Automating the Selection of Model-Implied Instrumental Variables

KENNETH A. BOLLEN
DANIEL J. BAUER
University of North Carolina at Chapel Hill

Recently, interest has grown in the use of instrumental variables (IVs) in estimating factor analysis and latent variable models such as structural equations models. Bollen (1996) suggested a two-stage least squares (2SLS) technique that makes use of model-implied IVs in estimating the measurement and latent variable models. Model-implied instrumental variables are the observed variables in the model that can serve as instrumental variables in a given equation. One difficulty inhibiting the practical use of the 2SLS estimator is identifying the model-implied IVs. The authors provide a simple procedure that identifies the model-implied IVs and a computer algorithm that can easily be implemented to automate the selection of IVs for simultaneous equations, factor analysis, and latent variable models.

Keywords: *Instrumental variables; structural equation models; two-stage least squares; algorithm*

The maximum likelihood (ML) estimator is the dominant and default estimator in software for structural equation modeling (SEM). Under ideal conditions of no excess multivariate kurtosis and a correctly specified model, there are good reasons for this choice. Under these conditions, the ML estimator is consistent and asymptotically un-biased, efficient, and normally distributed. Furthermore, we can consistently estimate its asymptotic standard errors (Browne 1984; Bollen 1989). However, the ideals of no excess kurtosis or no errors in model specification are more utopian than a description of the

AUTHORS' NOTE: *This work was funded in part by NIDA grant (DA13148) of the first author and a fellowship (DA06062) awarded to the second author. We would like to thank Patrick Curran, John Hipp, and the members of the Carolina Structural Equations Modeling Group for their valuable input throughout this project.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 32, No. 4, May 2004 425-452
DOI: 10.1177/0049124103260341

conditions researchers typically face. It is partly for this reason that researchers are investigating limited-information instrumental variable estimators for SEMs, some of which are asymptotically distribution free and more robust to specification errors (Bollen 2001; Cudeck 1991; Jennrich 1987). Testing for heteroscedasticity and functional form (Pesaran and Taylor 1999), specification errors (Davidson and Mackinnon 1993), and nonnested model testing (Oczkowski 2002) are other reasons that interest has increased in instrumental variable and two-stage least squares (2SLS) estimators. Another application of these estimators is to provide starting values for iterative estimation procedures. Instrumental variable estimators use instrumental variables (IVs), observed variables that are uncorrelated with the disturbances in an equation, to develop a consistent estimator of the parameters in that equation.

Econometrics has a long history of using IVs in simultaneous equations in which the observed variables are treated as if they have no measurement error (e.g., Johnston 1972; Bowden and Turkington 1984). To a lesser degree, econometricians have used IVs to take account of measurement error in a subset of the variables. Usually, these are presented in a single-equation case in which there are single indicators for each substantive variable. Madansky (1964) was perhaps the first to illustrate that researchers could apply IVs to exploratory factor analysis models, a measurement model closer to the applications of psychometricians than the uses of IVs in other disciplines. Hägglund (1982) further developed IV estimators for factor analysis under the assumptions of uncorrelated errors of measurement. Bollen (1989:412, 1996) extended this to confirmatory factor analysis models with correlated errors of measurement and further developed 2SLS to apply to both the latent variable and measurement models in SEM, with or without correlated errors across some equations. The *model-implied IVs* in this approach are observed variables in the model that satisfy the conditions for being IVs. The two most important conditions are (1) the IV must correlate with the endogenous variable that it will replace, and (2) the IV must not correlate with the (composite) disturbance of the estimated equation. Although the selection of the IVs is illustrated in examples (e.g., Bollen 1996; Hägglund 1982), no article has presented a method to automate the selection of IVs.

We emphasize that the model-implied IVs are determined by the specification of the model. An incorrect model can lead to an incorrect selection of IVs. Thus, the selection of model-implied IVs depends on the correctness of the model specification. Our aim is to describe methods by which model-implied IVs are selected, given a specific model structure. We are not providing a method that will lead to a better specified model.

Our primary purpose is to describe an algorithm to automatically select the model-implied IVs for an SEM. We first describe the algorithm, then discuss how it may be implemented in common statistical software. In the next section, we give the basis for the algorithm for the selection of model-implied IVs. After this, we briefly describe a series of programming procedures that can be used to implement the algorithm in statistical software with matrix capabilities (e.g., GAUSS, SAS/IML, SPSS, or STATA). We follow with an illustration of its implementation for an empirical SEM example and a general summary. An SAS/IML macro that implements the algorithm is in the appendix.

IVs SELECTION ALGORITHM

We begin by presenting a modified version of Jöreskog and Sörbom's (1993) LISREL notation for SEMs used in Bollen (2001). The latent variable model is

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

where $\boldsymbol{\eta}$ is a vector of the latent endogenous variables, $\boldsymbol{\alpha}_{\eta}$ is a vector of the intercept terms for the equations, $\mathbf{B}$ is the matrix of coefficients giving the impact of the latent endogenous variables on each other, $\boldsymbol{\xi}$ 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 $\boldsymbol{\zeta}$ is the vector of disturbances. We assume that $E(\boldsymbol{\zeta}) = 0$, $COV(\boldsymbol{\xi}', \boldsymbol{\zeta}) = 0$, and that $(\mathbf{I} - \mathbf{B})$ is nonsingular.

The two equations for the measurement model are

$$\mathbf{y} = \boldsymbol{\alpha}_{y} + \boldsymbol{\Lambda}_{y}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{2}$$

$$\mathbf{x} = \boldsymbol{\alpha}_{x} + \boldsymbol{\Lambda}_{x}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{3}$$

where $\mathbf{y}$ and $\mathbf{x}$ are vectors of the observed indicators of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, respectively; $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_x$ are intercept vectors; $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ are matrices of factor loadings or regression coefficients giving the impact of the latent $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ on $\mathbf{y}$ and $\mathbf{x}$, respectively; and $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are the unique components of $\mathbf{y}$ and $\mathbf{x}$. We assume that the unique components ($\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$) have expected values of zero and are uncorrelated with each other and with $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$.

To apply the 2SLS estimator to equations (1), (2), and (3) requires that each latent variable has a single observed variable to scale it, such that

$$\mathbf{y}_1 = \boldsymbol{\eta} + \boldsymbol{\varepsilon}_1 \tag{4}$$

and

$$\mathbf{x}_1 = \boldsymbol{\xi} + \boldsymbol{\delta}_1, \tag{5}$$

where $\mathbf{y}_1$ and $\mathbf{x}_1$ are the vectors of scaling indicators, and $\mathbf{y}_2$ and $\mathbf{x}_2$ are the vectors of the remaining nonscaling indicators. We can then reexpress equations (4) and (5) as

$$\boldsymbol{\eta} = \mathbf{y}_1 - \boldsymbol{\varepsilon}_1 \tag{6}$$

and

$$\boldsymbol{\xi} = \mathbf{x}_1 - \boldsymbol{\delta}_1, \tag{7}$$

and following Bollen (2001:122-24), we can rewrite the latent variable and measurement models as

$$\mathbf{y}_1 = \boldsymbol{\alpha}_\eta + \mathbf{B}\mathbf{y}_1 + \boldsymbol{\Gamma}\mathbf{x}_1 + \boldsymbol{\varepsilon}_1 - \mathbf{B}\boldsymbol{\varepsilon}_1 - \boldsymbol{\Gamma}\boldsymbol{\delta}_1 + \boldsymbol{\zeta}, \tag{8}$$

$$\mathbf{y}_2 = \boldsymbol{\alpha}_{\mathbf{y}_2} + \boldsymbol{\Lambda}_{\mathbf{y}_2}\mathbf{y}_1 - \boldsymbol{\Lambda}_{\mathbf{y}_2}\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2, \tag{9}$$

$$\mathbf{x}_2 = \boldsymbol{\alpha}_{\mathbf{x}_2} + \boldsymbol{\Lambda}_{\mathbf{x}_2}\mathbf{x}_1 - \boldsymbol{\Lambda}_{\mathbf{x}_2}\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2. \tag{10}$$

These equations are essential to choosing the model-implied IVs. A minimal condition for choosing IVs is that they must have a nonzero correlation with the predictor variables in the equation. This condition is easy to check by using sample estimates of the covariances of the IVs or by regressing the predictor variables on the IVs and checking

for nonzero R^2s. In addition to being nonzero, the R^2s should be nontrivial to avoid the problems that accompany weak IVs (see Bound, Jaeger, and Baker 1995). A more difficult condition to evaluate is that the IVs must be uncorrelated with the disturbance term in the equation in which they will be used.

