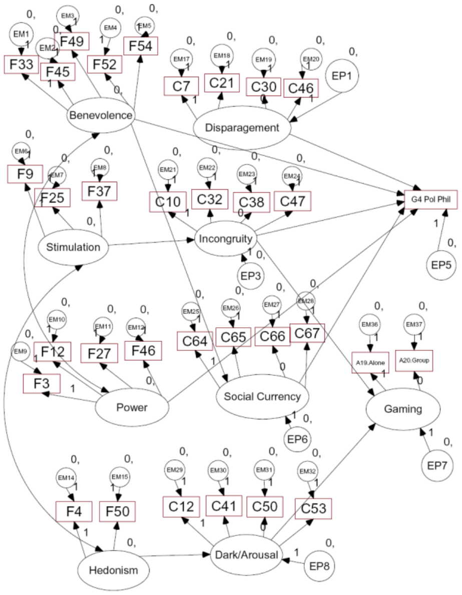
Neuendorf
Structural Equation Modeling

Structural equation modeling is useful in situations when we have a complicated set of relationships among variables as specified by theory. Two main methods have been employed to assess whether a complex and/or multi-step "causal" model is explained by the data at hand: (1) *Structural equation modeling* (SEM) solves multiple equations simultaneously, an improvement over (2) the older, often by-hand process of *path analysis*. SEM allows for the combining of a structural/theoretic model with a measurement model. As Hair et al. note, SEM is "an extension of several multivariate techniques we have already studied, most notably multiple regression and factor analysis."

Model:

In SEM, the researcher literally builds a model! e.g.,



Complex!! (See larger version near end of this handout)

The components of such an SEM model may include:

1. Exogenous latent constructs (ovals with no predictors)
2. Endogenous latent constructs (ovals with predictors)
3. Measured/observed variables (rectangles)
4. Errors of measurement (circles attached to measured variables)
5. Errors of prediction (circles attached to endogenous latent constructs or variables)
6. Causal paths (single-headed straight arrows)
7. Correlational links (double-headed curved arrows)

<u>Assumptions</u>:

1. Measures are at the I/R level, independent observations, and distributions are normal and multivariate normal.

2. Even though SEM allows for multi-step models, theoretic constructs are deemed either “exogenous” (similar to independent variables) and “endogenous” (similar to dependent variables). Exogenous means that the construct is not predicted by any other construct; endogenous means there's at least one causal predictor of that variable (i.e., there is at least one causal path leading to it, a single-headed arrow).

3. SEM allows for both (1) unmeasured, “latent” variables/constructs (structural/theoretic model only) and (2) measured, “observed” variables (measurement model or structural/theoretic model). In AMOS, the latent variables are diagrammed as ovals, and the measured/observed variables are rectangles. These rectangles represent actual measured variables in an SPSS data set; when building the model, you must link them up.

4. Relationships among the constructs and variables are linear.

5. Analysis is at the aggregate level. (e.g., In LISREL, the "data" consist of a correlation matrix--already aggregated over many respondents.)

6. The model and all its components need to be “identified”--generally, identification is related to the number of equations that AMOS needs to solve for the model, and the number of coefficients to be estimated. In essence, identification indicates that the multiple equations implied by the model are “solvable.” [The attached page from Asher uses algebraic equations as a useful analogy.] The issue of model identification is a very complex one--you can't just visually peruse a model and declare it "identified." Some treatments of identification examine two “practical” conditions that test for model identification: The order condition and the rank condition (see attached pages from Maruyama). For a fuller discussion, see Asher's *Causal Modeling*, Blalock's *Theory Construction* or Heise's *Causal Analysis*.

<u>Decisions to make</u>:

1. Will there be a structural (theoretic) model? Will there be a measurement model? Will both be included?

   Typically, both a structural/theoretic model and a measurement model will be included. An exception would be a structural/theoretic model in which all constructs are each measured by a single measure. Another exception would be a measurement model for CFA (confirmatory factor analysis), which may have no structural/theoretical causal paths specified.

