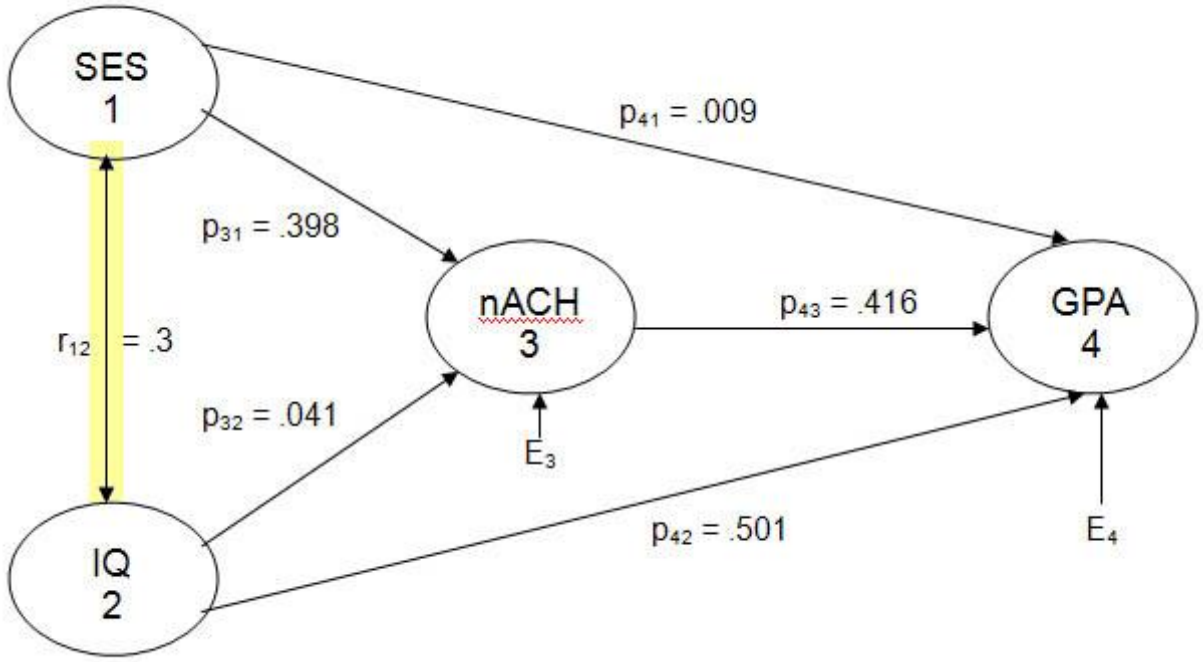
# An Introduction to Path Analysis©

Developed by Sewall Wright, path analysis is a method employed to determine whether or not a multivariate set of nonexperimental data fits well with a particular (a priori) causal model. Elazar J. Pedhazur (*Multiple Regression in Behavioral Research*, 2nd edition, Holt, Rinehard and Winston, 1982) has a nice introductory chapter on path analysis which is recommended reading for anyone who intends to use path analysis. This lecture draws heavily upon the material in Pedhazur's book.

Consider the path diagram presented in Figure 1.

**Figure 1**




Each oval represents a variable. We have data on each variable for each subject. In this diagram SES and IQ are considered to be **exogenous variables** -- their variance is assumed to be caused entirely by variables not in the causal model. The connecting line with arrows at both ends indicates that the correlation between these two variables will remain unanalyzed because we choose not to identify one variable as a cause of the other variable. Any correlation between these variables may actually be casual (1 causing 2 and/or 2 causing 1) and/or may be due to 1 and 2 sharing common causes. For example, having a certain set of genes may cause one to have the physical appearance that is necessary to obtain high SES in a particular culture and may independently also cause one to have a high IQ, creating a spurious correlation

between 1 and 2 that is totally due to their sharing a common cause, with no causal relationship between 1 and 2. Alternatively, some genes may cause only the physical appearance necessary to obtain high SES and high SES may cause high IQ (more money allows you to eat well, be healthy, afford good schools, etc., which raises your IQ). Alternatively, the genes may cause only elevated IQ, and high IQ causes one to socially advance to high SES. In this model we have chosen not to decide among these alternatives.

GPA and nAch are **endogenous variables** in this model -- their variance is considered to be explained in part by other variables in the model. Paths drawn to endogenous variables are directional (arrowhead on one end only). Variance in GPA is theorized to result from variance in SES, IQ, nAch, and extraneous (not in the model) sources. The influence of these extraneous variables is indicated by the arrow from $E_Y$. Variance in nAch is theorized to be caused by variance in SES, IQ, and extraneous sources.

Please note that the path to an endogenous variable must be unidirectional in path analysis. Were we to decide that not only does high SES cause high nAch but that also high nAch causes high SES, we could not use path analysis.

For each path to an endogenous variable we shall compute a path coefficient, $p_{ij}$, where "i" indicates the effect and "j" the cause. If we square a path coefficient we get the proportion of the affected variable's variance that is caused by the causal variable. The coefficient may be positive (increasing the causal variable causes increases in the dependent variable if all other causal variables are held constant) or negative (increasing causal variable decreases dependent variable).

A path analysis can be conducted as **a hierarchical (sequential) multiple regression analysis**. For each endogenous variable we shall conduct a multiple regression analysis predicting that variable (Y) from all other variables which are hypothesized to have direct effects on Y. We do not include in this multiple regression any variables which are hypothesized to affect Y only indirectly (through one or more intervening variables). The beta weights from these multiple regressions are the path coefficients shown in the typical figures that are used to display the results of a path analysis.

