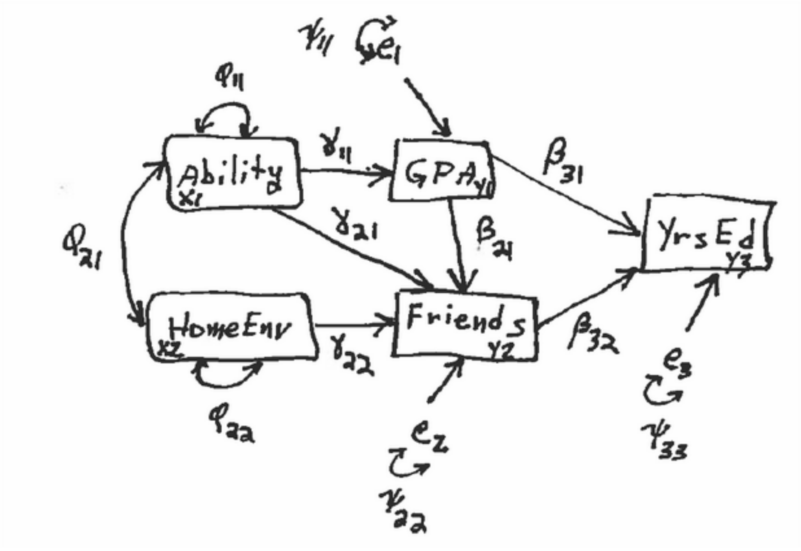
Specification Searches

When applying a path model to realistic problems (as opposed to small problems that are studied in class), it is practically inevitable that the model you first estimate and that appeared to be so ingenious when you initially began the study, does not fit. The question then becomes, What to do? A perfectly acceptable strategy is to investigate where the lack of fit occurs and modify the model so it performs better.

To be concrete, the model and data below are from a project that attempted to explain Years of Education (y_3) from Ability (x_1), the quality of the Home Environment (x_2), Grade Point Average in secondary school (y_1), and cohesion of a student's circle of Friends (y_2).



```
data EdLevel (type=corr); _type_=’corr’;
input _name_$ Ability HomeEnv GPA Friends YrsEd;
datalines;
  Ability    1   .   .   .   .
  HomeEnv   .29  1   .   .   .
  GPA       .59 .40  1   .   .
  Friends   .48 .36 .37  1   .
  YrsEd     .41 .48 .45 .61  1
                 ;
proc calis data=EdLevel nobs=200 residual;
  path Ability             -> GPA,
       Ability HomeEnv GPA -> Friends,
       GPA Friends         -> YrsEd;
run;
```

```
Predicting years of education
Observed variables
  Ability HomeEnv GPA Friends YrsEd
Correlation matrix
    1
  .29   1
  .59 .40   1
  .48 .36 .37   1
  .41 .48 .45 .61   1
Sample size 200
Relationships
  Ability             -> GPA
  Ability HomeEnv GPA -> Friends
  GPA Friends         -> YrsEd
LISREL OU rs mi wp
End of problem
```

Figure 1. Path diagram, SAS and LISREL code for Model 1 which predicts GPA, Friends Cohesion and Education Level from Ability and Hone Environment

The model specifies 6 regression coefficients plus 6 variances for $p = 5$ variables. Degrees of freedom $df = 5 \cdot 6/2 - 12 = 3$. The test of exact fit gives $T_{ml} = 35.3$ with upper tail probability $p < 0.0001$. *RMSEA = 0.23.* Parameter estimates are on the left side of Table 1. Residual correlations are on the right. These are defined as

$$\mathbf{R}_s = \mathbf{R} - \hat{\mathbf{P}}$$

where $\mathbf{R}$ is the sample correlation matrix and $\hat{\mathbf{P}}$ the reproduced matrix, and the estimated population correlation matrix, from the model. Two residuals are large, greater than 0.20. In particular residual correlations for (x_2, y_1) and $(x_2, y_{3)}$ are $[\mathbf{R}_s]_{32} = 0.23$, $[\mathbf{R}_s]_{52} = 026$, where $[\mathbf{R}_s]_{jk}$ is the residual for the $(j,k)th$ pair of variables. By all indications the model is not completely satisfactory.

Table 1. Parameter estimates, residual correlations and measures of fit for Model 1.

