.

Bollen, Kenneth A. Structural Equation Models with Latent Variables. New York: Wiley; 1989.

Bollen, Kenneth A.; Bauldry, Shawn. A Note on Algebraic Solutions to Identification. The Journal of Mathematical Sociology. forthcoming.

Bollen, Kenneth A.; Curran, Patrick J. Autoregressive Latent Trajectory (ALT) Models: A Synthesis of Two Traditions. Sociological Methods & Research. 2004; 32:336–383.

Bollen, Kenneth A.; Hoyle, Rick H. Perceived Cohesion: A Conceptual and Empirical Examination. Social Forces. 1990; 69:479–504.

Burt, Ronald S.; Fischer, Michael G.; Christman, Kenneth P. Structures of Well-Being: Sufficient Conditions for Identification of Restricted Covariance Models. Sociological Methods & Research. 1979; 8:111–120.

Burt, Ronald S.; Wiley, James A.; Minor, Michael J.; Murray, James R. Structure of Well-Being: Form, Content, and Stability Over Time. Sociological Methods & Research. 1978; 6:365–407.

Davidson, Russell; MacKinnon, James G. Estimation and Inference in Econometrics. New York: Oxford University Press; 1993.

Davis, Walter R. The FC1 Rule of Identification for Confirmatory Factor Analysis: A General Sufficient Condition. Sociological methods & Research. 1993; 21:403–437.

Duncan, Otis Dudley. Introduction to Structural Equation Models. New York: Academic Press; 1975.

Enders, Craig K. A Note on the Use of Missing Auxiliary Variables in Full Information Maximum Likelihood-Based Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal. 2008; 15:434–448.

Fisher, Franklin M. The Identification Problem in Econometrics. New York: McGraw Hill; 1966.

Geddes, KO.; Czapor, SR.; Labahn, G. Algorithms for Computer Algebra. Boston: Kluwer Academic Publishers; 1992.

Greene, William H. Econometric Analysis. 6. New York: Prentice Hall; 2008.

Jöreskog, Karl G.; Sörbom, Dag. LISREL 8: User's Reference Guide. Mooresville, IN: Scientific Software Inc; 1993.

Keesling, JW. PhD Dissertation. Department of Education, University of Chicago; 1972. Maximum Likelihood Approaches to Causal Analysis.

Kenny, David A. Correlation and Causality. New York: Wiley; 1979.

Long, J Scott. Confirmatory Factor Analysis: A Preface to LISREL. Beverly Hills: Sage University Press; 1983.

Maple Version 11 for Mac OS X. Waterloo, Ontario: 2007. [computer software]

McDonald RP, Krane WR. A Monte Carlo Study of Local Identifiability and Degrees of Freedom in the Asymptotic Likelihood Ratio Test. British Journal of Mathematical and Statistical Psychology. 1979; 32:121–131.

O'Brien, Robert M. Identification of Simple Measurement Models with Multiple Latent Variables and Correlated Errors. Sociological Methodology. 1994; 24:137–70.

Reilly, Terence. A Necessary and Sufficient Condition for Identification of Confirmatory Factor Analysis Models of Factor Complexity One. Sociological Methods & Research. 1995; 23:421–441.

Rigdon, Edward E. A Neccesary and Sufficient Identification Rule for Structural Models Estimated in Practice. Multivariate Behavioral Research. 1995; 30:359–383.

Rothenberg, Thomas J. Identification in Parametric Models. Econometrica. 1971; 39:577–591.

Schoenberg, Ronald. Identification and the Condition of the Information Matrix in the Maximum Likelihood Estimation of Structural Equation Models. Presented at the 1981 Annual Meeting of the American Sociological Association; 1981.

Wald, Abraham. A Note on the Identification of Econometric Relations. In: Koopmans, TC., editor. Statistical Inference in Dynamic Economic Models. New York: Wiley; 1950. p. 238-244.

Wiley, David E. The Identification Problem for Structural Equation Models with Unmeasured Variables. In: Goldberger, AS.; Duncan, OD., editors. Structural Equation Models in the Social Sciences. New York: Seminar Press; 1973. p. 69-83.