The only candidates for IVs are observed variables in $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{x}_1$, and $\mathbf{x}_2$, but we cannot use any of these that are correlated with the disturbances for a particular equation. Equations (8), (9), and (10) reveal that the error term for most of these equations will be a composite of several disturbances. For equation (8), the composite includes the equation error (i.e., the $\boldsymbol{\zeta}$ for the equation) and the unique variables (i.e., $\boldsymbol{\epsilon}$ or $\boldsymbol{\delta}$ in the composite disturbance) for any scaling indicators that enter the left- or right-hand side of the equation. The composite disturbances for equations (9) and (10) include unique variables that correspond to the indicator as well as to any scaling indicators that appear on the right-hand side of the equations. Any observed variable that correlates with the disturbances in this composite is *not* eligible to be an IV.[1]

The challenge is to determine whether an observed variable correlates with the composite disturbance, given the structure of the model. One solution derives from an examination of the *total effects* of each disturbance on each observed variable. Although it is rare to discuss the direct, indirect, or total effects of disturbances, each disturbance can be treated in the same way as other latent variables in the model, including determining the effects that it has on the observed variables. The model-implied IVs often change from equation to equation, so it is convenient to consider one equation at a time. First, among the pool of possible IVs, we eliminate any observed variable that is directly or indirectly affected by a disturbance or unique variable that appears in the composite disturbance for that equation. We also eliminate any observed variables that are affected by a disturbance or unique variable that *correlates* with a disturbance or unique variable that is part of the composite disturbance of the equation. Checking for the model-implied correlations between potential IVs and the disturbance terms of an SEM can be a tedious and error-prone process, especially in a large model in which many equations are simultaneously estimated.

Even for some smaller models, it can be difficult to ascertain the list of potential instrumental variables for a given equation. For instance,

consider the simultaneous equation model in Figure 1. Because the errors of the three endogenous variables are all correlated, it is easy to see that the three exogenous predictors—x_1, x_2, and x_3—are the only available instruments for the three equations. However, suppose that $COV(\zeta_1, \zeta_3)$ is zero. In that case, it is not obvious, but y_3 could serve as an instrument in the equation for y_1. If both the $COV(\zeta_1, \zeta_3)$ and the $COV(\zeta_2, \zeta_3)$ are zero, then y_3 is a suitable IV for the y_1 and y_2 equations. Similarly, consider the model of Bollen (1989:12-20, 334-35), reproduced in Figure 2. This model involves one exogenous latent factor ξ_1 with three indicators (x_1 to x_3), two endogenous latent factors η_1 and η_2 with four indicators each (y_1 to y_4 and y_5 to y_8, respectively), a number of correlated errors, and a specific causal model among the latent factors. For this model, it would be quite difficult to identify the instrumental variables for each equation without proceeding quite carefully through the selection steps given above. This suggests the need for an automated selection algorithm that can be implemented in standard statistical software.

A first step in creating an IV selection algorithm is to determine the total effects of each disturbance or unique variable on the observed variables in the model. Fortunately, work is available that provides the total effects of the disturbances and unique variables in SEMs

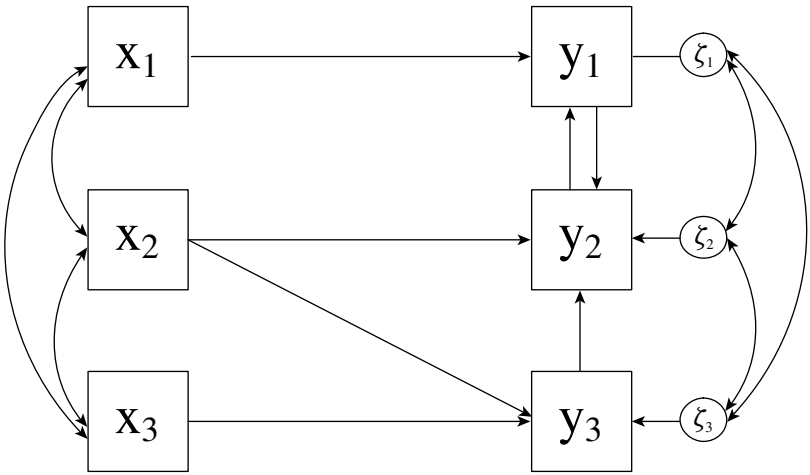

Figure 1: Simultaneous Equation Example

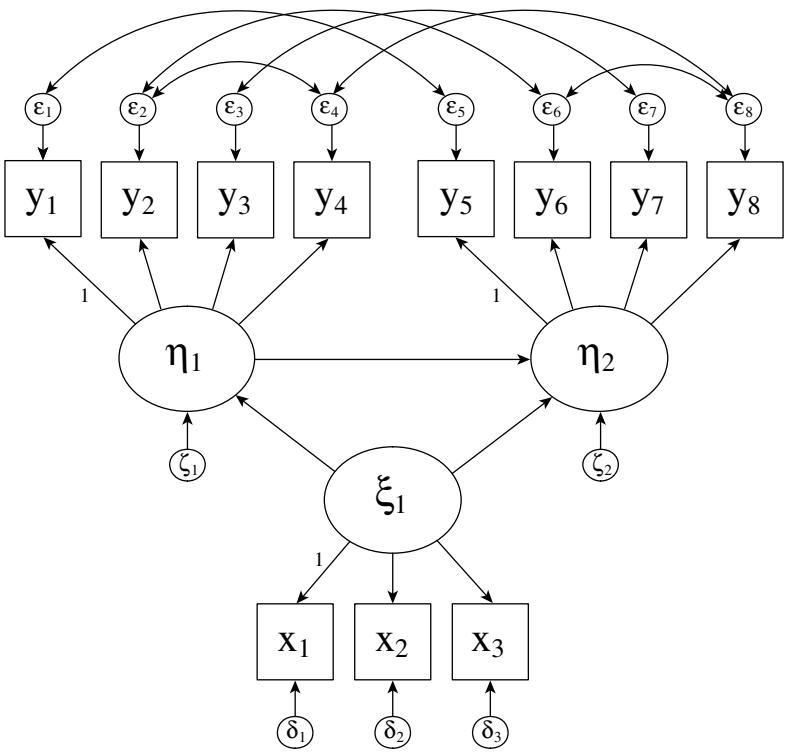

Figure 2: General Structural Equation Model Example

(see Bollen 1987, 1989:376-89). Using these results, the reduced-form equations for the **y** and **x** variables are

$$\mathbf{y} = \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha}_\eta + \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta} + \boldsymbol{\epsilon}, \tag{11}$$

$$\mathbf{x} = \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{12}$$

and the total effects of the disturbances on the observed variables are equal to the coefficient matrices for the disturbances or unique variables in these equations.[2] So the total effect of $\boldsymbol{\zeta}$ on **y** is $\boldsymbol{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}$, the total effect of $\boldsymbol{\epsilon}$ on **y** is **I**, and the total effect of $\boldsymbol{\delta}$ on **x** is **I**. Implicit in these equations are total effects of **0** for $\boldsymbol{\epsilon}$ on **x**, $\boldsymbol{\zeta}$ on **x**, and $\boldsymbol{\delta}$ on **y**.

With these total effects, we can outline the steps for finding the model-implied IVs. For each equation from (8), (9), or (10), we must

1. identify which disturbances or unique components are in the equation;
2. calculate the total effects of $\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$ on each observed variable in the model;
3. eliminate as IVs any observed variables that have a nonzero total effect originating from the disturbances or unique components from (1) above;
4. of the remaining observed variables, eliminate any observed variable affected by a disturbance or unique component that correlates with a disturbance or unique component identified from (1).

By necessity, the remaining observed variables are the model-implied IVs. We now describe in more detail how to program this algorithm for selecting IVs.

PROGRAM FOR SELECTING IVs

The program proceeds in several steps, each of which can be implemented in any programming language capable of manipulating matrices, such as GAUSS, SAS/IML, SPSS, or STATA. We describe these steps in a general way here so that they can be easily implemented in any of these languages. The appendix provides an SAS/IML macro that implements this algorithm. This program is written in a modular form so that each module corresponds to a specific programming task or step of the algorithm.