$\hat{\mathbf{B}}$

	y_1	y_2	y_3
y_1 GPA	0	0	0
y_2 Friends	.05	0	0
y_3 YrsEd	.26	.52	0

$\hat{\boldsymbol{\Gamma}}$

	x_1	x_2
y_1	.59	$\boxed{0}$
y_2	.38	.23
y_3	0	$\boxed{0}$

Residual correlationss, $\mathbf{R}_s = \mathbf{R} - \hat{\mathbf{P}}$

	x_1	x_2	y_1	y_2	y_3
x_1 Ability	00				
x_2 HomeEnv	00	00			
y_1 GPA	00	23	00		
y_2 Friends	00	01	05	01	
y_3 YrsEd	01	26	03	02	02

$\hat{\boldsymbol{\Psi}}$

	y_1	y_2	y_3
y_1 GPA	.61		
y_2 Friends	0	.72	
y_3 YrsEd	0	0	.58

$\hat{\boldsymbol{\Phi}}$

	x_1	x_2
x_1	1.0	
x_2	.29	1.0

$T_{ml} = 35.3$, $df = 3$, $p < 10^{-5}$
$RMSEA = 0.23$

With path analysis it is easy to modify a poorly fitting model.

(A) Add a path between the two variables with the large residual correlation.

(B) If the residual correlation is between two endogenous variables, allow their residuals variables, e_j and e_k, to covary.

Of course,

> The direction of the path, and even the decision as to whether the path should be included at all, have to make scientific sense.

Everyone realizes its easy to kid yourself about this. "I knew all along that path should have been there! Silly me for having forgotten to include it in the original model." Be prepared to justify the modification to your skeptical officemate. And to all you skeptical officemates out there, please grill your colleague about the changes. Mercilessly.

There is no path between HomeEnv and GPA or HomeEnv and YrsEd, $\gamma_{12} = \gamma_{32} = 0$. These three variables also correspond to the largest residual correlations. A positive HomeEnv is such a fundamental demographic variable that it seems reasonable to assume $HomeEnv \longrightarrow GPA$ and $HomeEnv \longrightarrow YrsEd$ rather than the other direction. The original model has $df = 3$ so we can afford to add two parameters and after the modification still have $df = 1$.

The modified model with its two new paths is Model 2. The residual correlations, $T_{ml} = 0.104$ and $RMSEA = 0$ are now all fine. Estimates are in Table 2. Estimates of the new paths are both significant $\hat{\gamma}_{12} = 0.25$, $se(\hat{\gamma}_{12}) = 0.06$, $\hat{\gamma}_{32} = 0.24$, $se(\hat{\gamma}_{32}) = 0.06$

Table 2. Parameter estimates, residual correlations and measures of fit for Model 2.

$\hat{\mathbf{B}}$

	y_1	y_2	y_3
y_1 GPA	0	0	0
y_2 Friends	.05	0	0
y_3 YrsEd	.18	.45	0

$\hat{\mathbf{\Gamma}}$

	x_1	x_2
y_1	.52	$\boxed{.25}$
y_2	.38	.23
y_3	0	$\boxed{.24}$

Residual correlationss, $\mathbf{R}_s = \mathbf{R} - \hat{\mathbf{P}}$

	x_1	x_2	y_1	y_2	y_3
x_1 Ability	00				
x_2 HomeEnv	00	00			
y_1 GPA	00	00	00		
y_2 Friends	00	00	00	00	
y_3 YrsEd	01	00	00	00	00

$\hat{\mathbf{\Psi}}$

	y_1	y_2	y_3
y_1 GPA	.59		
y_2 Friends	0	.71	
y_3 YrsEd	0	0	.52

$\hat{\mathbf{\Phi}}$

	x_1	x_2
x_1	1.0	
x_2	.29	1.0

$T_{ml} = 0.104$, $df = 1$, $p < 0.74$
$RMSEA = 0.0$

A Dangerous Button

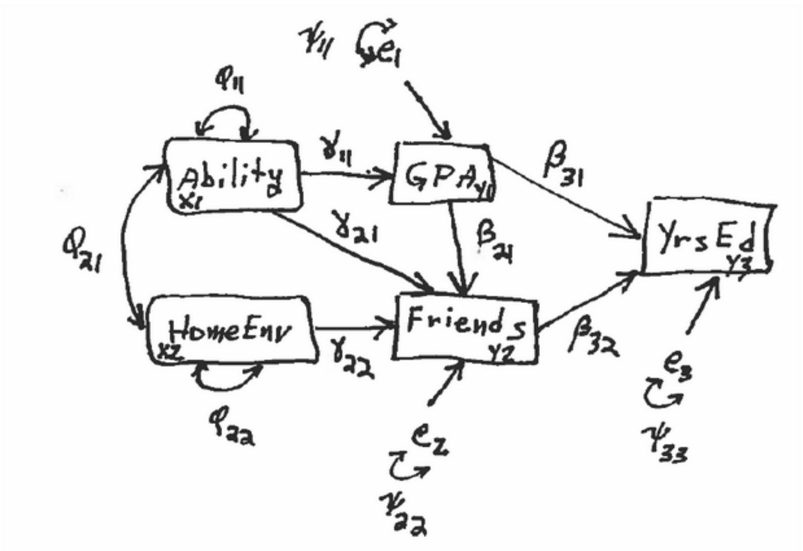
The world's most dangerous button is any button connected to a bomb. After that, another dangerous button is the button connected to an SEM computer program labeled Modification Indices. In SAS the keyword "MOD" on the CALIS line pushes the button. With LISREL the keyword "MI" does it. AMOS actually has the button of death as a handy option. All SEM computer programs have one.