2. What will be the causal paths among the theoretic constructs? Will there be any correlational links specified between exogenous variables? Any correlational links between errors of prediction on the endogenous variables? Any correlational links between measurement errors in the measurement model?

3. Will there be any nonrecursive "feedback loops" in the model?

4. What estimation procedure will be used? LISREL offers maximum likelihood (default), two-stage least squares and several other options. AMOS offers maximum likelihood (default), generalized least squares and several other options.

Statistics:

1. Goodness of fit indices (absolute, incremental, or parsimony). Hair has a fine description of the choices on pp. 648-652. Arbuckle and Wothke have formulas and a classification scheme. Generally, these goodness of fit indices assess whether the model as a whole is "good" (well specified, fitting the data at hand), sort of like R^2 in multiple regression but here for everything going on in the model. Some of the more commonly used goodness of fit measures are:

   Chi-square: we want it to be small, nonsig.
   GFI: absolute goodness of fit index; we want it to be large, near 1
   AGFI: adjusted GFI; we want it to be large, over .90
   NFI: normed fit index; an incremental goodness of fit index; a recent "practical criterion of choice" (Byrne); we want it to be large, over .90
   CFI: comparative fit index; an incremental goodness of fit index; like the NFI with sample size taken into account; we want it to be large, over .90
   RMSEA: Root mean square error of approximation; avg. correlation among residuals; we want it to be small, substantially smaller than the original variable intercorrelations in the raw data matrix; Byrne cites several sources and presents standards of .05 or smaller as a good fit, .08 or smaller as a reasonable fit, and .08-.10 as "mediocre fit."

2. Path coefficients--are essentially like partial regression coefficients; they show the unique (partial), unstandardized or standardized contribution of one variable to another's variance. For each coefficient, LISREL also provides a SE and a t-test to test whether it differs significantly from 0. AMOS provides a SE and a C.R. (critical ratio), which is Estimate/SE, and is equivalent to LISREL's t. In both cases, anything greater than 1.96 is considered significant. AMOS will show the actual level of *p*.

3. Multiple R^2s--one for each endogenous construct/variable. Indicates the proportion of the variance of that variable that is explained by the model.

4. Degrees of freedom = (roughly) # of elements in the correlation matrix - # of parameters to be estimated. It cannot be less than zero (if so, the model in total is not identified). Notice df is not related to n.

5. Modification indices--for each unspecified potential path or link in a model, LISREL or

AMOS can calculate how much the Chi-square goodness of fit indicator will be reduced by adding that path or link. If the incremental improvement is significant, you may wish to consider adding the path/link and rerunning. In AMOS, you cannot obtain modification indices if you have any missing data.

6. IF you have a measurement model--"construct reliabilities" and "average variance extracted" may be hand-calculated from loadings (relationships between measured variables and structural constructs). See Hair p. 687 for formulae.

References

Arbuckle, J. L. (2003). *AMOS 5.0 update to the AMOS user's guide.* Chicago, IL: SmallWaters Corp.

Arbuckle, J. L. (2007). *AMOS 16.0 user's guide.* Chicago, IL: SPSS, Inc. Accessed from: http://www.uni-muenster.de/imperia/md/content/ziv/service/software/spss/handbuecher/englisch/amos16.0_user_s_guide.pdf

Arbuckle, J. L., & Wothke, W. (1995-99). *AMOS 4.0 user's guide*. Chicago, IL: SmallWaters Corp.

Asher, H. D. (1983). *Causal modeling* (2nd ed.). Newbury Park, CA
Publications.

Blalock, H. M. Jr. (Ed.). (1985). *Causal models in the social sciences* (2nd ed.). Chicago, IL: Aldine Publishing.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL*. Baltimore, MD: The Johns Hopkins University Press.

Heise, D. R. (1975). *Causal analysis*. New York: Wiley.

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage Publications.

Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage Publications.

[Attachments: 3 pages from Asher, 2 pages from Maruyama; 2 pages from AMOS]

4/19

From: Asher, H. D. (1983). *Causal modeling* (2nd ed.). Newbury Park, CA: Sage Publications.