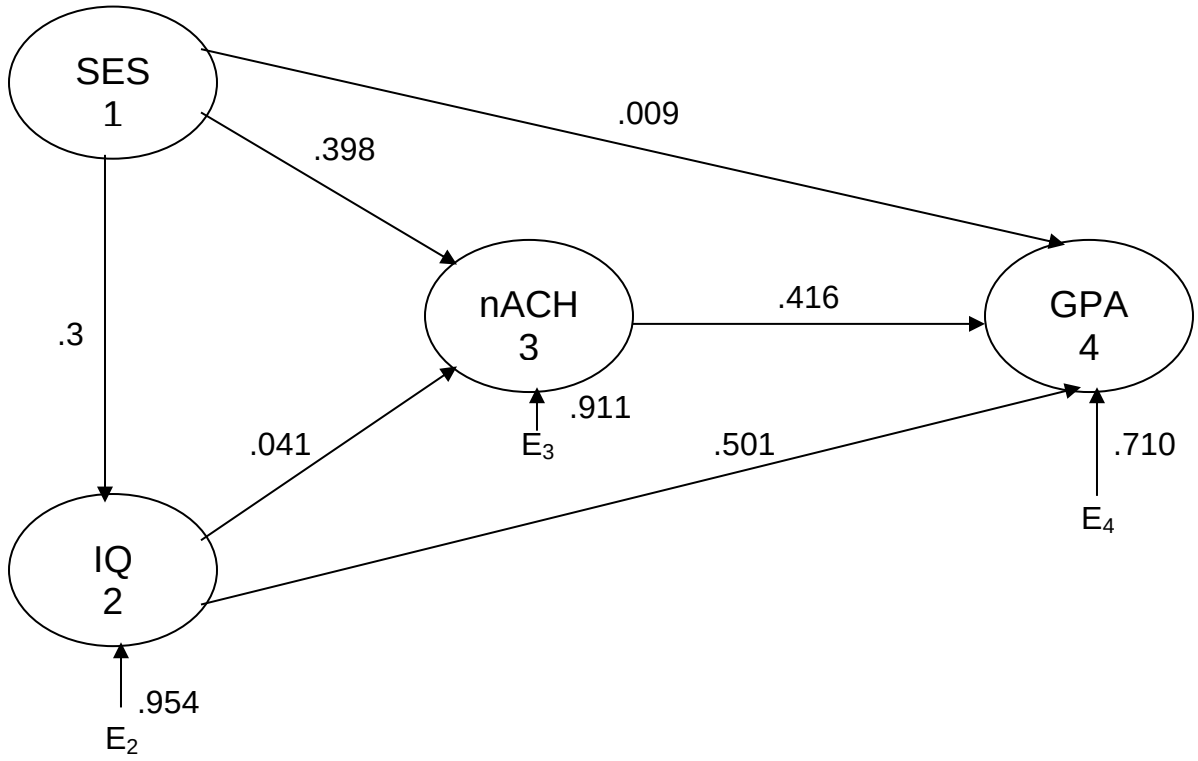
Consider these data from Pedhazur:

| | IQ | nAch | GPA |
|---|---|---|---|
| SES | .300 | .410 | .330 |
| IQ | | .160 | .570 |
| nAch | | | .500 |

For our analysis, let us make one change in Figure 1: Make IQ an endogenous variable, with SES a cause of variance in IQ (make unidirectional arrow from SES to

IQ). Our revised model is illustrated in Figure 1A, to which I have added the path coefficients computed below.

**Figure 1A**




Obtain and run Path-1.sas from my SAS Programs page. Here is the code that produced the coefficients for the model in the figure above:

```
PROC REG;
 Figure_1_GPA: MODEL GPA = SES IQ NACH;
 Figure_1_nACH: MODEL NACH = SES IQ;
```

**Parameter Estimates for Predicting GPA**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **SES** | **1** | **0.00919** | 0.11881 | 0.08 | 0.9387 |
| **IQ** | **1** | **0.50066** | 0.10978 | 4.56 | <.0001 |
| **NACH** | **1** | **0.41613** | 0.11481 | 3.62 | 0.0007 |

Our diagram indicates that GPA is directly affected by SES, IQ, and nAch. We regress GPA on these three causal variables and obtain $R^2_{4.123}$ = .49647, $\beta_{41.23} = p_{41}$ = .009, $\beta_{42.13} = p_{42}$ = .501, and $\beta_{43.12} = p_{43}$ = .416.

| R-Square | 0.4965 |
|---|---|

The path coefficient from extraneous variables is $\sqrt{1-R^2_{4.123}} = \sqrt{1-.49647} = .710$.

We see that GPA is directly affected by IQ, nAch, and extraneous variables much more than by SES, but we must not forget that SES also has indirect effects (through IQ & nAch) upon GPA. We shall separate direct from indirect effects later.

**Parameter Estimates for Predicting nAch**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **SES** | **1** | **0.39780** | 0.13934 | 2.85 | 0.0064 |
| **IQ** | **1** | **0.04066** | 0.13934 | 0.29 | 0.7717 |

Achievement motivation is affected by both SES and IQ in our model, and these causes are correlated with one another. We regress nAch on these two causal variables and obtain $R^2_{3.12}$ = .1696, $\beta_{31.2} = p_{31}$ = .398, and $\beta_{32.1} = p_{32}$ = .041.

| R-Square | 0.1696 |
|---|---|

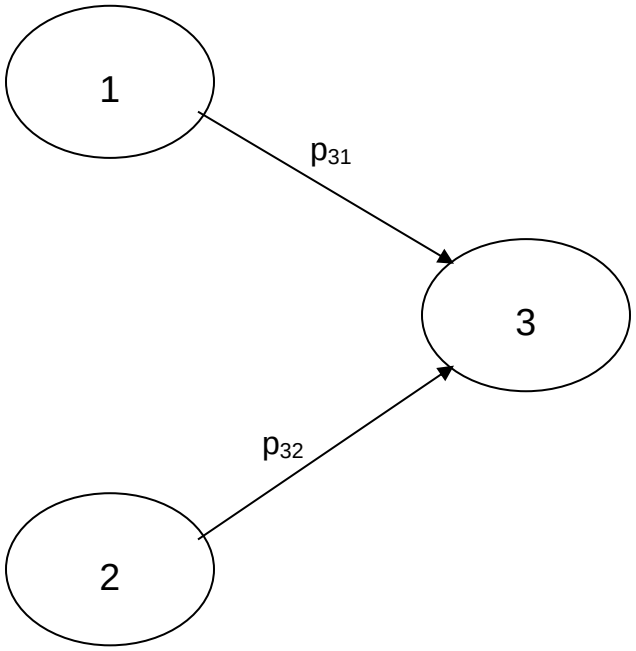
The path coefficient from $E_3$ to nAch is $\sqrt{1-R^2_{3.12}} = \sqrt{1-.1696} = .911$.