The numerical results you get after pushing the button varies with the software but the information is similar. For every fixed element of every matrix used by the model, a statistic called the modification index is calculated. The modification index is the expected drop in T_{ml} if the fixed element were set free. SEM programs also display an estimate of what the newly freed parameter will be. In a large problem there are gobs of these statistics.

To illustrate, go back to Model 1 above. The fixed elements are the zeros in $\mathbf{B}$, $\mathbf{\Gamma}$ and $\mathbf{\Psi}$. All elements of $\mathbf{\Phi}$ are already in use. SAS displays the largest modification indices and these are shown in Table 3. The parameters labeled "a" are impossible in the general version of path analysis we have been using, but they are valid in the SAS system. In Table 1, Model 1 gives $T_{ml} = 35.3$. The column labeled "Change to T_{ml}" shows what will happen to this statistic when the associated modification is added to the model.

Table 3. The largest modification statistics for Model 1 from SAS

	Path or	Change		Expected
Parameter	Covariance	to T_{ml}	Significance	Estimate
γ_{32}	HomeEnv → YrsEd	23.6	$< .0001$	0.28
γ_{12}	HomeEnv → GPA	17.5	$< .0001$	0.35
a	GPA → HomeEnv	17.5	$< .0001$	0.25
a	YrsEd → HomeEnv	14.6	$< .0001$	0.22
ψ_{32}	YrsEd $\frown$ Friends	5.48	0.02	-0.24
γ_{31}	Ability → YrsEd	2.49	0.12	-0.18
ψ_{31}	YrsEd $\frown$ GPA	0.06	0.80	-0.02



The first two lines of Table 3 show that the paths HomeEnv → YrsEd and HomeEnv → GPA will produce the largest improvements in fit. These are the same changes I suggested earlier, but they are indicated here based purely on the data. It can be helpful to see the changes suggested by the Modification Indices when they are relationships you did not anticipate or even imagine. By definition, EDA is exploratory in path analysis as in other statistical methods, and good science is all about investigating relationships in data that are ad hoc and serendipitous.

On the other hand, you don't really need for me to issue a warning about why the Modification Indices button is dangerous. The main danger occurs when the software indicates a change to the model, such as a new path, that produces a giant improvement in fit but which is really illogical given the scientific context. To take advantage of the improvement in fit, an investigator offers some lame justification. For example, in this problem YrsEd is considered the most dependent of the dependent variables, the ultimate effect of all the antecedent causes. Consequently in Table 3 the suggested path that goes "backward" from YrsEd to HomeEnv, with an expected change of $T_{ml} = 14.6$, is preposterous.

Karl Jöreskog knows more about SEM than anyone. In the world.

> SEM works best when it is applied to designed studies based on a definite theory and with a clear objective. The initial model of the investigator need not be correct or best for the data. SEM has often been applied to more exploratory situations in which the initial model is set up more or less arbitrarily and then successively modified, perhaps numerous times, so as to improve the parsimony and fit of the model. This process has been termed a specification search. The goal of the search procedure is to find a model which fits the data well and in which all parameters have real significance and substantive meaning. After such an exploratory search it is important that the final model is cross validated on a different data set.
>
> A typical step in the specification search procedure involves the examination and assessment of fit of the current model, in particular, the t-values of estimated parameters and the modification indices for fixed parameters. The modification of the model may involve (a) elimination of parameters with small t-values, or (b) adding parameters with a large modification index. These steps must be taken with very careful judgment.

It is best to change only one parameter in each step but it does not have to be the one with the largest modification index. If it makes more sense from a scientific point of view to free a parameter with a smaller modification index then this should be done. Eliminating a parameter on the basis of its t-value may also be ill advised especially in a small sample. Even nonsignificant parameters may be of practical importance. If the scientific theory suggests that a particular parameter should be included in the model, it is probably better to retain it even though it is nonsignificant, which may occur because sample size is too small to detect its real significance.

In exploratory situations with many variables and weak or nonexisting substantive theory, SEM is probably not a useful tool. (Jöreskog & Sörbom, 1994. *LISREL 8 Users' Reference Guide*, pp. 274-275.)