AN INTUITIVE LOOK AT IDENTIFICATION

Before formally discussing identification, it may be helpful to consider the topic from an intuitive perspective by employing analogies to algebraic systems of equations. If we think of the problem of identification as having sufficient information to come up with a unique solution(s) for a set of unknowns, then we might by analogy say that equations 3.9 and 3.10 below represent an exactly identified system since there are just as many linearly independent equations as unknowns, thereby yielding the unique solution set $X = 2$ and $Y = 1$.

$$2X + 3Y = 7 \qquad [3.9]$$

$$X - 4Y = -2 \qquad [3.10]$$

Equations 3.11 and 3.12 might be considered an underidentified system since there are fewer linearly independent equations than unknowns. That is, while there are two equations in two unknowns, the second equation is simply a multiple of the first which means that it does not contribute any information in solving for X and Y. Hence, we in effect have one equation in two unknowns. This means that an infinite set of solutions exists; for example, $X = 2$ and $Y = 1$, $X = 3.5$ and $Y = 0$, and $X = 5$, and $Y = -1$ are all solution sets for equations 3.11 and 3.12. If we are trying to come up with interpretable estimates of some unknowns, then the situation of infinite solutions is obviously fatal to any sound inferences.

$$2X + 3Y = 7 \qquad [3.11]$$

$$4X + 6Y = 14 \qquad [3.12]$$

Finally, equations 3.13 through 3.15 can be viewed as an overidentified system since there are more equations than unknowns. If we used equations 3.13 and 3.14 to solve, we would get $X = 3$ and $Y = -1$. Equations 3.13 and 3!5 generate solutions of $X = 12$ and $Y = 17$, while equations 3.14 and 3.15 yield $X = 6/11$ and $Y = -2/11$. While there is only a finite set of solutions here, it still is the case that different pairs of equations give very dissimilar results, again an unsatisfactory situation for making inferences on the basis of the estimates. In equations 3.13 through 3.15, the excess equation is not consistent with the other two.

$$2X - Y = 7 \qquad [3.13]$$

$$X + 3Y = 0 \qquad [3.14]$$

$$3X - 2Y = 2 \qquad [3.15]$$

From: Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage Publications.

Model Identification

Unlike recursive path models without measurement error that always will be identified, there is no guarantee that a unique solution can be obtained for nonrecursive models. Some nonrecursive models can be underidentified and therefore not solvable. To ensure identification, certain conditions need to be met. Those conditions can be met when some of the predictor variables do *not* have direct paths to certain endogenous variables. The term frequently used to describe such variables is **instrumental variable** or **instrument.** A predictor variable serves as an instrument for an endogenous variable and helps to identify its equation, provided that variable has a direct path to other endogenous variables but *not* to the variable of interest. For a model to be identified, each equation needs to have as many instruments (variables without direct paths) as there are variables in reciprocal relationships. Furthermore, as is explained in the next section of this chapter when the rank condition for identification is described, the instruments have to be distributed in particular ways for each dependent variable to have a solvable equation.

Consider the illustration in Figure 6.2. How could we tell whether or not it is identified? Consider first estimating the model including the dashed line path. There are five variables (therefore, 5 × 4 / 2 = 10 degrees of freedom) and exactly 10 paths, suggesting that the model might be identified. Furthermore, X_3 is an instrument for the Y_1 equation. But notice also that all three of the exogenous variables have arrows directly to Y_2, which means that the endogenous variable has no instruments, and therefore its equation is not identified. Once the dashed line path is dropped, however, X_1 becomes an instrument for the equation of Y_2 and the model becomes identified.

There are two conditions that must be met to ensure identification. Before presenting these two conditions, however, it should be noted that, particularly for complex models, ensuring identification may be very difficult. In principle, however, the computer programs that analyze structural equation models should provide tests for model identification. If the proposed model is underidentified, then the program should not be able to generate a complete solution. Specifically, calculation of confidence intervals requires inverting the matrix of estimates. A matrix called the information matrix (see, e.g., Jöreskog & Sörbom, 1988), which is based on the matrix of estimates, should be singular and noninvertible for an underidentified model, with the result that confidence intervals cannot be produced for the estimated parameters. Although this should provide a surefire test of model identification, there is waffling about identification because exceptions seem to have been found. Therefore, readers concerned about complex models are referred to the works of Bollen (e.g., 1989) and his colleagues as well as Rigdon (1995).