We see that nAch is more strongly caused by SES than by IQ, and that extraneous variables exert great influence.

Now consider the path to IQ from SES. Since there is only one predictor variable in this model, the path coefficient is the simple (zero-order) *r* between IQ and SES, which is .300. This would also be the case if the Y variable were theorized to be affected by two independent causes (see Figure 2, in which our model theorizes that the correlation between 1 and 2 equals 0).

The path coefficient from extraneous variables to IQ is the residual of the SES-IQ correlation, $\sqrt{1-r^2_{2.1}} = \sqrt{1-.09} = .954$.

**Figure 2**




$r_{12} = 0$

$p_{31} = r_{31}$

$p_{32} = r_{32}$

Note that the program contains the correlation matrix from Pedhazur. I decided to use an *N* of 50, but did not enter means and standard deviations for the variables, so the parameter estimates that SAS produces are standardized (the slope is a beta).

## Decomposing Correlations

The correlation between two variables may be decomposed into four components:

1. the **direct effect** of X on Y,
2. the **indirect effect** of X (through an intervening variable) on Y,
3. **an unanalyzed component** due to our not knowing the direction of causation for a path, and
4. a **spurious component** due to X and Y each being caused by some third variable or set of variables in the model.

Consider first the correlations among the variables in **Figure 1**.

**The correlation between SES and IQ, $r_{12}$**, will be **unanalyzed** because of the bi-directional path between the two variables.

**The correlation between SES and nAch, $r_{13}$** = .410 is decomposed into:

- **$p_{31}$, a direct effect**, SES to nAch, which we already computed to be .398, and
- **$p_{32}r_{12}$, an unanalyzed component**, SES to IQ to nAch, whose size = .041(.3) = .012. -- SES could indirectly affect nAch if SES causes changes in IQ which in turn cause changes in nAch, but we do not know the nature of the causal relationship between SES and IQ, so this component must remain unanalyzed.
- When we sum these two components, .398 + .012, we get the value of the original correlation, .410.

**The correlation between IQ and nAch, $r_{23}$** = .16, is decomposed into:

- **$p_{32}$**, the **direct** effect, = .041 and
- **$p_{31}r_{12}$**, an **unanalyzed** component, IQ to SES to nAch, = .398(.3) = .119.
- Summing .041 and .119 gives the original correlation, .16.

**The SES - GPA correlation, $r_{14}$** =.33 is decomposed into:

- **$p_{41}$**, the **direct** effect, = .009.
- **$p_{43}p_{31}$**, the **indirect** effect of SES through nAch to GPA, = .416(.398) = .166.
- **$p_{42}r_{12}$**, SES to IQ to GPA, is **unanalyzed**, = .501(.3) = .150.
- **$p_{43}p_{32}r_{12}$**, SES to IQ to nAch to GPA, is **unanalyzed**, = .416(.041)(.3) = .005.
- When we sum .009, .166, ,150, and .155, we get the original correlation, .33.
- The **total effect** (or **effect coefficient**) of X on Y equals the sum of X's direct and indirect effects on Y -- that is, .009 + .166 = .175.

**The IQ - GPA correlation, $r_{24}$**, =.57 is decomposed into:

- **$p_{42}$**, a **direct** effect, = .501.
- **$p_{43}p_{32}$**, an **indirect** effect through nAch to GPA, = .416(.041) = .017.
- **$p_{41}r_{12}$**, **unanalyzed**, IQ to SES to GPA, .009(.3) = .003
- **$p_{43}p_{31}r_{12}$, unanalyzed**, IQ to SES to nAch to GPA, = .416(.398)(.3) = .050.
- The original correlation = .501 + .017 + .003 .050 = .57.

**The nAch - GPA correlation, $r_{34}$** = .50, is decomposed into:

- **$p_{43}$,** the **direct** effect, = .416
- and a **spurious component** due to nAch and GPA sharing common causes SES and IQ
  - **$p_{41}p_{31}$,** nAch to SES to GPA, = (.009)(.398).
  - **$p_{41}r_{12}p_{32}$**, nAch to IQ to SES to GPA, = (.009)(.3)(.041).
  - **$p_{42}p_{32}$**, nAch to IQ to GPA, = (.501)(.041).

- $\boldsymbol{p_{42}r_{12}p_{31}}$, nAch to SES to IQ to GPA, = (.501)(.3)(.398).
- These spurious components sum to .084. Note that in this decomposition elements involving $r_{12}$ were classified spurious rather than unanalyzed because variables 1 and 2 are common (even though correlated) causes of variables 3 and 4.

Here is a summary of the decomposition of correlations from Figure 1:

$r_{12}$ = unanalyzed

$r_{13} = p_{31} + p_{32}r_{12}$

DE U

$r_{23} = p_{32} + p_{31}r_{12}$

DE U

$r_{14} = p_{41} + p_{43}p_{31} + (p_{43}p_{32}r_{12} + p_{42}r_{12})$

DE IE U

$r_{24} = p_{42} + p_{43}p_{32} + (p_{41}r_{12} + p_{43}p_{31}r_{12})$

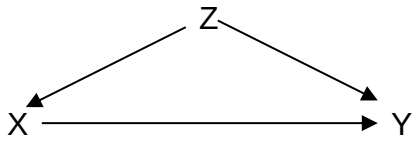
DE IE U

$r_{34} = p_{43} + (p_{41}p_{31} + p_{41}r_{12}p_{32} + p_{42}p_{32} + p_{42}r_{12}p_{31})$

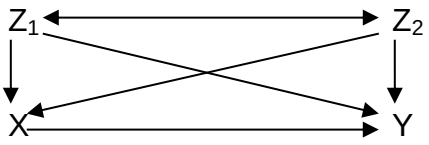
DE S

Are you sufficiently confused yet? I get confused doing these decompositions too. Here is a relatively simple set of instructions to help you decide whether a path is direct, indirect, spurious, or unanalyzed: Put your finger on the affected variable and trace back to the causal variable. Now,

- If you cross only one arrowhead, head first, you have a direct effect.
- If you cross two or more arrowheads, each head first, you have an indirect effect.
- If you cross a path that has arrowheads on both ends, the effect is unanalyzed (or possibly spurious)
- Only cross a path not head first when you are evaluating a spurious effect -- that is, where a pair of variables is affected by a common third variable or set of variables. For example, some of the correlation between X and Y below is due to the common cause Z.