Before implementing the algorithm outlined above, it is necessary to define the variables in the model and the model structure. First, each observed variable must be assigned a unique index number and placed into the appropriate vector (i.e., $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{x}_1$, or $\mathbf{x}_2$). The scaling variables for the latent factors must be designated at this point. An indicator must be selected that loads only on the latent variable, and among such indicators, it should be the one hypothesized to be most closely related to the latent variable. The model structure may be defined in several ways. One option would be to input the predictors for each equation directly. An alternative is to define the pattern of fixed and free elements in the regression coefficient matrices $\boldsymbol{\Lambda}_{\mathbf{y}_2}$, $\boldsymbol{\Lambda}_{\mathbf{x}_2}$, $\mathbf{B}$, and $\boldsymbol{\Gamma}$. Because the latter approach is more familiar to most

SEM users, it is the one that we pursue here. The elements in the regression coefficient matrices are dummy coded, with a 0 indicating that the parameter is not estimated or is fixed at 0, or a 1 indicating that the parameter is estimated or fixed to a nonzero value.[3] Note that the intercept vectors $\boldsymbol{\alpha}_\eta$, $\boldsymbol{\alpha}_y$ and $\boldsymbol{\alpha}_x$ do not affect the selection of IVs and so are not required. The last piece of information needed to run the algorithm are the nonzero elements of $\boldsymbol{\Theta}_\varepsilon$, $\boldsymbol{\Theta}_\delta$, and $\boldsymbol{\Psi}$ where these are the covariance matrices for $\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$, and $\boldsymbol{\zeta}$ respectively. As before, elements that are fixed at 0 are designated with a 0. However, unlike the regression coefficients, each nonzero parameter in the disturbance/uniqueness matrices must be assigned a unique index number (note that because these matrices are symmetric, these numbers will be duplicated for off-diagonal elements).

With this information in hand, we can proceed to Step 1 of the algorithm outlined above, identifying the composite disturbance of each equation. To do so, we must determine the model-implied predictors in each equation, which in turn involves scanning the regression coefficient matrices for nonzero elements. As can be seen in equations (8), (9), and (10), the relevant regression coefficient matrices vary depending on whether the dependent variable in the equation is in $\mathbf{y}_1$, $\mathbf{y}_2$, or $\mathbf{x}_2$. For instance, to identify the predictors for the i element of $\mathbf{y}_1$, or y_i, we first scan the row of the $\mathbf{B}$ matrix associated with y_i for nonzero values. Whenever a nonzero value is found, we use its column position j to identify the particular predictor in $\mathbf{y}_1$, or y_j, associated with that coefficient. Then the row of $\boldsymbol{\Gamma}$ associated with y_i is scanned for nonzero elements, using their column positions j to identify the particular predictor in $\mathbf{x}_1$, or x_j, associated with that coefficient. The same approach is used to identify the model-implied predictors for variables in $\mathbf{y}_2$ and $\mathbf{x}_2$, except that the relevant regression coefficient matrices are $\boldsymbol{\Lambda}_{\mathbf{y}_2}$ and $\boldsymbol{\Lambda}_{\mathbf{x}_2}$, respectively. It is convenient to stack the predictor arrays for each equation into a matrix $\mathbf{P}$ and to identify each row by setting the first column position to the index number of the dependent variable of the equation. Predictor arrays will vary in length, so some must be "padded" with 0s for concatenation into a single matrix. Each of these operations is demonstrated in the "Predictors" module of the program in the appendix.

We now have sufficient information to determine the composite disturbance for each equation in the model. Recall that the equations for the dependent variables (e.g., equations (8), (9), and (10)) each

include uniqueness terms associated with the observed dependent variables, as well as the uniquenesses and disturbances associated with their predictors. Similarly, each row of **P** contains the number of the dependent variable in the first column and the numbers of the predictors in the remaining columns. Thus, the variable numbers in each row of **P** can be used to identify the disturbances of the equation. If the variable number is in $\mathbf{y}_1$ or $\mathbf{y}_2$, then the index of the corresponding uniqueness parameter from $\boldsymbol{\Theta}_\varepsilon$ is added to the composite disturbance array, designated $\mathbf{C}_i$ for equation i. On the other hand, if the variable is in $\mathbf{x}_2$, the index of the corresponding uniqueness parameter in $\boldsymbol{\Theta}_\delta$ is added to $\mathbf{C}_i$. For each dependent variable from $\mathbf{y}_1$, it is also necessary to locate the disturbance of the corresponding $\boldsymbol{\eta}$ and add the index number of the ζ to $\mathbf{C}_i$. For subsequent manipulation, each $\mathbf{C}_i$ is stacked in a matrix **C**, with each row identified by the dependent variable index, as was done for the **P** matrix. This step of the algorithm is illustrated in the "Composite" module of the program in the appendix.

Step 2 of the algorithm is to determine the total effects of the disturbances and uniquenesses on each variable in the model. This step involves the construction of several "effects" matrices, **T**, which contain the index numbers of the disturbances or uniquesnesses that affect the variables in the model. Begin by considering the effects of the disturbances/uniquenesses on the **y** variables. The total effects of $\boldsymbol{\epsilon}$ on **y**, or $\mathbf{T}_{\mathbf{y}\boldsymbol{\varepsilon}}$, are composed entirely of direct effects. Specifically, each ε has a direct effect on the corresponding y variable. The unique index numbers for the $\boldsymbol{\varepsilon}$ are contained on the diagonal of $\boldsymbol{\Theta}_\varepsilon$, and thus $\mathbf{T}_{\mathbf{y}\boldsymbol{\varepsilon}}$ may be constructed by simply extracting these elements from $\boldsymbol{\Theta}_\varepsilon$ and placing them into a column vector. The effects of $\boldsymbol{\zeta}$ on $\mathbf{y}_1$, or $\mathbf{T}_{\mathbf{y}_1\boldsymbol{\zeta}}$, are calculated as $(\mathbf{I} - \mathbf{B})^{-1}$, and the effects of $\boldsymbol{\zeta}$ on $\mathbf{y}_2$, or $\mathbf{T}_{\mathbf{y}_2\boldsymbol{\zeta}}$, are calculated as $\boldsymbol{\Lambda}_{\mathbf{y}_2}(\mathbf{I} - \mathbf{B})^{-1}$. Nonzero elements in $\mathbf{T}_{\mathbf{y}_1\boldsymbol{\zeta}}$ and $\mathbf{T}_{\mathbf{y}_2\boldsymbol{\zeta}}$ must then be replaced by the corresponding parameter numbers from $\boldsymbol{\Psi}$. The index vectors for $\mathbf{y}_1$ and $\mathbf{y}_2$ are then appended to $\mathbf{T}_{\mathbf{y}_1\boldsymbol{\zeta}}$ and $\mathbf{T}_{\mathbf{y}_2\boldsymbol{\zeta}}$, respectively, so that the first column of the matrices identifies the variable affected by each disturbace/uniqueness. The resulting matrices are then concatenated together and sorted by the variable index to create $\mathbf{T}_{\mathbf{y}\boldsymbol{\zeta}}$. The sorting is necessary to place this matrix into the same index order as $\mathbf{T}_{\mathbf{y}\boldsymbol{\varepsilon}}$. $\mathbf{T}_{\mathbf{y}\boldsymbol{\zeta}}$ and $\mathbf{T}_{\mathbf{y}\boldsymbol{\varepsilon}}$ are then concatenated together to form a single effects matrix for the disturbances/uniquenesses on the **y** variables, designated $\mathbf{T}_{\mathbf{y}\boldsymbol{\zeta}\boldsymbol{\varepsilon}}$.