Wolfram Research, Inc. Mathematica Version 6 for Mac OS X. [computer software]. Champaign, IL: Wolfram Research, Inc; 2007.

## Biographies

Kenneth Bollen is Director of the Odum Institute for Research in Social Science and the Immerwahr Professor of Sociology at the University of North Carolina at Chapel Hill. He is a Fellow of American Association for the Advancement of Science and winner of the Lazarsfeld Award for Methodological Contributions from the American Sociological Association. ISI lists him among the World's Most Cited Authors in the Social Sciences. Three of his ASR articles were recognized as among the most cited in ASR's history. He is co-author of Latent Curve Models (2006, Wiley), author of Structural Equation Models with Latent Variables (1989, Wiley) and over 100 papers. Bollen's research areas include structural equation models, population studies, and democratization.

Shawn Bauldry is a graduate student in the departments of Sociology and Statistics at the University of North Carolina at Chapel Hill. Bauldry's research areas include structural equation models, stratification, and the sociology of education.

## 9 Appendix: Mathematica Code

In this appendix we provide a brief overview of the code we used for the examples in this paper. We have examined all of the models using both Mathematica and Maple (the two most popular CASs) (Wolfram Research, Inc. Mathematica 2007; Maple 2007). We present the Mathematica code, but all of the commands have quite similar analogues in Maple.

In all of the examples we begin by writing a function to generate the implied moment matrices.xiii We accomplish this by defining the following equations for different partitions of the implied covariance matrix and the implied mean vector corresponding with *y* and *x* observed variables:

```
YMat[Ly_, Ph_, G_, Id_, B_, Ps_, ThE_]:=
Ly.Inverse[(Id - B)].(G.Ph.Transpose[G] + Ps).
    Transpose[Inverse[(Id - B)]]. Transpose[Ly] +
ThE
XYMat[Lx_, Ph_, G_, Id_, B_, Ly_]:=
Lx.Ph.Transpose[G].Transpose[Inverse[(Id - B)]].
    Transpose[Ly]
XMat[Lx_, Ph_, ThD_]:= Lx.Ph.Transpose[Lx] + ThD
MeanVecY[Ay_, Ly_, Id_, B_, AEta_, G_, MXi_]:= Ay +
Ly.Inverse[(Id - B)].(AEta + G.MXi)
MeanVecX[Ax_, Lx_, MXi_]:= Ax + Lx.MXi
```

Once these functions are defined, we need to define the various inputs for the functions (i.e., the parameters for each model). For the Duncan (1975) model we define the following parameter matrices:

```
xLy = IdentityMatrix[3]
xLx = IdentityMatrix[2]
xId = IdentityMatrix[3]
xB = { {0,b12,0}, {b21,0,0}, {b31, b32,0} }
xG = { {g11,0}, {g21,0}, {0, g32} }
xPh = { {ph11, ph21}, {ph21, ph22} }
xPs = { {ps11, ps21,0}, {ps21, ps22,0}, {0,0, ps33}
}
xThE = DiagonalMatrix[{0,0,0}]
xThD = DiagonalMatrix[{0,0}]
```

For the subjective class model we define the following:

```
xLy = IdentityMatrix[3]
xLx = IdentityMatrix[2]
xId = IdentityMatrix[3]
xB = { {0, b12,0}, {b21,0,0}, {b31, b32,0} }
xG = { {g11,0}, {0, g22}, {0,0} }
xPh = { {ph11, ph21}, {ph21, ph22} }
xPs = DiagonalMatrix[{ps11, ps22, ps33}]
xThE = DiagonalMatrix[{th11, th22,0}]
xThD = DiagonalMatrix[{0,0}]
```

For Enders' (2008) auxiliary data model we define the following:

[xiii]For a few models, there are alternative forms for the implied moment matrices. For example, Bollen and Curren (2004) derive a different expression for the implied moment matrices of ALT models.

```
xLy = { {1,0,0,0}, {0,1,0,0}, {0,0,1,0} }
xLx = { {0,1,0,0,0,0,0,1}, {0,0,1,0,0,0,0, l28},
{0,0,0,1,0,0,0, l38},
          {1,0,0,0,0,0,0,0} }
xId = IdentityMatrix[4]
xB = { {0,0,0,1}, {0,0,0, b24}, {0,0,0, b34},
{0,0,0,0} }
xG = { {0,0,0,0,1,0,0,0}, {0,0,0,0,0,1,0,0},
{0,0,0,0,0,0,1,0},
        {0,0,0,0,0,0,0, g48} }
xPh = { {ph11, ph21, ph31, ph41, ph51, ph61, ph71,0},
{ph21, ph22,0,0,0,0,0,0},
          {ph31,0, ph33,0,0,0,0,0},
{ph41,0,0, ph44,0,0,0,0},
         {ph51,0,0,0, ph55,0,0,0},
{ph61,0,0,0,0, ph66,0,0},
         {ph71,0,0,0,0,0, ph77,0},
{0,0,0,0,0,0,0,0, ph88} }
xPs = DiagonalMatrix[{0,0,0, ps44}]
xThE = DiagonalMatrix[{0,0,0}]
xThD = DiagonalMatrix[{0,0,0,0}]
xAy = { {0}, {0}, {0} }
xAx = { {0}, {ax2}, {ax3}, {0} }
xAEta = { {0}, {aEta2}, {aEta3}, {aEta4} }
xMXi = { {mXi1}, {0}, {0}, {0}, {0}, {0}, {0},
{mXi8} }
```

Finally, for the Bollen and Hoyle (1990) model we define the following:

```
xLy = { {1,0}, {l21,0}, {l31,0}, {0,1}, {0, l52},
{0, l62} }
xId = IdentityMatrix[2]
xB = { {0, b21}, {b21,0} }
xG = DiagonalMatrix[{0,0}]
xPh = DiagonalMatrix[{0,0}]
xPs = DiagonalMatrix[{ps11, ps22}]
xThE =
DiagonalMatrix[{th11, th22, th33, th44, th55, th66}]
```

Once the functions for the implied moment matrices are defined and the parameters are entered, one can substitute the parameters into the functions to find the specific implied moment matrices for a given model. The next step of our analysis involves checking local identification using Wald's Rank Rule. To do this, we define a vector of equations based on the nonredundant elements of the implied moment matrices, form the Jacobian, and check the rank of the Jacobian. In defining the vector of equations we make use of the functions we already defined and reference specific elements of the matrices using the row and column locations. The "D" operator in Mathematica takes the derivative of the first expression (in our case, a vector of equations) with respect to the parameters in the second expression (in our case, all of the parameters in the model). The procedure for doing this is

the same for all models. The following commands implement this procedure for the Duncan (1975) model:

```
EqVec =
{YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[1,1]],
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[2,1]],
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[3,1]],
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[2,2]],
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[3,2]],
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[3,3]],
                 XYMat[xLx, xPh, xG, xId, xB, xLy][[1,1]],
                 XYMat[xLx, xPh, xG, xId, xB, xLy][[2,1]],
                 XYMat[xLx, xPh, xG, xId, xB, xLy][[1,2]],
                 XYMat[xLx, xPh, xG, xId, xB, xLy][[2,2]],
                 XYMat[xLx, xPh, xG, xId, xB, xLy][[1,3]],
                 XYMat[xLx, xPh, xG, xId, xB, xLy][[2,3]],
                 XMat[xLx, xPh, xThD][[1,1]],
                 XMat[xLx, xPh, xThD][[2,1]],
                 XMat[xLx, xPh, xThD][[2,2]] }
Jcb = D[EqVec,{
{b12, b21, b31, b32, g11, g21, g32, ph11, ph21,
            ph22, ps11, ps21, ps22, ps33} }]
MatrixRank[Jcb]
```