- An effect that includes a bidirectional path can be considered spurious rather than unanalyzed if both of the variables in the bidirectional path are causes of both of the variables in the correlation being decomposed, as illustrated below:



Next, consider **Figure 1A**, with SES a cause of variance in IQ.

The $r_{12}$ is now $p_{21}$, the **direct** effect of SES on IQ.

**The correlation between SES and nAch, $r_{13}$** = .410 is decomposed into:

- $p_{31}$**, a direct effect**, SES to nAch, .398, and
- $p_{32}p_{21}$**, an indirect effect**, SES to IQ to nAch, whose size = .041(.3) = .012. -- Note that this indirect effect was an unanalyzed component of $r_{13}$ in the previous model.
- The **total effect** (or **effect coefficient**) of X on Y equals the sum of X's direct and indirect effects on Y. For SES to nAch, the effect coefficient = .398 + .012 = .410 = $r_{13}$. Note that making SES a cause of IQ in our model only slightly increased the effect coefficient for SES on IQ (by .012).

**The correlation between IQ and nAch, $r_{23}$** = .16, is decomposed into:

- $p_{32}$, the **direct** effect, = .041 and
- $p_{31}p_{21}$, a **spurious** component, IQ to SES to nAch, = .398(.3) = .119. Both nAch and IQ are caused by SES, so part of the $r_{23}$ must be spurious, due to that shared common cause rather than to any effect of IQ upon nAch. This component was unanalyzed in the previous model.

**The SES - GPA correlation, $r_{14}$** =.33 is decomposed into:

- $p_{41}$, the **direct** effect, = .009.
- $p_{43}p_{31}$, the **indirect** effect of SES through nAch to GPA, = .416(.398) = .166.
- $p_{42}p_{21}$, the **indirect** effect of SES to IQ to GPA, .501(.3) = .150.
- $p_{43}p_{32}p_{21}$, the **indirect** effect of SES to IQ to nAch to GPA, = .416(.041)(.3) = .005.
- The indirect effects of SES on GPA total to .321. The total effect of SES on GPA = .009 + .321 = .330 = $r_{14}$. Note that the indirect and total effects of SES upon GPA are greater in this model than in the previous model. Considering SES a cause of variance in IQ moved what otherwise would be SES' unanalyzed effects into its indirect effects.

**The IQ - GPA correlation, $r_{24}$**, =.57 is decomposed into:

- $p_{42}$, a **direct** effect, = .501.
- $p_{43}p_{32}$, an **indirect** effect through nAch to GPA, = .416(.041) = .017.
- $p_{41}p_{21}$, **spurious**, IQ to SES to GPA, .009(.3) = .003 (IQ and GPA share the common cause SES).

- $p_{43}p_{31}p_{12}$, **spurious**, IQ to SES to nAch to GPA, .416(.398)(.3) = .050 (the common cause also affects GPA through nAch).
- The total effect of IQ on GPA = DE + IE = .501 + .017 = .518 = $r_{24}$ less the spurious component.

**The nAch - GPA correlation, $r_{34}$** = .50, is decomposed in exactly the same way it was in the earlier model.

Here is a summary of the decompositions for the correlations in Figure 1A:

$r_{12} = p_{21}$ $\qquad$ $r_{13} = p_{31} + p_{32}p_{21}$ $\qquad$ $r_{23} = p_{32} + p_{31}p_{21}$

DE IE $\qquad$ DE S

$r_{14} = p_{41} + (p_{43}p_{31} + p_{43}p_{32}p_{21} + p_{42}p_{21})$ $\qquad$ $r_{24} = p_{42} + p_{43}p_{32} + (p_{41}p_{21} + p_{43}p_{31}p_{21})$

DE IE $\qquad$ DE IE S

$r_{34} = p_{43} + (p_{41}p_{31} + p_{41}p_{21}p_{32} + p_{42}p_{32} + p_{42}p_{21}p_{31})$

DE S

## Overidentified Models

Consider a simple three variable model where $r_{12} = r_{23} = .50$ and $r_{13} = .25$. In Figure 3A and Figure 3B are two different "**just-identified**" models of the causal relationships among these three variables. In a just-identified model (aka "saturated model" there is a direct path (not through an intervening variable) from each variable to each other variable. In such a model the fit between the data and the model will always be perfect. Note that the decomposed correlations for both models can be used to "reproduce" the original correlations perfectly, even though the two models present quite different pictures of the casual relationships among the variables.