Calculating the total effects of the disturbances/uniquenesses on the **x** variables is much simpler. The only effects on **x** are direct effects from $\boldsymbol{\delta}$, so $\mathbf{T}_{\mathbf{x}\boldsymbol{\delta}}$ can be formed like $\mathbf{T}_{\mathbf{y}\boldsymbol{\varepsilon}}$ by simply extracting the diagonal elements from $\boldsymbol{\Theta}_{\delta}$ and placing them into a column vector. The index vectors for $\mathbf{x}_1$ and $\mathbf{x}_2$ are then combined into a single vector, sorted in index order, and appended to $\mathbf{T}_{\mathbf{x}\boldsymbol{\delta}}$ to form the index column for this matrix. Finally, $\mathbf{T}_{\mathbf{y}\boldsymbol{\zeta}\boldsymbol{\varepsilon}}$ and $\mathbf{T}_{\mathbf{x}\boldsymbol{\delta}}$ are combined into a single matrix **T**, and the rows are sorted by the variable index. Because $\mathbf{T}_{\mathbf{y}\boldsymbol{\zeta}\boldsymbol{\varepsilon}}$ and $\mathbf{T}_{\mathbf{x}\boldsymbol{\delta}}$ are likely to have different numbers of columns, it may be necessary to "pad" one matrix with columns of 0s to make it conformable for concatenation. The "Effects" module of the program in the appendix illustrates this step of the algorithm.

In Step 3 of the algorithm, an initial set of potential instruments for each equation is defined by selecting variables that are unaffected by the disturbances or uniquenesses in that equation. This step is accomplished by searching the total effects matrix **T** for the disturbance/uniqueness numbers that appear in each row of the composite disturbance matrix **C**. Selecting row i from **C**, we examine each column from $j = 2$ to J, where J is the total number of columns in **C** (recall that column 1 contains the dependent variable index for the equation). Each nonzero number C_{ij} corresponds to a disturbance/uniqueness parameter in $\boldsymbol{\Theta}_{\varepsilon}$, $\boldsymbol{\Theta}_{\delta}$, or $\boldsymbol{\Psi}$ that appears in the disturbance array for equation i. These index numbers are compared to those that appear in each row p from **T** in columns $q = 2$ to Q, where Q is the total number of columns in **T** (recall that column 1 contains the index number of the variable affected by the disturbances). Each nonzero number T_{pq} corresponds to a disturbance/uniqueness parameter that influences variable p. Of key importance, if at any point $C_{ij} = T_{pq}$, then variable p is ineligible as an instrument for equation i. If, however, $C_{ij} \neq T_{pq}$ for all j and q, then the variable index T_{p1} is added to the "potential instrumental variable" array, designated $\mathbf{PIV}_i$ for equation i. Each $\mathbf{PIV}_i$ is indexed by the number of the dependent variable C_{i1}, and the arrays are stacked into a single matrix **PIV**. To do so, it may be necessary to "pad" some of the $\mathbf{PIV}_i$ arrays with extra 0s to equalize their lengths. This step of the algorithm is illustrated in the "Potential_IV" module of the program in the appendix.

The final step of the algorithm, Step 4, finalizes the list of eligible instrumental variables. In this step, variables designated as

potential instruments for a given equation in Step 3 are reexamined to evaluate whether they are correlated with any of the parameters in the composite disturbance of that equation. We begin by setting the instrumental variable matrix **IV** equal to the **PIV** matrix defined in Step 3. Each row i from **PIV** is then selected and examined from column $j = 2$ to J, where J is the total number of columns in **PIV**. Each element PIV_{ij} references a potential instrumental variable for equation i. If PIV_{ij} is in $\mathbf{y}_1$ or $\mathbf{y}_2$, $\boldsymbol{\Theta}_\varepsilon$ is scanned to determine whether the ε affecting PIV_{ij} is correlated with any $\boldsymbol{\varepsilon}$ in the composite disturbance for equation i. Alternatively, if PIV_{ij} is in $\mathbf{x}_1$ or $\mathbf{x}_2$, $\boldsymbol{\Theta}_\delta$ is scanned to determine whether the δ affecting PIV_{ij} is correlated with any $\boldsymbol{\delta}$ in the composite disturbance for equation i. Finally, if PIV_{ij} is in $\mathbf{y}_1$, it is also necessary to scan $\boldsymbol{\Psi}$ to determine whether the ζ associated with PIV_{ij} is correlated with any $\boldsymbol{\zeta}$ in the composite disturbance for equation i. If any of these disturbances/uniquenesses are correlated, then the variable index in IV_{ij} is replaced by a 0. The remaining nonzero elements in **IV** are the model-implied instrumental variables for the equations in the model.

EXAMPLE

In this section, we illustrate the IV selection algorithm with an empirical example. The example considers the relationship between industrialization and political democracy among developing countries from 1960 to 1965. Further description of the model is available in Bollen (1989:12-20, 334-35). Figure 2 provides a path diagram of the model. We begin by relabeling the variables and the disturbance/uniqueness parameters with numbers, as shown in Figure 3. In this new diagram, each variable has been assigned a unique variable number, and each nonzero element in the disturbance/uniqueness matrices has been assigned a unique parameter number. For ease of reference, the uniqueness of each observed variable has been assigned the same number as that observed variable. The scaling variables for η_1, η_2, and ξ_1 are variables 1, 5, and 9, respectively. Variables 2 to 4 and 6 to 8 constitute $\mathbf{y}_2$, and variables 10 to 11 constitute $\mathbf{x}_2$. The "Main" module of the example SAS program in the appendix demonstrates how this model would be input by the user.

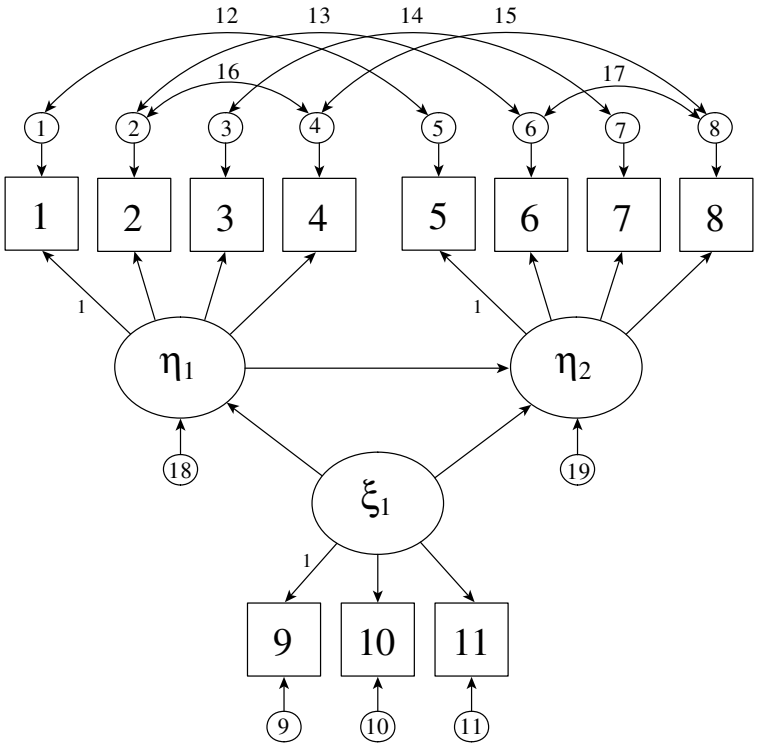

Figure 3: Industrialization and Democracy Example to Illustrate Algorithm

Step 1 of the algorithm defines the composite disturbance of each equation and arrays them into the matrix

$$\mathbf{C} = \left[\begin{array}{c|cccc} 1 & \mathbf{18} & \mathbf{1} & \mathbf{9} & 0 \\ 5 & \mathbf{19} & \mathbf{5} & \mathbf{1} & \mathbf{9} \\ 2 & \mathbf{2} & \mathbf{1} & 0 & 0 \\ 3 & \mathbf{3} & \mathbf{1} & 0 & 0 \\ 4 & \mathbf{4} & \mathbf{1} & 0 & 0 \\ 6 & \mathbf{6} & \mathbf{5} & 0 & 0 \\ 7 & \mathbf{7} & \mathbf{5} & 0 & 0 \\ 8 & \mathbf{8} & \mathbf{5} & 0 & 0 \\ 10 & \mathbf{10} & \mathbf{9} & 0 & 0 \\ 11 & \mathbf{11} & \mathbf{9} & 0 & 0 \end{array}\right]. \tag{13}$$

The rows of matrix **C** correspond to the dependent variables in equations (8), (9), and (10). The first column is partitioned from the rest of the row to indicate that this column contains the indices of the dependent variables of the model equations, while the remaining columns contain the indices of the disturbances/uniquenesses for those equations. The numbers contained in the matrix correspond to the variables and disturbance/uniqueness labels given in Figure 3. The disturbance/uniqueness parameter numbers are printed in bold so that they will stand out from the zeros, which are merely used as place holders in the matrix. Variables 1 and 5 are the scaling indicators, and thus rows 1 and 2 of **C** give the composite disturbances for the latent variable model in equation (8). For instance, row 1 of this matrix indicates that the equation for η_1, scaled by variable 1, contains parameters 18, 1, and 9, which correspond to the disturbance of η_1 and the uniquenesses of variables 1 and 9. Similarly, the rows indexed by variables 2 to 4 and 6 to 8 give the composite disturbances for the measurement model in equation (9). Variables 10 and 11 are from $\mathbf{x}_2$, so the last two rows identify the composite disturbances for the measurement model in equation (10). There is no row corresponding to variable 9, the scaling indicator for ξ, because variable 9 is contained in $\mathbf{x}_1$ and so is not a dependent variable in equation (8), (9), or (10).