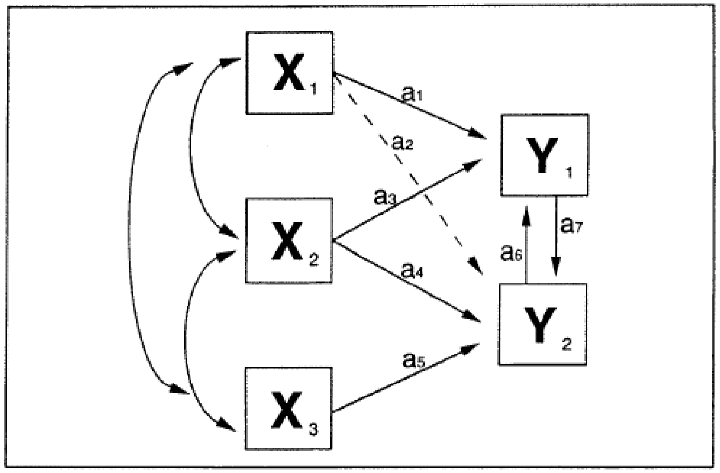


Figure 6.2. Nonrecursive Path Model to Illustrate Model Identification

The treatment of identification issues for manifest variable models that I find most understandable is the one presented by Namboodiri, Carter, and Blalock (1975, pp. 502-505). I will try to model my description after theirs. The first condition, which they called the **order condition,** is a necessary but not a sufficient condition for identification. It requires that for any system of N endogenous variables (which therefore means that there will be N equations, one for each endogenous variable), **a particular equation will be identified only if at least $N - 1$ variables are left out of that equation** (i.e., their regression weights are set to 0). For the Y variables in Figure 6.2, with two endogenous variables the two equations are

$$Y_1 = a_1{*}X_1 + a_3{*}X_2 + 0{*}X_3 + a_6{*}Y_2 + e_1 \qquad (6.1)$$
$$Y_2 = a_2{*}X_1 + a_4{*}X_2 + a_5{*}X_3 + a_7{*}Y_1 + e_2. \qquad (6.2)$$

The residuals can be ignored because they go to the left side of the equation, whereas the dependent variables join the other variables on the right side of the equation (the signs on the coefficients also are inconsequential and can be ignored), yielding

$$-e_1 = a_1{*}X_1 + a_3{*}X_2 + 0{*}X_3 - 1{*}Y_1 + a_6{*}Y_2$$
$$-e_2 = a_2{*}X_1 + a_4{*}X_2 + a_5{*}X_3 + a_7{*}Y_1 - 1{*}Y_2.$$

In terms of the **order condition**, each equation needs to have (2 – 1) variables omitted from the equation. The first equation is fine because X_3 is omitted, whereas the second equation fails to meet the order condition. Once the a_2 coefficient is set to zero, that equation also meets the order condition for identification.

The second condition, more restrictive than the first and both a necessary and a sufficient condition for identification, is called the **rank condition.** Given a system of *N* dependent variables, for the rank condition to be satisfied for a particular equation, it must be possible to form at least one nonzero determinant of rank *N* – 1 from the coefficients of the variables omitted from that equation. Using the last set of the preceding equations, with the residuals isolated from all other variables, follow these three steps.

1. Form a matrix from the coefficients (signs again can be ignored). For the example, it would be as follows:

	X_1	X_2	X_3	Y_1	Y_2
Y_1	a_1	a_3	0	1	a_6
Y_2	0	a_4	a_5	a_7	1

2. To test for identification of a particular equation, delete from the matrix (a) the row of that equation and (b) all columns that do not have a zero in the row of the equation of interest.
3. Find a nonzero determinant of rank *N* – 1 from the remaining values.