**Figure 3A**

**Model A:** $r_{12} = r_{23} = .50$ and $r_{13} = .25$

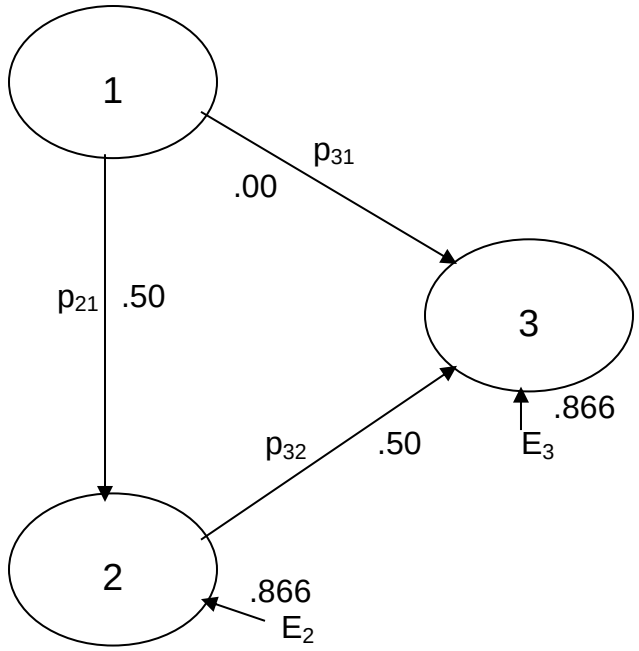
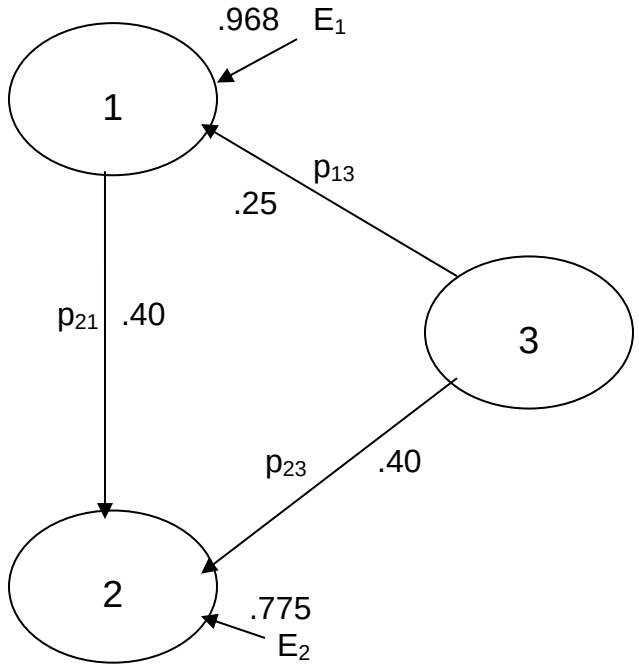



**Figure 3B**

**Model B:** $r_{12} = r_{23} = .50$ and $r_{13} = .25$



| | Model A | Model B |
|---|---|---|
| $r_{12}$ = .50 | $p_{21}$ = .50 | $p_{21} + p_{23}p_{13}$ |
| | .50 | .40 + .10 |
| | DE | DE + spurious component |
| $r_{13}$ = .25 | $p_{31} + p_{32}p_{31}$ | $p_{31}$ |
| | 0.00 + .25 | .25 |
| | DE + IE | DE |
| $r_{23}$ = .50 | $p_{32} + p_{31}p_{21}$ | $p_{23} + p_{21}p_{13}$ |
| | .50 + 0.00 | 40 + .10 |
| | DE + spurious | DE + IE |

In Figure 4A and Figure 4B are two "**overidentified**" models, models in which at least one pair of variables are not connected to each other by direct paths. In Model A it is hypothesized that 1 causes 2, 2 causes 3, and there is no direct effect of 1 on 3. In B it is hypothesized that 2 causes both 1 and 3 with no nonspurious effect between 1 and 3. Both models attempt to explain the $r_{13}$ in the absence of any direct effect of 1 on 3.

**Figure 4A**

**Model A:** $r_{12} = r_{23} = .50$ and $r_{13} = .25$

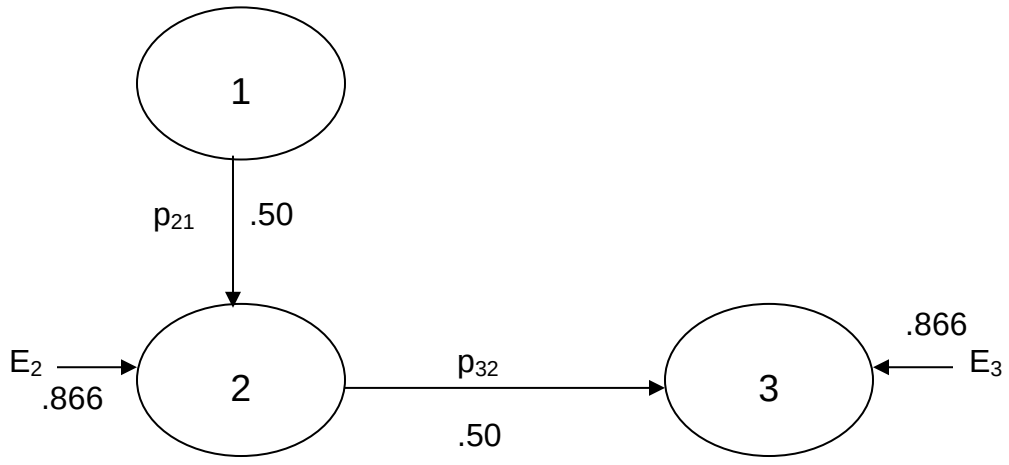
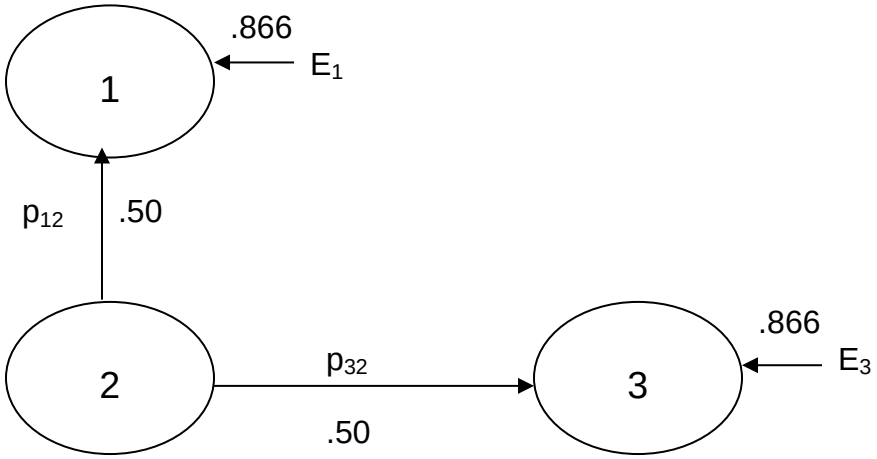