In Step 2, the total effects of the disturbances/uniqueness on the observed variables are calculated to be

$$\mathbf{T} = \left[\begin{array}{c|ccc} 1 & \mathbf{1} & \mathbf{18} & 0 \\ 2 & \mathbf{2} & \mathbf{18} & 0 \\ 3 & \mathbf{3} & \mathbf{18} & 0 \\ 4 & \mathbf{4} & \mathbf{18} & 0 \\ 5 & \mathbf{5} & \mathbf{18} & \mathbf{19} \\ 6 & \mathbf{6} & \mathbf{18} & \mathbf{19} \\ 7 & \mathbf{7} & \mathbf{18} & \mathbf{19} \\ 8 & \mathbf{8} & \mathbf{18} & \mathbf{19} \\ 9 & \mathbf{9} & 0 & 0 \\ 10 & \mathbf{10} & 0 & 0 \\ 11 & \mathbf{11} & 0 & 0 \end{array}\right]. \tag{14}$$

This matrix is also partitioned, with the first column indicating the variable index and the remaining columns indicating the parameter indices of disturbances that have a nonzero total effect on the variable.

The disturbance parameters are printed in bold to stand out from the place-holding 0 elements. Note that there are several submatrices present in **T**. For instance, the first eight rows of column 2 represent $\mathbf{T}_{y\varepsilon}$, while the last three rows are $\mathbf{T}_{x\delta}$. In addition, $\mathbf{T}_{y\zeta}$ is contained in the first eight rows of columns 2 and 3.

Step 3 of the algorithm compares matrices **T** and **C** to provide us with an initial set of potential instruments for each equation, given as

$$\mathbf{PIV} = \left[\begin{array}{cc|cccccccc} 1 & | & \mathbf{10} & \mathbf{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & | & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{10} & \mathbf{11} & 0 & 0 & 0 & 0 \\ 2 & | & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 3 & | & \mathbf{2} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 4 & | & \mathbf{2} & \mathbf{3} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 6 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 7 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{6} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 8 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{6} & \mathbf{7} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 10 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{11} \\ 11 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{10} \end{array}\right]. \tag{15}$$

This matrix is partitioned such that the first column indicates the number of the dependent variable in the equation, and the remaining columns index the observed variables that are not affected by any of the disturbances for that equation. The variable indices are printed in bold to differentiate them from the place-holding 0s in the matrix. Not all of these variables are eligible as instruments, however, as some are correlated with the disturbances for the equation.

In Step 4, the indices of ineligible variables in **PIV** are replaced with 0s, leaving only the variable indices of the model-implied IVs, given as

$$\mathbf{IV} = \left[\begin{array}{cc|cccccccc} 1 & | & \mathbf{10} & \mathbf{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & | & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{10} & \mathbf{11} & 0 & 0 & 0 & 0 \\ 2 & | & \mathbf{3} & 0 & 0 & 0 & \mathbf{7} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 3 & | & \mathbf{2} & \mathbf{4} & 0 & \mathbf{6} & 0 & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 4 & | & 0 & \mathbf{3} & 0 & \mathbf{6} & \mathbf{7} & 0 & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 6 & | & 0 & 0 & \mathbf{3} & \mathbf{4} & \mathbf{7} & 0 & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 7 & | & 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{6} & \mathbf{8} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 8 & | & 0 & \mathbf{2} & \mathbf{3} & 0 & 0 & \mathbf{7} & \mathbf{9} & \mathbf{10} & \mathbf{11} \\ 10 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{11} \\ 11 & | & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7} & \mathbf{8} & \mathbf{10} \end{array}\right]. \tag{16}$$

This matrix is partitioned in the same way as **PIV**. The variables that meet all of the criteria to be used as instruments for the different equations are printed in bold. For instance, row 1 of this matrix indicates that the equation for η_1, scaled by variable 1, has two eligible instruments, variables 10 and 11. Row 2 indicates that the equation for η_2, scaled by variable 5, has five eligible instruments, variables 2, 3, 4, 10, and 11. Rows 3 through 8 indicate the instruments for the measurement equations of $\mathbf{y}_2$, and the last two rows indicate the instruments for the measurement equations of $\mathbf{x}_2$.

As an example of how these matrices are constructed and used to specify the model-implied IVs of each equation, we trace through the selection algorithm for a single variable in the model, variable 2 of Figure 3. As indicated by row 2 of the **T** matrix, the model implies that variable 2 is affected both by its unique residual, indexed by the number 2, and by the disturbance of η_1, indexed by the number 18. In the **C** matrix, these numbers appear in the composite disturbances of the equations for variables 1 and 2. As such, the **PIV** matrix lists only variables 3 through 11 as potential instruments for the equation. However, we must also rule out variables that are affected by a disturbance or uniqueness that *correlates* with a disturbance or uniqueness in the composite disturbance for the equation. Row 3 of the **C** matrix indicates that the composite disturbance for the equation includes uniquenesses 1 and 2. Scanning through $\mathbf{\Theta}_\varepsilon$ (or visual examination of the model in Figure 3) reveals that uniqueness 1 is correlated with uniqueness 5, ruling out variable 5 as an instrument. Furthermore, uniqueness 2 is correlated with 4 and 6, ruling out variables 4 and 6 as instruments. Thus, in the **IV** matrix, variables 4 to 6 are "zeroed" out as eligible instruments for the equation for variable 2.

SUMMARY

One obstacle to the broader implementation of the 2SLS estimator in SEM software is that the procedure requires selection of IVs for each equation in the model. This article outlines a general algorithm for selecting the IVs for each equation in an SEM. Our appendix provides an IV selection program in Proc IML in SAS to illustrate its implementation. The same algorithm could be implemented in many matrix

programming languages and should be adaptable to existing SEM packages. We also should point out that although the algorithm works for models with latent variables, it also applies to standard simultaneous equation models in which latent variables are not present such as is typical in econometrics. In the simultaneous equation situation typical of econometrics, in which all disturbances are correlated and no measurement error is allowed, then this algorithm is not needed since only the exogenous variables in the system are IVs. But in more complicated models with some, but not all, disturbances uncorrelated or when multiple indicators and measurement error are present, then this algorithm should prove helpful.

One limitation of our work is that we have not discussed the selection of IVs in models with interactions of latent variables. Bollen (1995) and Bollen and Paxton (1998) give rules for selecting IVs in these models. Another limitation is that we do not discuss the use of 2SLS in models in which all possible scaling indicators load on two or more factors such as occurs in multitrait multimethod models. However, the algorithm that we provide here will cover the majority of applications that occur in practice and should make it easier to use the 2SLS estimator with SEMs.

It should be remembered that the algorithm derives *model-implied* IVs. If the model is misspecified, some of the IVs chosen by using our IV selection strategy could be in error. In the case of full-information estimators such as maximum likelihood, this misspecification could have systemwide ramifications. For 2SLS, the effects of the misspecification are restricted to the misspecified equations and those equations with the incorrect IVs. We recommend that researchers use tests for the suitability of the IVs in overidentified equations. For example, Basmann (1957, 1960) has an overidentification test that appears to perform well and is available in Proc Syslin in SAS. Other tests are available as well (e.g., Anderson 1951; Anderson and Rubin 1949; Sargan 1958). The null hypothesis of these tests is that all IVs are uncorrelated with the composite disturbance of the equation in which they are being used. A significant test statistic suggests that at least one IV is not suitable.