Concretely, for Y_1 the entire first row (the Y_1 row) is deleted, as are the first (X_1), second (X_2), fourth (Y_1), and fifth (Y_2) columns, leaving [a_5], which happens to be a 1 × 1 matrix with a nonzero determinant unless a_5 happens to be exactly 0. For Y_2, the entire second row is deleted, as are the second through fifth columns, leaving [a_1], another 1 × 1 matrix with a nonzero determinant unless a_1 is exactly 0. Because both a_5 and a_1 are being estimated, they are expected to be nonzero. If so, the modified Figure 6.2, with the dashed path from X_1 to Y_2 omitted, is an identified model. As suggested earlier, X_1 serves as an instrument for Y_2 and X_3 as an instrument for Y_1.

A few final points about identification are in order. First, if the *X* variables are highly intercorrelated, then it may make little sense to argue that one *X* can readily be dropped from each equation given that they share much common variance and are not easily distinguishable one from another. Ideally, instruments are basically independent of other exogenous variables. An important point is that although instruments are essential for attaining model identification, in some instances it may be very difficult to find variables that meet the requirements of good instruments. Second, what if the two endogenous variables "shared" the same instrument, for example, if in Figure 6.2 we were to put a_2 back into the model and remove a_5. The answer is that the rank condition no longer could be satisfied because the 1 × 1 matrices would be 0. The important point here is that **each endogenous variable in a reciprocal relationship needs its own separate instruments.**

From: Asher, H. D. (1983). *Causal modeling* (2nd ed.). Newbury Park, CA: Sage Publications.

CAUSAL MODELING

SECOND EDITION

HERBERT B. ASHER
Ohio State University

SAGE PUBLICATIONS
Beverly Hills / London / New Delhi

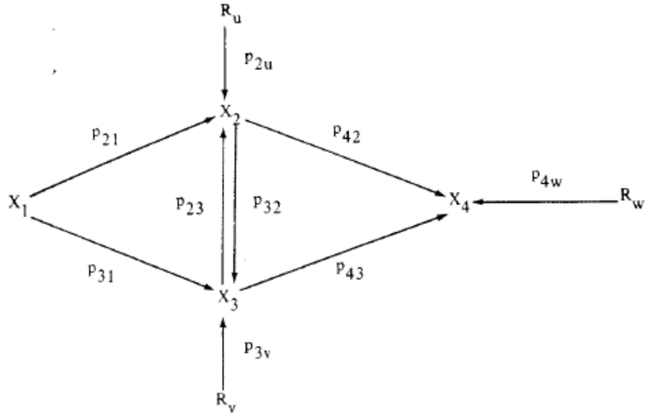
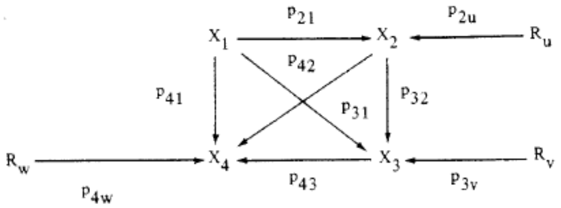


Figure 4: The Miller and Stokes Representation Model with Residual Variables and Path Coefficients Included

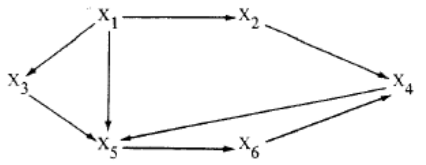


$$X_2 = p_{21}X_1 + p_{2u}R_u \quad [2.5]$$

$$X_3 = p_{31}X_1 + p_{32}X_2 + p_{3v}R_v \quad [2.6]$$

$$X_4 = p_{41}X_1 + p_{42}X_2 + p_{43}X_3 + p_{4w}R_w \quad [2.7]$$

Figure 8: A Four-Variable Recursive Model and Structural Equations



X_1: father's sociological characteristics
X_2: father's party identification
X_3: respondent's sociological characteristics
X_4: respondent's party identification
X_5: respondent's partisan attitudes
X_6: respondent's vote