**Figure 4B**

**Model B:** $r_{12} = r_{23} = .50$ and $r_{13} = .25$



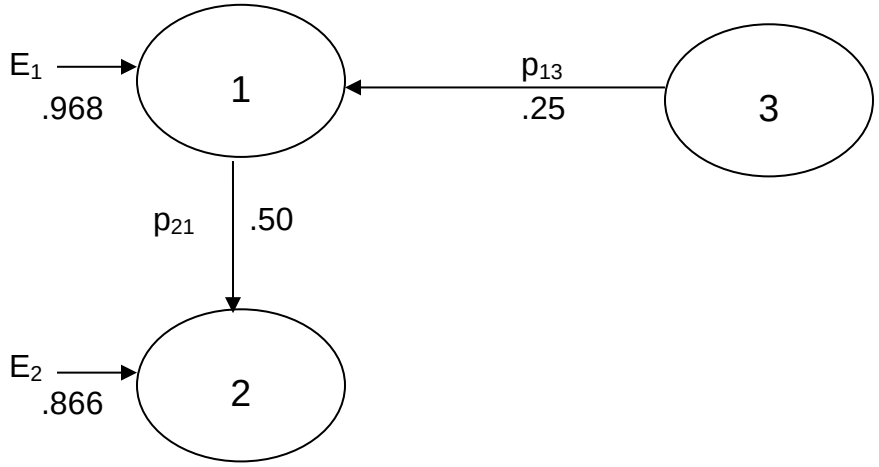

- In Model A, $r_{13} = p_{32}p_{21} = .5(.5) = .25$ = IE of 1 through 2 on 3.
- In Model B, $r_{13} = p_{32}p_{21} = .5(.5) = .25$ = spurious correlation between 1 and 3 due to their sharing 2 as a common cause.

One may attempt to determine how well an overidentified path model fits the data by checking how well the original correlation matrix can be reproduced using the path coefficients. Any just-identified model will reproduce the correlation matrix perfectly. However, two radically different overidentified models may reproduce the correlation matrix equally well. For our overidentified models A and B, $r_{13} = p_{32}p_{21}$ in A = .25 = $p_{32}p_{21}$ in B. That is, A and B fit the data equally well (perfectly). Being able to reproduce the correlation matrix should lend support to your model in the sense that you have attempted to refute it but could not, but it does not in any way "prove" your model to be correct -- other models may pass the test just as well or even better!

Consider another overidentified model, based on the same data used for all models in Figures 3 and 4. As shown in Figure 5, this model supposes that 3 affects 1 directly and 2 only indirectly. That is, $r_{23}$ is due only to an indirect effect of 3 on 2. The $r_{23}$ here decomposes to $p_{21}p_{13} = (.50)(.25) = .125$, the IE of 3 through 1 on 2 -- but the original $r_{23} = .50$. This model does such a lousy job of reproducing $r_{23}$ that we conclude that it is not supported by the data.

**Figure 5**

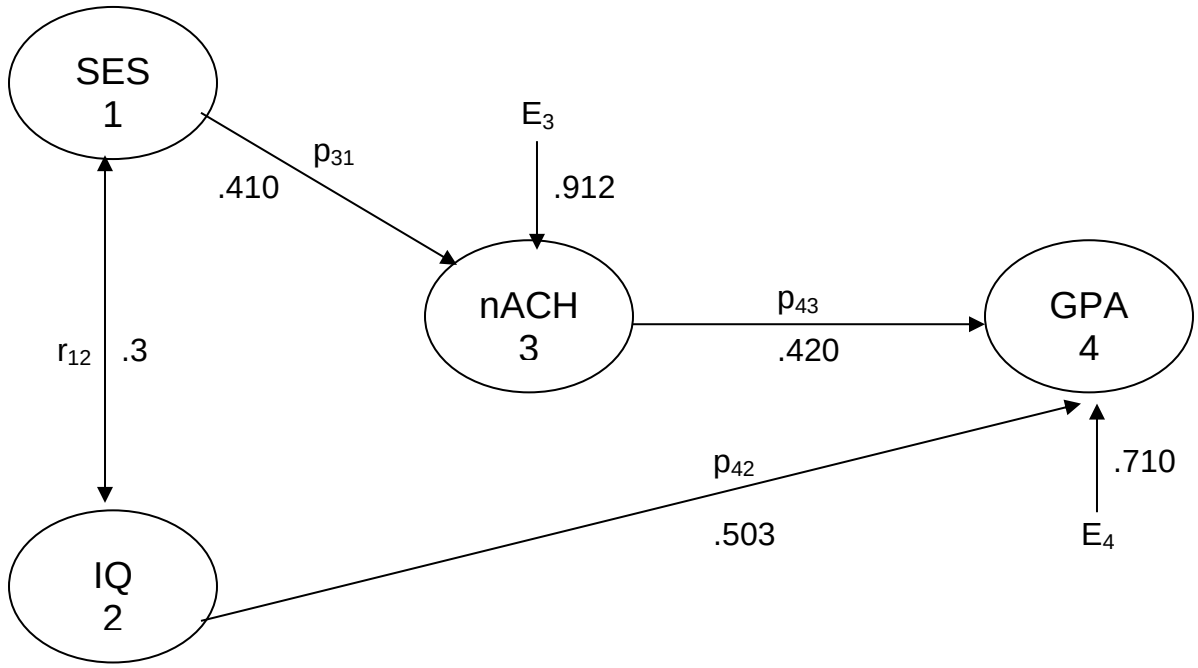
$r_{12} = r_{23} = .50$ and $r_{13} = .25$