Assuming the model is locally identified, our final step is to attempt to obtain an algebraic solution. We accomplish this by defining each non-redundant element of the implied moment matrices as an equation and then requesting a solution for all, or a subset, of the equations in terms of all of the parameters. This process is also the same for every model, so we illustrate it with just the subjective class model.

```
eq1 = sig11 = =
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[1,1]]
eq2 = sig21 = =
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[2,1]]
eq3 = sig31 = =
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[3,1]]
eq4 = sig22 = =
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[2,2]]
eq5 = sig32 = =
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[3,2]]
eq6 = sig33 = =
YMat[xLx, xPh, xG, xId, xB, xPs, xThE][[3,3]]
eq7 = sig41 = =
XYMat[xLx, xPh, xG, xId, xB, xLy][[1,1]]
eq8 = sig51 = =
XYMat[xLx, xPh, xG, xId, xB, xLy][[2,1]]
eq9 = sig42 = =
XYMat[xLx, xPh, xG, xId, xB, xLy][[1,2]]
eq10 = sig52 = =
XYMat[xLx, xPh, xG, xId, xB, xLy][[2,2]]
```

```
eq11 = sig43 = =
XYMat[xLx, xPh, xG, xId, xB, xLy][[1,3]]
eq12 = sig53 = =
XYMat[xLx, xPh, xG, xId, xB, xLy][[2,3]]
eq13 = sig44 = = XMat[xLx, xPh, xThD][[1,1]]
eq14 = sig54 = = XMat[xLx, xPh, xThD][[2,1]]
eq15 = sig55 = = XMat[xLx, xPh, xThD][[2,2]]
Solve[
{eq1, eq4, eq6, eq13, eq15, eq14, eq2, eq3, eq7, eq8, eq9, eq10, eq11, eq12},
{b12, b21, b31, b32, g11, g22, ph11, ph21, ph22, ps11, ps 22, ps33, th11,
th22}]
```