Another point to emphasize is that the algorithm chooses all model-implied IVs for an equation. There are simulation and analytical results that suggest there are sometimes advantages to using a subset

of the possible IVs. For example, Bound et al. (1995) show that using variables only weakly related to the endogenous explanatory variables as instrumental variables can lead to inconsistent estimates, even if there is only a weak relationship between the instrument and the error term in the structural equation. In addition, analytical work by Mariano (1977) illustrates that the absolute bias of the estimator is "an increasing function of the degree of overidentification" (see also Magdalinos 1985). We thus recommend that the algorithm be used to identify all possible IVs, from which a smaller set may be used in the actual estimation process.

APPENDIX

```
/*************************************************************************/
/*                                                                       */
/* Program:           Automated IV Selection                             */
/* By:                Daniel Bauer and Kenneth Bollen                    */
/* Last Revised:      09/04/2003                                         */
/*                                                                       */
/*************************************************************************/


options nocenter;
title 'Automated IV Selection';


PROC IML;

/*************************************************************************/
/* MODULE: FINAL_IV                                                      */
/*                                                                       */
/* Create IV = Matrix with one row for every equation in the model. Column    */
/*       position 1 contains the dependent variable index. Remaining          */
/*       columns contain indices of  variables that are eligible              */
/*       instruments for that equation.                                       */
/*                                                                       */
/*************************************************************************/


START FINAL_IV;

* MAKE FINAL SELECTION OF INSTRUMENTAL VARIABLES FOR EACH EQUATION;

* Check to see if potential instruments are affected by
  any disturbances correlating with any disturbance in the
  composite disturbance of equation;

IV = PIV;
DO i = 1 to NROW(PIV);
  DO p = 2 to NCOL(PIV);
    IF PIV[i,p] > 0 THEN
    DO j = 2 to NCOL(Comp);
```

```
      DO q = 2 to NCOL(Total);
        IF (Comp[i,j] > 0) & (Total[PIV[i,p],q] > 0) THEN DO;
          * check if element from composite and element from
            total in same matrix;
          * if so, check if covariance between disturbances = 0;
          IF ANY(ThetaE=Comp[i,j]) & ANY(ThetaE=Total[PIV[i,p],q]) THEN DO;
            DO r = 1 to NROW(ThetaE);
              DO c = 1 to NCOL(ThetaE);
                IF Comp[i,j] = ThetaE[r,c] THEN loc1 = r;
                IF Total[PIV[i,p],q] = ThetaE[r,c] THEN loc2 = c;
              END;
            END;
            IF ThetaE[loc1,loc2] > 0 THEN DO;
               IV[i,p] = 0;
            END;
          END;
          IF ANY(ThetaD=Comp[i,j]) & ANY(ThetaD=Total[PIV[i,p],q]) THEN DO;
            DO r = 1 to NROW(ThetaD);
              DO c = 1 to NCOL(ThetaD);
                IF Comp[i,j] = ThetaD[r,c] THEN loc1 = r;
                IF Total[PIV[i,p],q] = ThetaD[r,c] THEN loc2 = c;
              END;
            END;
            IF ThetaD[loc1,loc2] > 0 THEN IV[i,p] = 0;
          END;
          IF ANY(Psi=Comp[i,j]) & ANY(Psi=Total[PIV[i,p],q]) THEN DO;
            DO r = 1 to NROW(Psi);
              DO c = 1 to NCOL(Psi);
                IF Comp[i,j] = Psi[r,c] THEN loc1 = r;
                IF Total[PIV[i,p],q] = Psi[r,c] THEN loc2 = c;
              END;
            END;
            IF Psi[loc1,loc2] > 0 THEN IV[i,p] = 0;
          END;
        END;
      END;
    END;
  END;
END;
PRINT IV;


* Print IV matrix;
PRINT '****** MODEL-IMPLIED INSTRUMENTAL VARIABLES ******';
PRINT 'DV = Dependent Variable; IVs = Instrumental Variables';
PRINT '(Numbers Correspond to User-Defined Indices for Variables)';
IV2 = IV;                  *Sorting by DV index;
IV[rank(IV[,1]),]=IV2;
DO i = 1 to NROW(IV);
  DV = IV[i,1];
  IVs = {0};
  DO j = 2 to NCOL(IV);
    IF IV[i,j] > 0 THEN
      IVs = IVs||IV[i,j];
  END;
  IF NCOL(IVs) > 1 THEN IVs = IVs[1,2:NCOL(IVs)];
  PRINT DV IVs;
END;
```

```
FINISH;

/*****************************************************************************/
/* MODULE: POTENTIAL_IV                                                      */
/*                                                                           */
/* Create PIV = Matrix of potential instrumental variables. Column position 1  */
/*        contains the variable index of the dependent variable from         */
/*        the equation. The remaining column positions contain the indices   */
/*        of variables that are not affected by disturbances in composite    */
/*        disturbance of the equation.                                       */
/*                                                                           */
/***************************************************************************** */

START POTENTIAL_IV;

* CREATE MATRIX PIV OF POTENTIAL INSTRUMENTAL
  VARIABLES FOR EACH EQUATION;
DO i = 1 to NROW(COMP);
  PIVi = COMP[i,1];
  DO p = 1 to NROW(Total);
    inst = 1;
    DO j = 2 to NCOL(COMP);
      IF (COMP[i,j] > 0) & (Inst = 1) THEN
        DO q = 2 to NCOL(Total);
          IF COMP[i,j] = TOTAL[p,q] THEN Inst = 0;
        END;
    END;
    IF inst = 1 THEN PIVi = PIVi||TOTAL[p,1];
  END;
  *array PIVi into matrix PIV accounting for fact that they vary in length;
  IF i = 1 THEN PIV = PIVi;
  ELSE DO;
    IF NCOL(PIV) = NCOL(PIVi) THEN;
    ELSE DO;
      IF NCOL(PIV) > NCOL(PIVi) THEN DO;  *Pad PIVi with zeros to be conformable;
        Add = J(1,(NCOL(PIV)-NCOL(PIVi)),0);
        PIVi = PIVi||Add;
      END;
      ELSE DO;                            *Pad PIV with zeros to be conformable;
        Add = J(NROW(PIV),(NCOL(PIVi)-NCOL(PIV)),0);
        PIV = PIV||Add;
      END;
    END;
    PIV = PIV//PIVi;
  END;
END;

PRINT PIV;

FINISH;

/*****************************************************************************/
/* MODULE: EFFECTS                                                           */
/*                                                                           */
/* Create TOTAL = Matrix with one row for every variable in the model.       */
/*        Column position 1 contains the variable index. Remaining           */
/*        columns contain indices of disturbances that have a non-zero       */
/*        total effect on that variable.                                     */
/*                                                                           */
/*****************************************************************************/
```

```
START EFFECTS;

/***** calculate total effects of disturbances on observed variables *****/

* FIRST CALCULATE TOTAL EFFECTS OF ERRORS ON Ys;
  IF y1 > 0 THEN DO;

  *calculate total effects of epsilon on ys;
  TotalE = VECDIAG(ThetaE);

  *calculate total effects of zeta on y1 through Beta;
  TotalZ_y1 = INV(I(NCOL(y1))-Beta);

  *calculate total effects of zeta on y2 through Lambda;
  IF y2 > 0 THEN DO;
    TotalZ_y2 = ly2*TotalZ_y1;
  END;

  * Replacing non-zero elements in TotalZ_y1 and TotalZ_y2 with appropriate zeta;
  DO i = 1 to NROW(TotalZ_y1);
    DO j = 1 to NCOL(TotalZ_y1);
      IF TotalZ_y1[i,j] ^= 0 THEN TotalZ_y1[i,j] = Psi[j,j];
    END;
  END;
  IF y2 > 0 THEN DO i = 1 to NROW(TotalZ_y2);
    DO j = 1 to NCOL(TotalZ_y2);
      IF TotalZ_y2[i,j] ^= 0 THEN TotalZ_y2[i,j] = Psi[j,j];
    END;
  END;

  *Combine TotalZ_y1 and TotalZ_y2 and sort by index to match up with TotalE;
  TotalZ_y1 = (y1`)||TotalZ_y1;                *Indexing TotalZ_y1;
  IF y2 > 0 THEN DO;
    TotalZ_y2 = (y2`)||TotalZ_y2;              *Indexing TotalZ_y2;
    TotalY = TotalZ_y1//TotalZ_y2;
  END;
  ELSE DO;
    TotalY = TotalZ_y1;
  END;
  TotalY2 = TotalY;
  TotalY[rank(TotalY[,1]),]=TotalY2;
  TotalY = TotalY||TotalE;

  IF x1 = 0 THEN DO;
    Total = TotalY;
  END;
END;

* CALCULATE TOTAL EFFECTS OF ERRORS ON Xs;
IF x1 > 0 THEN DO;
  IF x2 > 0 THEN TotalX = (x1`)//(x2`); *Creating Index Column of TotalX;
    ELSE TotalX = x1`;
  TotalD = VECDIAG(ThetaD);
  TotalX2 = TotalX;                            *Sorting Index Column;
  TotalX[rank(TotalX[,1]),]=TotalX2;
  TotalX = TotalX||TotalD;                     *Adding total effects of delta;

  IF y1 = 0 THEN DO;
    Total = TotalX;
  END;
END;
```

```
* COMBINE TOTALX AND TOTALY INTO ONE MATRIX;
* Need to pad TotalX or TotalY with 0s so equivalent # columns in each;
IF y1 > 0 & x1 > 0 THEN DO;
  IF NCOL(TotalY) > NCOL(TotalX) THEN DO;
    Add = J(NROW(TotalX),(NCOL(TotalY)-NCOL(TotalX)),0);
    TotalX = TotalX||Add;
  END;
  ELSE DO;
    Add = J(NROW(TotalY),(NCOL(TotalX)-NCOL(TotalY)),0);
    TotalY = TotalY||Add;
  END;
Total = TotalY//TotalX;
END;

* Sort Total;
Total2 = Total;
Total[rank(Total[,1]),]=Total2;

PRINT TOTAL;
FINISH;

/*****************************************************************************/