Let us now consider an overidentified model based on Figure 1. As shown in Figure 6, we hypothesize no direct effects of SES on GPA or of IQ on nAch, and we choose not to directionalize the SES - IQ relationship. Since nAch is now caused only by SES and $E_3$, $p_{31} = r_{13}$. Since GPA is now directly affected only by nAch and IQ (and $E_4$), we regress GPA on nAch and IQ and obtain $\beta_{42.3} = p_{42} = .503$ and $\beta_{43.2} = p_{43} = .420$ **Path-1.sas** includes computation of the path coefficients for the model in Figure 6:

```
Figure_6: MODEL GPA = IQ NACH;
```

**Figure 6**



Now, to see how well this overidentified model fits the data, we reproduce the original correlations (*rr* = reproduced correlation coefficient, *r* = original coefficient)

$rr_{12} = r_{12} = .3$ $rr_{13} = p_{31} = .41 = r_{13}$

U DE

$rr_{14} = p_{43}p_{31} + p_{42}r_{12} = .172 + .151 = .\mathbf{323}$ $r_{14} = \mathbf{.330}$

IE U

$rr_{23} = p_{31}r_{12} = .\mathbf{123}$ $r_{23} = \mathbf{.160}$

U

$rr_{24} = p_{42} + p_{43}p_{31}r_{12} = .503 + .052 = .\mathbf{555}$ $r_{24} = \mathbf{.570}$

DE U

$rr_{34} = p_{43} + p_{42}r_{12}p_{31} = .420 + .062 = .\mathbf{482}$ $r_{34} = \mathbf{.500}$

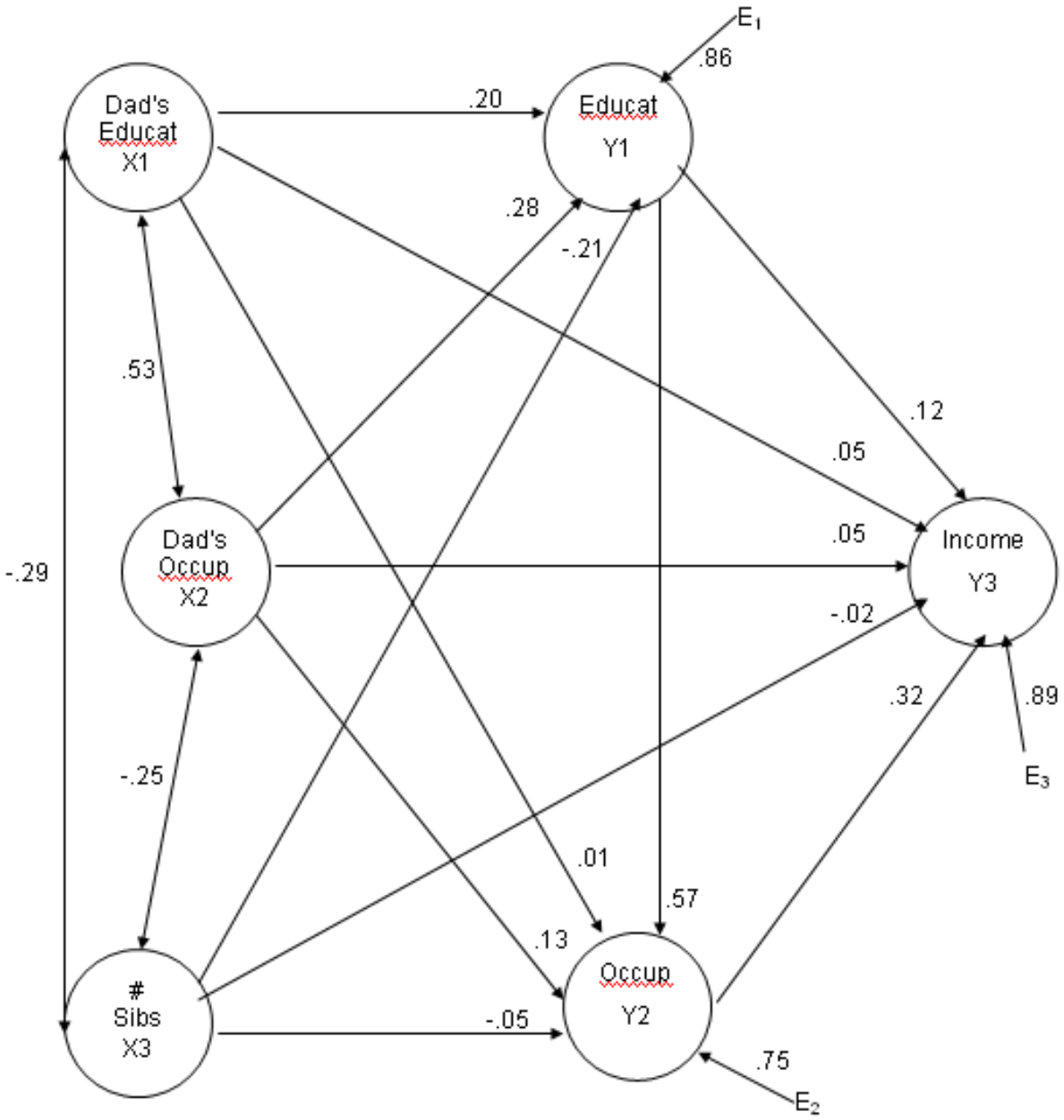
DE S

Note that there are relatively small discrepancies between the reproduced correlations and the original correlations, indicating that the model fits the data well.

See output from PROC CALIS showing analysis of the saturated model and the overidentified model.

Path analysis models are frequently much more complex than those we have covered so far. Such a complex model is presented in Figure 7. The X's are Fathers' education, Fathers' occupation, and number of siblings. The Y's are subjects' education, occupation, and income. Subjects were "non-Negro men, 35-44 age group." Path coefficients to $Y_1$ were obtained by regressing $Y_1$ on $X_1$, $X_2$, and $X_3$; to $Y_2$ by regressing $Y_2$ on $Y_1$, $X_1$, $X_2$, and $X_3$; and to $Y_3$ by regressing $Y_3$ on $Y_1$, $Y_2$, $X_1$, $X_2$, and $X_3$. Run **Path-1.sas** to see the computation of the path coefficients for the models in Figures 1, 6, and 7.