Prediction Equations	Actual Values
$r_{41\cdot23} = 0$	0.017
$r_{61\cdot2345} = 0$	–0.019
$r_{32\cdot1} = 0$	0.101
$r_{52\cdot134} = 0$	0.032
$r_{62\cdot1345} = 0$	0.053
$r_{43\cdot12} = 0$	0.130
$r_{63\cdot1245} = 0$	–0.022
$r_{64\cdot1235} = 0$	0.365

Figure 9: The First Goldberg Model with Simon-Blalock Predictions and Actual Values

Blalock (1964: 80) suggests two general strategies of model revision:

(1) make revisions early in the model so that the effects of these revisions will be transmitted throughout the system of relationships, and
(2) revise where the discrepancies are greatest between the predicted and actual values.

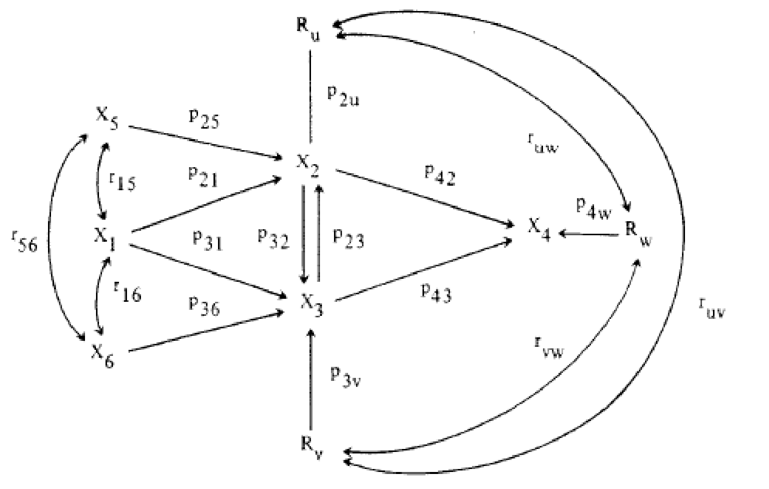


Figure 18: The Nonrecursive Miller and Stokes Representation Model with Two Hypothetical Exogenous Variables

The structural equations and matrix of coefficients for this revised model are given below.

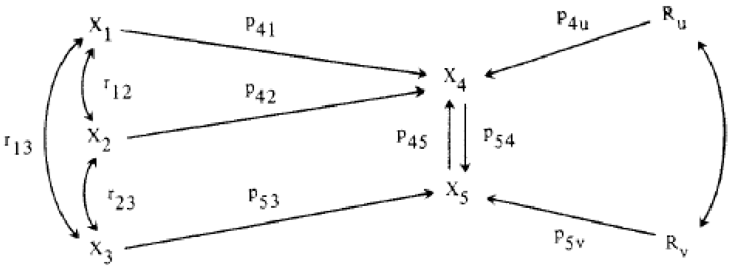
$$-p_{21}X_1 + X_2 - p_{23}X_3 - p_{25}X_5 = p_{2u}R_u \quad [3.22]$$

$$-p_{31}X_1 - p_{32}X_2 + X_3 - p_{36}X_6 = p_{3v}R_v \quad [3.23]$$

$$- p_{42}X_2 - p_{43}X_3 + X_4 = p_{4w}R_w \quad [3.24]$$

$$\begin{bmatrix} -p_{21} & 1 & -p_{23} & 0 & -p_{25} & 0 \\ -p_{31} & -p_{32} & 1 & 0 & 0 & -p_{36} \\ 0 & -p_{42} & -p_{43} & 1 & 0 & 0 \end{bmatrix}$$

Note that the addition of X_5 and X_6 to the model does not change the number of equations. Hence, the order condition requires that each equation excludes k-1 variables where k refers to the number of linear equations, in this case three. Thus, all three equations satisfy the order