/* MODULE: COMPOSITE                                                         */
/*                                                                           */
/* Create COMP = Matrix with one row for every equation in the model. Column */
/*        position 1 contains the dependent variable index. Remaining        */
/*        columns contain indices of disturbances that have a non-zero       */
/*        total effect on that variable.                                     */
/*                                                                           */
/***************************************************************************** */

START COMPOSITE;

* First need to keep track of position of each variable index in y or x series;
* Sorting Y and X vectors into index order;

IF y1 > 0 THEN DO;
  VecY = y1`;
  IF y2 > 0 THEN VecY = VecY//y2`;
  OrderY = VecY;
  OrderY[rank(OrderY),]=VecY;
  OrderY = Rank(OrderY)||OrderY;
END;

IF x1 > 0 THEN DO;
  VecX = x1`;
  IF x2 > 0 THEN VecX = VecX//x2`;
  OrderX = VecX;
  OrderX[rank(OrderX),]=VecX;
  OrderX = Rank(OrderX)||OrderX;
END;

* Find Composite Disturbance for each equation;
DO i = 1 to NROW(PRED);              *for each DV, pull predictors from PRED;
  COMPi = PRED[i,1];                 *set first column in row of DIST to dv index;
  IF y1 > 0 THEN DO j = 1 to NCOL(y1);
                                     *if DV is in y1, add corresponding zeta;
    IF COMPi = y1[j] THEN DO;
      COMPi = COMPi||Psi[j,j];
    END;
  END;
```

```
  DO j = 1 to NCOL(PRED);       *examine each variable in equation (including dv);
    IF y1 > 0 THEN DO k = 1 to NROW(OrderY);
                                *identify source of variable and find its error;
      IF PRED[i,j] = OrderY[k,2] THEN
        COMPi = COMPi||ThetaE[OrderY[k,1],OrderY[k,1]];
    END;
    IF x1 > 0 THEN DO k = 1 to NROW(OrderX);
      IF PRED[i,j] = OrderX[k,2] THEN
        COMPi = COMPi||ThetaD[OrderX[k,1],OrderX[k,1]];
    END;
  END;
  *array COMPi into matrix COMP accounting for fact that they vary in length;
  IF i = 1 THEN COMP = COMPi;
  ELSE DO;
    IF NCOL(COMP) = NCOL(COMPi) THEN;
    ELSE DO;
      IF NCOL(COMP) > NCOL(COMPi) THEN DO;
                                      *Pad COMPi with zeros to be conformable;
        Add = J(1,(NCOL(COMP)-NCOL(COMPi)),0);
        COMPi = COMPi||Add;
      END;
      ELSE DO;                        *Pad COMP with zeros to be conformable;
        Add = J(NROW(COMP),(NCOL(COMPi)-NCOL(COMP)),0);
        COMP = COMP||Add;
      END;
    END;
    COMP = COMP//COMPi;
  END;
END;

PRINT COMP;

FINISH;

/*****************************************************************************/
/* MODULE: PREDICTORS                                                        */
/*                                                                           */
/* Create PRED = Matrix with one row for every equation in the model. Column */
/*        position 1 contains the dependent variable index. Remaining        */
/*        columns contain varaible indices of the predictors in the          */
/*        equation.                                                          */
/*                                                                           */
/*****************************************************************************/

START PREDICTORS;

*Constructing PRED matrix of predictors for which IVs are needed;

*Create PREDi for each variable in y1 (y1_i);
*Search Beta and Gamma matrix for y1s and x1s that influence y1_i;
IF y1 > 0 THEN DO i = 1 to NROW(Beta);
  PREDi = y1[i];                                  *Starting with DV index;
  DO j = 1 to NCOL(Beta);
    IF Beta[i,j] ^= 0 THEN DO;
      PREDi = PREDi||y1[j];
    END;
  END;
  IF x1 > 0 THEN
    DO j = 1 to NCOL(Gamma);
      IF Gamma[i,j] ^= 0 THEN DO;
        PREDi = PREDi||x1[j];
```

```
      END;
    END;
  *Add PREDi to larger PRED matrix here;
  IF i=1 THEN DO;
    PRED = PREDi;
  END;
  ELSE DO;
    *array PREDi into matrix PRED accounting for fact that they vary in length;
    IF NCOL(PRED) ^= NCOL(PREDi) THEN DO;
      IF NCOL(PRED) > NCOL(PREDi) THEN DO;
                                          *Pad PREDi with zeros to be conformable;
        Add = J(1,(NCOL(PRED)-NCOL(PREDi)),0);
        PREDi = PREDi||Add;
      END;
      ELSE DO;                            *Pad PRED with zeros to be conformable;
        Add = J(NROW(PRED),(NCOL(PREDi)-NCOL(PRED)),0);
        PRED = PRED||Add;
      END;
    END;
    PRED = PRED//PREDi;
  END;
END;

*Create PREDi for y2, concatenate with PRED;
IF y2 > 0 THEN DO i = 1 to NCOL(y2);
  PREDi = y2[i];                              *Starting with DV index;
  DO j = 1 to NCOL(y1);
    IF ly2[i,j] > 0 THEN DO;
      PREDi = PREDi||y1[,j];
    END;
  END;
  *array PREDi into matrix PRED accounting for fact that they vary in length;
  IF NCOL(PRED) = NCOL(PREDi) THEN;
  ELSE DO;
    IF NCOL(PRED) > NCOL(PREDi) THEN DO; *Pad PREDi with zeros to be conformable;
      Add = J(1,(NCOL(PRED)-NCOL(PREDi)),0);
      PREDi = PREDi||Add;
    END;
    ELSE DO;                               *Pad PRED with zeros to be conformable;
      Add = J(NROW(PRED),(NCOL(PREDi)-NCOL(PRED)),0);
      PRED = PRED||Add;
    END;
  END;
  PRED = PRED//PREDi;
END;

*Create PREDi for x2, concatenate with PRED;
IF x2 > 0 THEN DO i = 1 to NCOL(x2);
  PREDi = x2[i];                              *Starting with DV index;
  DO j = 1 to NCOL(x1);
    IF lx2[i,j] > 0 THEN DO;
      PREDi = PREDi||x1[,j];
    END;
  END;
  IF (y1 = 0) & (i=1) THEN DO;
    PRED = PREDi;
  END;
  ELSE DO;
    *array PREDi into matrix PRED accounting for fact that they vary in length;
    IF NCOL(PRED) = NCOL(PREDi) THEN;
    ELSE DO;
```

```
      IF NCOL(PRED) > NCOL(PREDi) THEN DO;
                                        *Pad PREDi with zeros to be conformable;
        Add = J(1,(NCOL(PRED)-NCOL(PREDi)),0);
        PREDi = PREDi||Add;
      END;
      ELSE DO;                          *Pad PRED with zeros to be conformable;
        Add = J(NROW(PRED),(NCOL(PREDi)-NCOL(PRED)),0);
        PRED = PRED||Add;
      END;
    END;
    PRED = PRED//PREDi;
  END;
END;

PRINT PRED;

FINISH;

/*****************************************************************************/
/* MODULE: MAIN                                                              */
/*                                                                           */
/* User assigns index numbers to variables and disturbances and defines the  */
/* model                                                                     */
/*                                                                           */
/*****************************************************************************/

START MAIN;

/* INDEX VARIABLES IN THE MODEL                                              */
/* Create row vectors designating index of scaling and nonscaling variables. */
/* Numbers will subsequently be used to reference variables. Use {0} if no   */
/* variables are in the vector.                                              */
/* Note that the order is arbitrary (ys do not have to precede xs etc), but  */
/* the specified order must be followed in other matrices. For example,      */
/* ThetaE(1,1)references the error variance of whichever y has the first     */
/* position (lowest number) in the variable list for y1 and y2.              */

y1 = {1 5};            * scaling y's;
y2 = {2 3 4 6 7 8};    * non-scaling y's;
x1 = {9};              * scaling x's (and exogenous x's);
x2 = {10 11};          * non-scaling x's;

/* SET UP PARAMETER MATRICES                                                 */
/* Parameter matrices should be specified in binary -- non-zero elements     */
/* labeled 1; zero elements labeled 0. Set null matrices to {0}.             */

/* Create loading matrix for y2 and x2 variables = lambda y2 and lambda x2   */

  ly2 = {1 0,
         1 0,
         1 0,
         0 1,
         0 1,
         0 1};

  lx2 = {1,
         1};

/* Create matrices of structural parameters Beta and Gamma                   */
/* Note - may be necessary to use values other than 1 to code non-zero effects */
/*        in Beta if you get an error that 'Matrix should be non-singular'   */
```

```
  Beta = {0 0,
          1 0};

  Gamma = {1,
           1};

/* INDEX DISTURBANCES                                                          */
/* Each unique non-zero element in ThetaE, ThetaD and Psi should be given a    */
/* unique index number. It is helpful to label the diagonals of ThetaE and     */
/* ThetaD with the same numers that index the Y and X variables that the       */
/* elements correspond to for easy reference later. If there are no parameters */
/* in a matrix, set it to {0}. Note that upper and lower off-diagonal matrices */
/* should contain the same numbers.                                            */

  ThetaE = { 1  0  0  0 12  0  0  0,
             0  2  0 16  0 13  0  0,
             0  0  3  0  0  0 14  0,
             0 16  0  4  0  0  0 15,
            12  0  0  0  5  0  0  0,
             0 13  0  0  0  6  0 17,
             0  0 14  0  0  0  7  0,
             0  0  0 15  0 17  0  8};

  ThetaD = { 9  0  0,
             0 10  0,
             0  0 11};

  Psi = {18 0,
         0 19};

RUN PREDICTORS;
RUN COMPOSITE;
RUN EFFECTS;
RUN POTENTIAL_IV;
RUN FINAL_IV;

FINISH;
RUN;
QUIT;
```