**Figure 7**




## Trimming Models

Starting with a just-identified model, one may want to do some "**theory trimming**" -- that is, evaluate whether or not dropping one or more of the paths would substantially reduce the fit between the model and the data. For example, consider the model in Figure 7. We regressed $Y_1$ on $X_1$, $X_2$, and $X_3$ to obtain the path coefficients

(Beta weights) to $Y_1$. If we wished to use statistical significance as our criterion for retaining a coefficient, we could delete any path for which our multiple regression analysis indicated the Beta was not significantly different from zero. One problem with this approach is that with large sample sizes even trivially small coefficients will be "statistically significant" and thus retained. One may want to include a "meaningfulness" criterion and/or a minimum absolute value of Beta for retention. For example, if $|\beta| < .05$, delete the path; if $.05 < |\beta| < .10$ and the path "doesn't make (theoretical) sense," delete it. In Figure 7 all paths to Education are large enough to pass such a test.

Consider Occupation regressed on Education, Dad's Education, Dad's Occupation, and Sibs. The Beta for Dad's Education is clearly small enough to delete. We might also decide to delete the path from Sibs if it does not seem sensible to us that one's occupation be directly affected by the number of sibs one has. Of course, deleting one predictor will change the Beta weights of remaining predictors, so one may want to delete one variable at a time, evaluating all remaining predictors at each step. If you are eliminating two or more predictors at one stage (for example, Dad's Education and Sibs to Occupation) and wish to test whether simultaneously eliminating them significantly affects the model at that stage, use a partial *F* test.

For Income regressed on all other variables, it appears that the direct paths from Dad's Education, Dad's Occupation, and Sibs could all be eliminated. Notice, however, that we have evaluated this model in three separate stages. It is possible that eliminating one path might not have a significant effect on the model at that stage, but nonetheless might have a significant effect on the entire model. Later I shall show you how to test the entire model.

Perhaps the most serious criticism of the sort of theory trimming described above is that it is applied a posteriori. Some argue that the data should not be allowed to tell one which hypotheses to test and which not to test. One should always feel more comfortable with a priori trimming, that is, trimming that is supported on theoretical grounds. To some extent, that is the function of the "meaningfulness" criterion I mentioned earlier. We might have a priori suspicions that a particular path is not important, but include it in our just identified model anyhow, later trimming it away if the data results in it receiving the low Beta we expected.

## Evaluating Trimmed Models

Once one has trimmed away some paths (a priori or a posteriori) from a just identified model, e is left with an overidentified model. Specht (On the evaluation of causal models, *Social Science Research*, 1975, *4*, 113-133) developed a goodness of fit statistic (*Q)* to measure how well a trimmed model fits the data (the reproduced correlation coefficients differ little from the observed correlation coefficients) and a test statistic (*W*) to test the null hypothesis that the trimmed model fits the data as well as does the just-identified model. If our trimmed model is not significantly different from the just-identified model, then we feel more comfortable with it. The flaw in this is that even a poorly fitting model will not differ significantly from the just-identified model if sample size (and thus power) is low, and a good fitting model will differ significantly from the just-identified model if sample size (and power) is large. Specht's Q seems to have

been replaced by more modern goodness of fit statistics. I have not seen it in modern statistical packages. If you would like to learn a bit about Specht's Q and W, see my document Specht's Goodness of Fit Estimate and Test for Path Analyses . See http://davidakenny.net/cm/fit.htm for a summary of some commonly used modern goodness of fit indices.

## PROC CALIS

These days path analyses are typically conducted using software that makes it a lot easier to do the analysis than what I have shown above using Proc Reg. In SAS, Proc Calis can be used to conduct the complete analysis, including computation of the effect coefficients, which ounce were tediously computed by matrix algebra. Please do look at the Proc Calis document linked above. I shall include part of the output here.

The model in Figure 1 is saturated -- that is, there is a direct path from each variable to each other variable. With a saturated model, the fit between data and model will be perfect. The observed correlations can be perfectly reconstructed from the decomposed correlation. The chi-square that tests the null that the fit is perfect will have a value of zero, indicating perfect fit. Here is the goodness of fit output:

| **Fit Summary** | | |
|---|---|---|
| **Modeling Info** | **N Observations** | 10000 |
| | **N Variables** | 4 |
| | **N Moments** | 10 |
| | **N Parameters** | 10 |
| | **N Active Constraints** | 0 |
| **Absolute Index** | **Chi-Square** | 0.0000 |
| | **Chi-Square DF** | 0 |
| | **Pr > Chi-Square** | . |
| | **Root Mean Square Residual (RMSR)** | 0.0000 |
| | **Goodness of Fit Index (GFI)** | 1.0000 |

Since this is a saturated model, the fit is perfect.

Now I remove from the model the direct path from IQ to nACH and from SES to GPA. The model is no longer saturated. Look at the goodness of fit statistics now:

| Fit Summary | | |
|---|---|---|
| **Modeling Info** | **N Observations** | 10000 |
| | **N Variables** | 4 |
| | **N Moments** | 10 |
| | **N Parameters** | 8 |
| | **N Active Constraints** | 0 |
| **Absolute Index** | **Chi-Square** | 19.3988 |
| | **Chi-Square DF** | 2 |
| | **Pr > Chi-Square** | <.0001 |
| | **Goodness of Fit Index (GFI)** | 0.9990 |
| | **RMSEA Estimate** | 0.0295 |

Although we reject the null hypothesis of perfect fit, this is meaningless, given the very large sample size. Notice that our fit statistics are still excellent.

Return to Wuensch's Stats Lessons Page

This document may be cited in this fashion:

Wuensch, K. L. (2016). *An introduction to path analysis*. Retrieved from http://core.ecu.edu/psyc/wuenschk/MV/SEM/Path.pdf on <provide date here>.