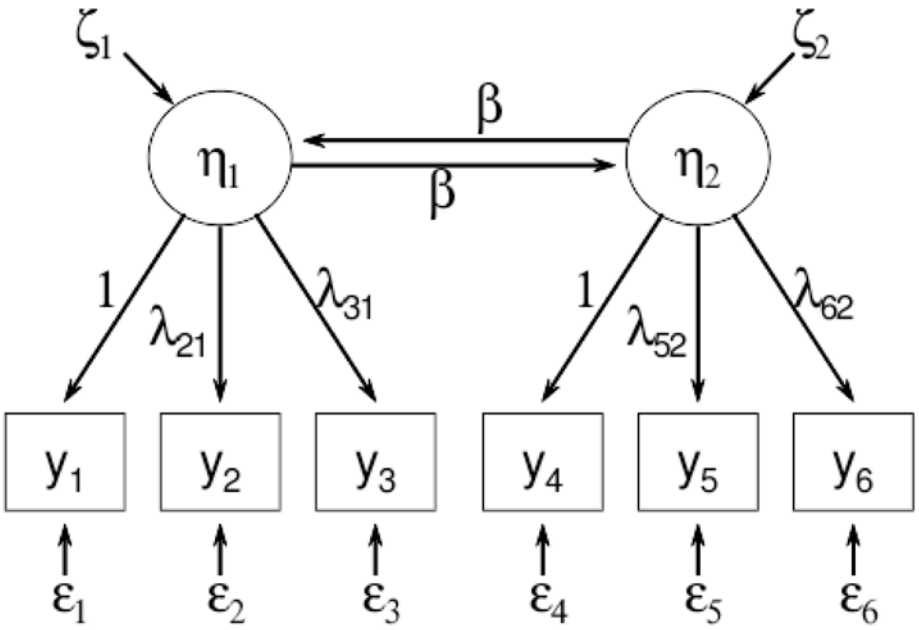


**Figure 1.**
Bollen and Hoyle (1990) Model.

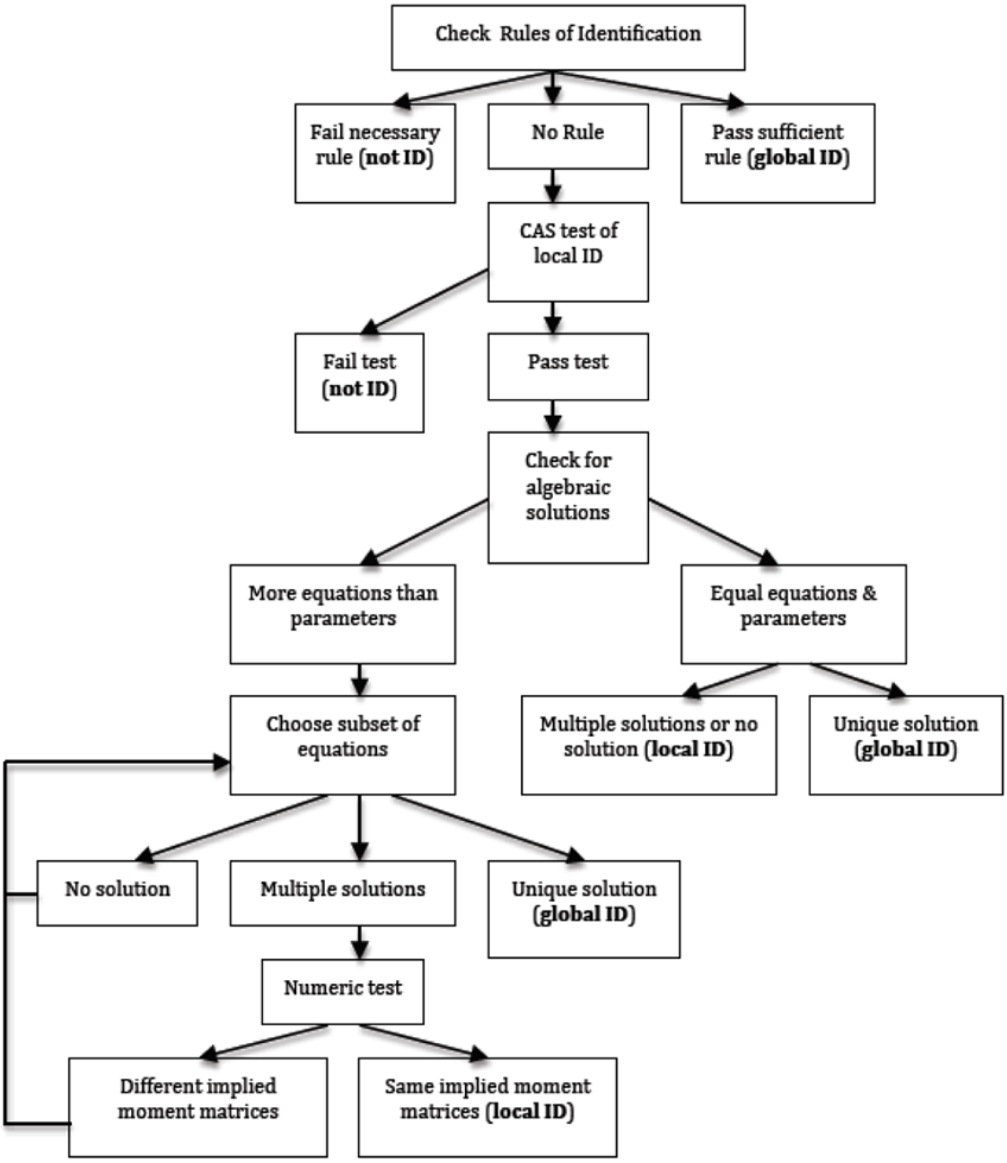


**Figure 2.**

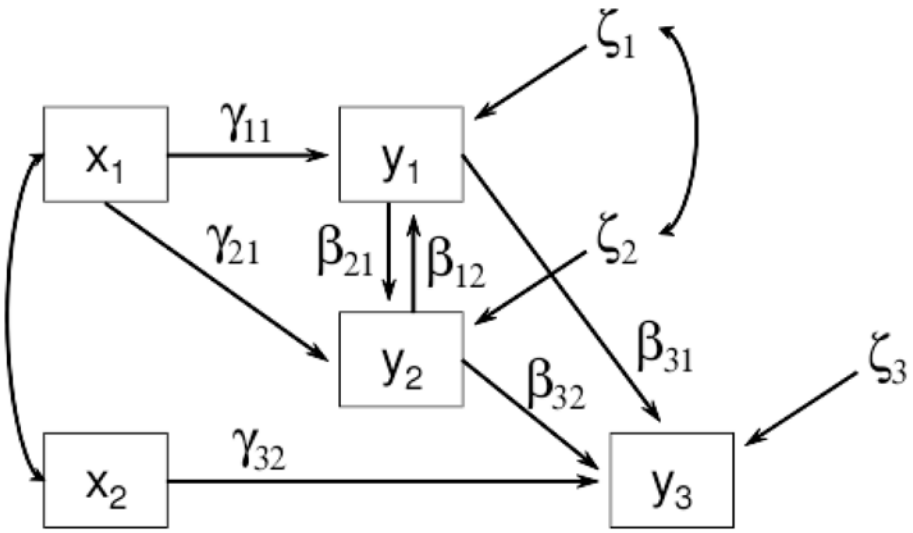