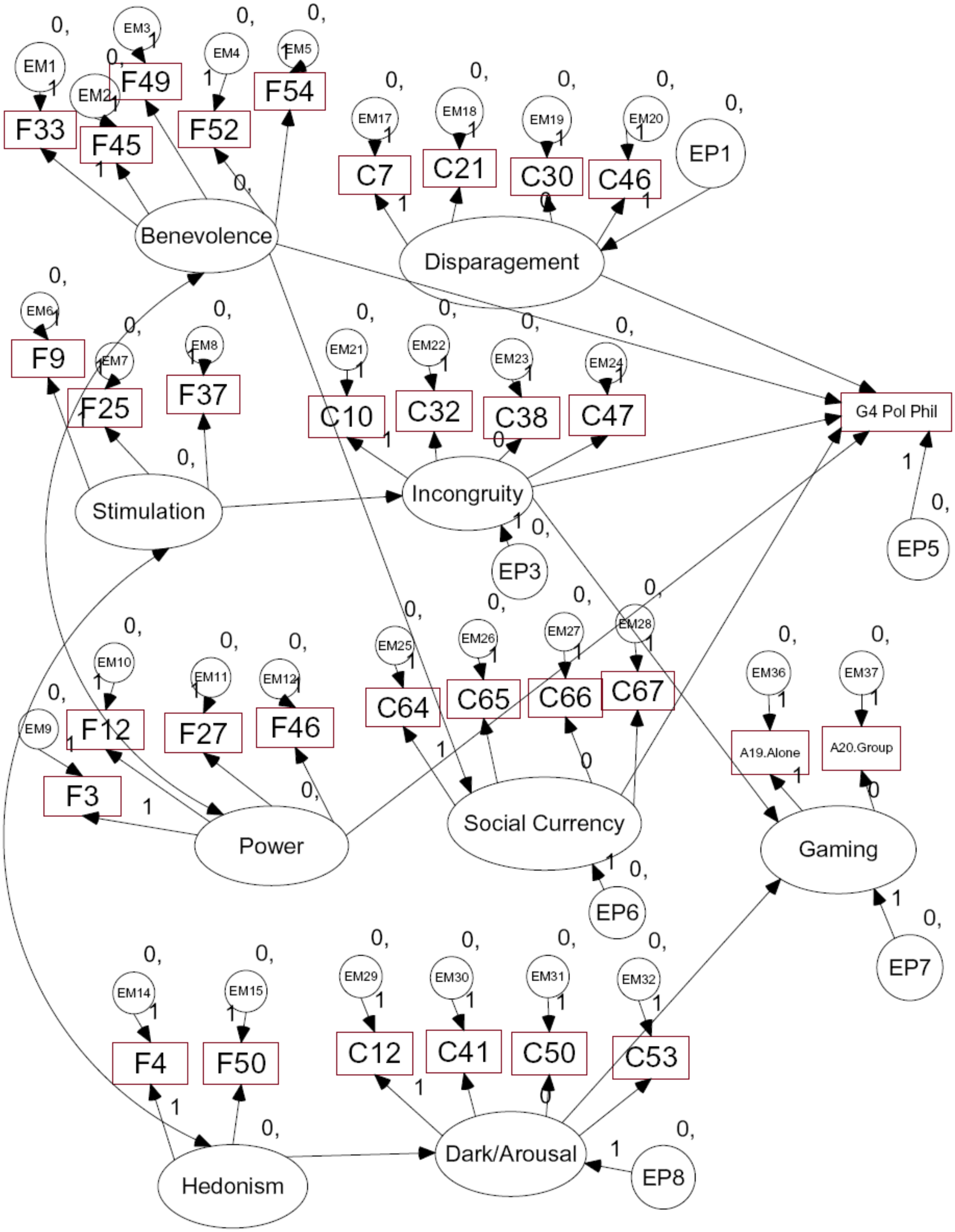


$$X_4 = p_{41}X_1 + p_{42}X_2 + p_{45}X_5 + p_{4u}R_u \quad [3.30]$$

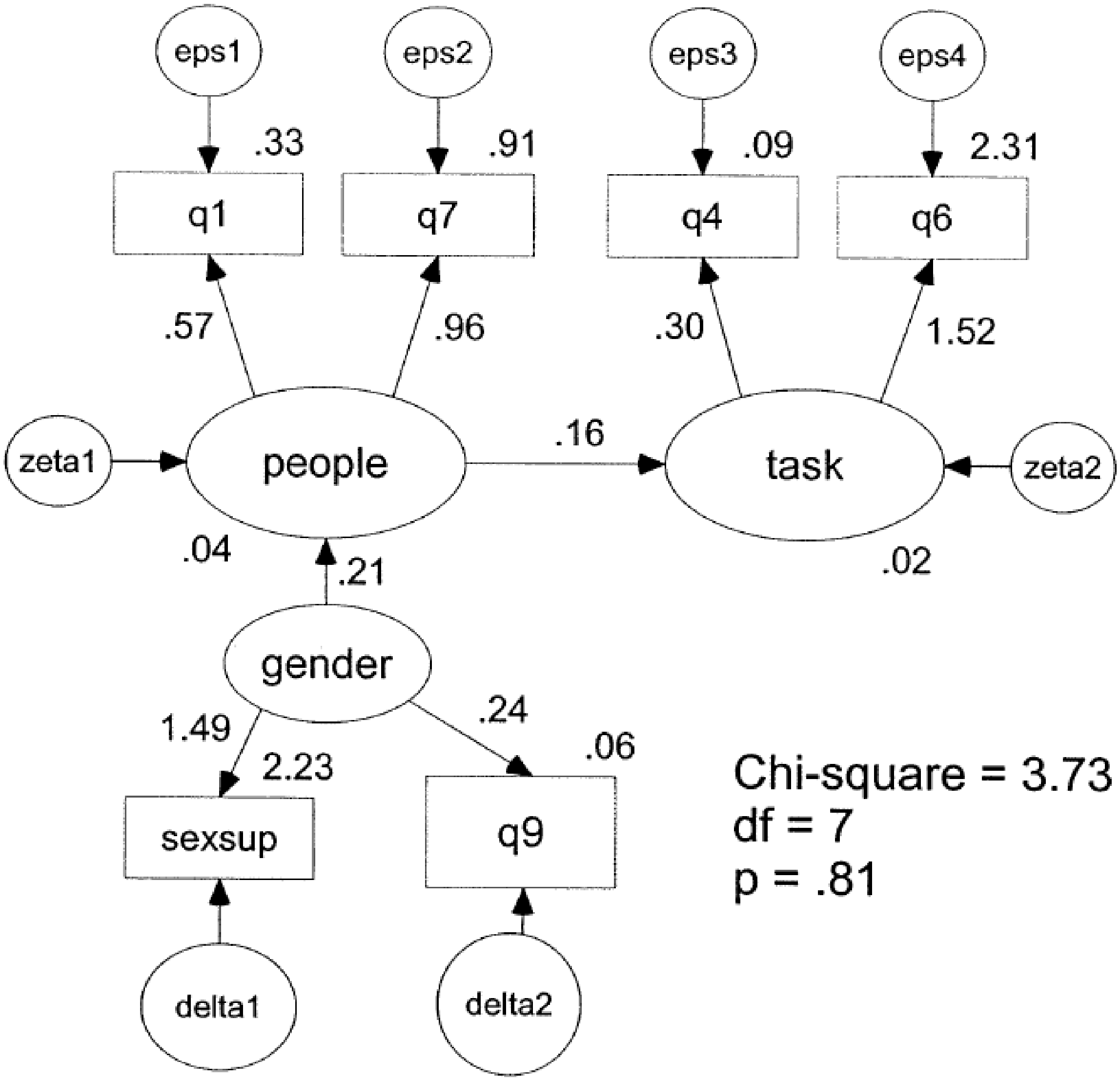
$$X_5 = p_{53}X_3 + p_{54}X_4 + p_{5v}R_v \quad [3.31]$$

Figure 21: Overidentification and Two-Stage Least Squares

Sample AMOS input.



Sample AMOS output.



Grover's Data/Kim's Model