NOTES

1. We remind the reader that the selection of the instrumental variables (IVs) is conditional on the model specification. If an equation is overidentified such that there are more IVs than the minimum needed for that equation, then there are overidentification tests available that can point to problems in the model specification. We return to this issue in the summary.

2. See Bollen (1987) for stability conditions that are typically applied when defining total and indirect effects.

3. This dummy coding scheme may occasionally lead $\mathbf{I} - \mathbf{B}$ to be noninvertable (if any row or column of $\mathbf{I} - \mathbf{B}$ can be expressed as a linear combination of the other rows or columns). In such a situation, the user could adopt a nonsense coding scheme, inserting other nonzero values in place of some of the 1s in $\mathbf{B}$ to remove the linear redundancy.

REFERENCES

Anderson, T. W. 1951. "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions." *Annals of Mathematical Statistics* 22:327-51.

Anderson, T. W. and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20:46-63.

Basmann, R. L. 1957. "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation." *Econometrica* 25:77-83.

———. 1960. "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics." *Journal of the American Statistical Association* 55:650-59.

Bollen, Kenneth A. 1987. "Total, Direct, and Indirect Effects in Structural Equation Models." *Sociological Methodology* 17:37-69.

———. 1989. *Structural Equations With Latent Variables*. New York: John Wiley.

———. 1995. "Structural Equation Models That Are Nonlinear in Latent Variables: A Least Squares Approach." *Sociological Methodology* 25:223-51.

———. 1996. "An Alternative Two Stage Least Squares (2SLS) Estimator for Latent Variable Equations." *Psychometrika* 61:109-21.

———. 2001. "Two-Stage Least Squares and Latent Variable Models: Simultaneous Estimation and Robustness to Misspecifications." Pp. 119-38 in *Structural Equation Modeling: Present and Future*, edited by R. Cudeck, S. du Toit, and D. Sorbom. Lincolnwood, IL: Scientific Software.

Bollen, Kenneth A. and Pamela Paxton. 1998. "Two-Stage Least Squares Estimation of Interaction Effects." In *Interaction and Nonlinear Effects in Structural Equation Modeling*, edited by R. E. Schumaker and G. A. Marcoulides. Mahwah, NJ: Lawrence Erlbaum.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90:443-50.

Bowden, Roger J. and Darrell A. Turkington. 1984. *Instrumental Variables*. New York: Cambridge University Press.

Browne, Michael W. 1984. "Asymptotic Distribution Free Methods in Analysis of Covariance Structures." *British Journal of Mathematical and Statistical Psychology* 37:62-83.

Cudeck, Robert 1991. "Noniterative Factor Analysis Estimators, With Algorithms for Subset and Instrumental Variable Selection." *Journal of Educational Statistics* 16:35-52.

Davidson, Russell and James Mackinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Hägglund, Gösta. 1982. "Factor Analysis by Instrumental Variables." *Psychometrika* 47:209-22.

Jennrich, Robert I. 1987. "Tableau Algorithms for Factor Analysis by Instrumental Variable Method." *Psychometrika* 52:469-76.

Johnston, John. 1972. *Econometric Methods*. New York: McGraw-Hill.

Jöreskog, Karl G. and Dag Sörbom. 1993. *LISREL 8*. Mooresville, IN: Scientific Software.

Madansky, Albert. 1964. "Instrumental Variables in Factor Analysis." *Psychometrika* 29:105-13.

Magdalinos, Michael A. 1985. "Selecting the Best Instrumental Variables Estimator." *Review of Economic Studies* 52:473-85.

Mariano, Roberto S. 1977. "Finite Sample Properties of Instrumental Variable Estimators of Structural Coefficients." *Econometrica* 45:487-96.

Oczkowski, Edward. 2002. "Discriminating Between Measurement Scales Using Non-Nested Tests and 2SLS: Monte Carlo Evidence." *Structural Equation Modeling* 9:103-25.

Pesaran, M. Hashem and Larry W. Taylor. 1999. "Diagnostics for IV Regressions." *Oxford Bulletin of Economics and Statistics* 61:255-81.

Sargan, J. D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26:393-415.

Kenneth A. Bollen is the director of the Odum Institute for Research in Social Science and the H. R. Immerwahr Distinguished Professor of Sociology at the University of North Carolina at Chapel Hill. He was the recipient of the Lazarsfeld Award for Methodological Contributions in Sociology in 2000. The ISI named him among the World's Most Cited Authors in the Social Sciences. His primary area of statistical research is structural equation modeling and latent curve models. He is the author of Structural Equation Models With Latent Variables *and more than 90 papers. Recent articles have been published in* Sociological Methods & Research, Sociological Methodology, Social Forces, Sociology of Education, Annual Review of Psychology, Population Studies, *and* Studies in Comparative International Development.

Daniel J. Bauer is an assistant professor in the Department of Psychology at North Carolina State University. His quantitative research focuses on structural equation models, multilevel models, and mixture models, with a particular emphasis on the analysis of change. His recent publications have appeared in Psychological Methods, Journal of Educational and Behavioral Statistics, *and* Social Forces.