**Figure 3.**
Duncan Model (1975)

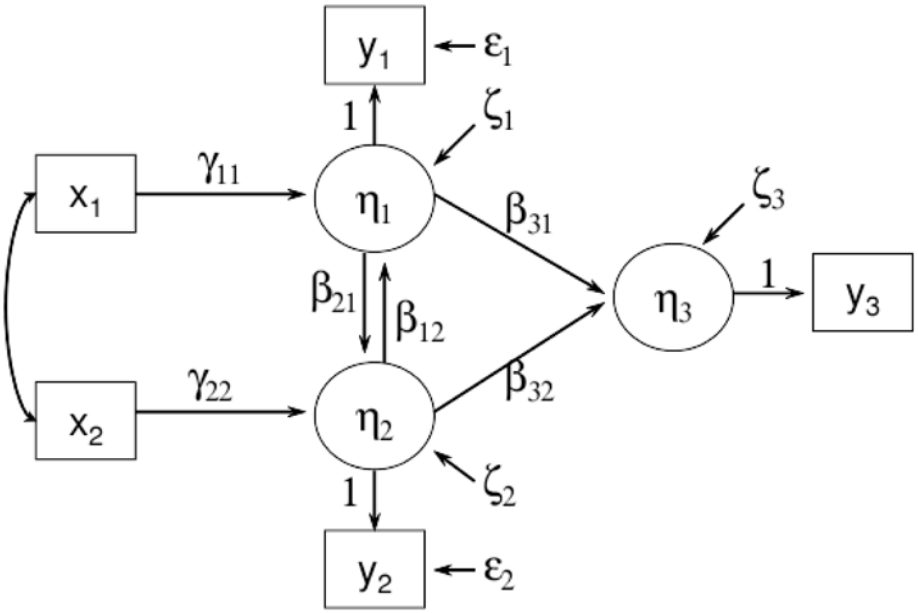


**Figure 4.**
Subjective Class Model

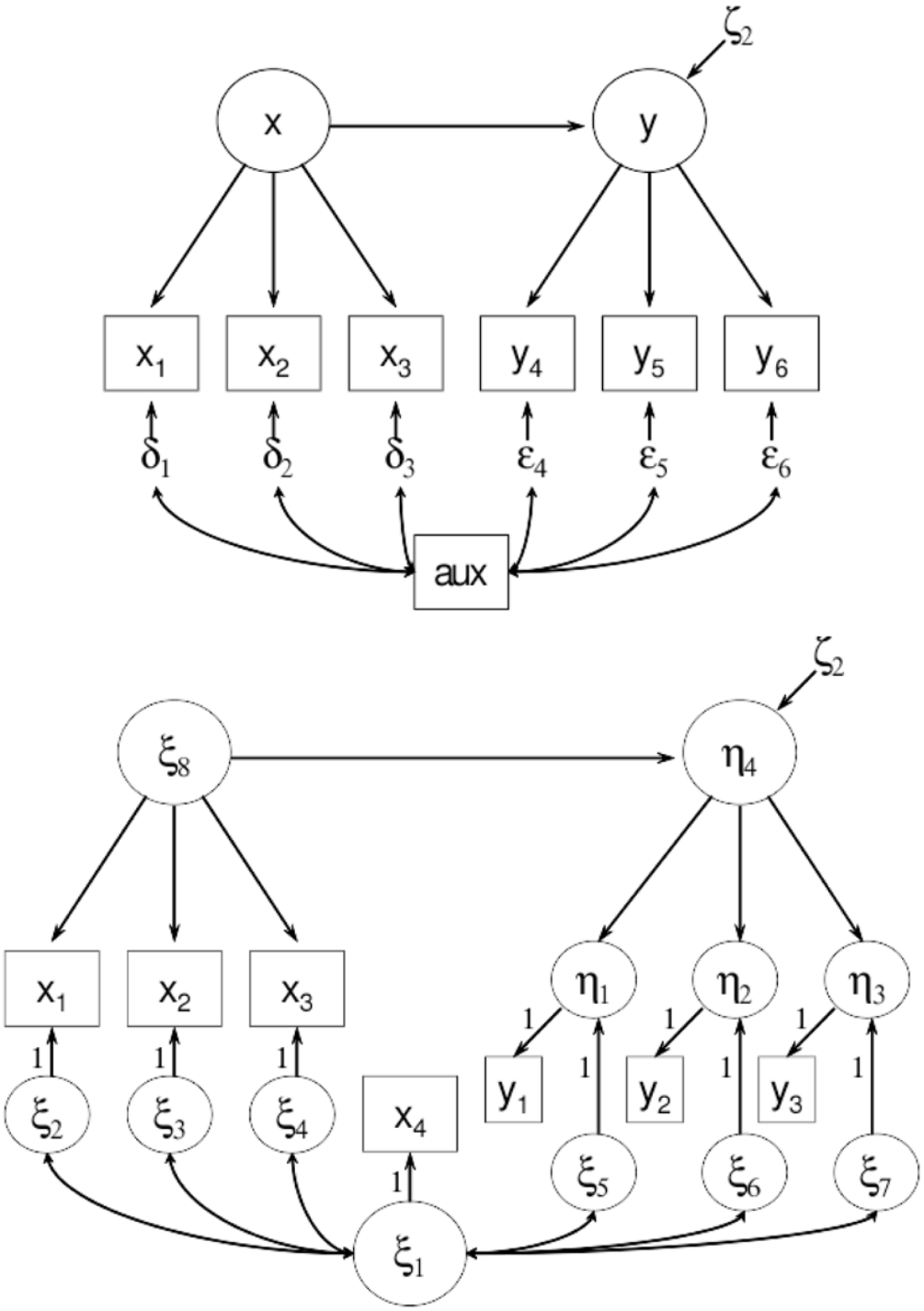


**Figure 5.**
Panel A: Enders (2008) Model
Panel B: Enders (2008) Model

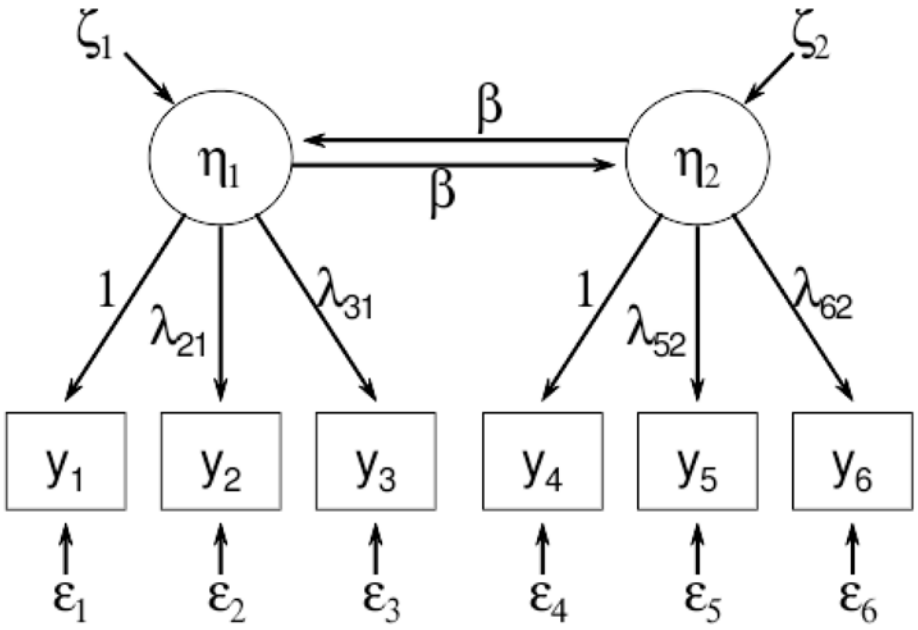


**Figure 6.**
Bollen and Hoyle (1990) Model.