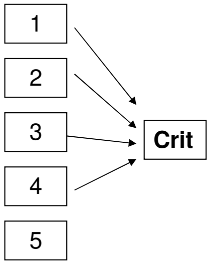
Introduction to Path Analysis

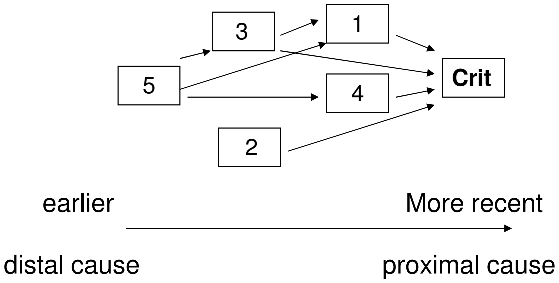
- Ways to “think about” path analysis
- Path coefficients
- A bit about direct and indirect effects
- What path analysis can and can’t do for you…
- Measured vs. manifested → the “when” of variables
- Some ways to improve a path analysis model
- About non-recursive cause in path models
- Mediation analyses
- Model Identification & Testing

One way to “think about” path analysis is as a way of “sorting out” the colinearity patterns amongst the predictors – asking yourself what may be the “structure” -- temporal &/or causal relationships -- among these predictors that produces the pattern of colinearity.

“Structure” of a MR model – with hypotheses about which predictors will contribute



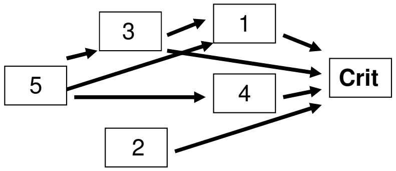
A proposed structure for the colinearity among the predictors and how they relate to the criterion – with hypotheses about which paths will contribute



Where do the path coefficients come from?

One way is to run a series of multiple regressions…

for each analysis: a variable with arrows pointing at it will be the criterion variable and each of the variables having arrows pointing to it will be the predictors



1. Crit = 3 Pred = 5
2. Crit = 1 Preds = 3 & 5
3. Crit = 4 Pred = 5
4. Crit = Crit Preds = 1, 2, 3 & 4

The path coefficients are the β weights from the respective regression analyses (remember that β = r for bivariate models)

What path analysis can and can't accomplish…

Cans -- for a given structural model you can…

- evaluate the contribution of any path or combination of paths to the overall fit of that structural model
- help identify sources of suppressor effects (indirect paths)

Can'ts

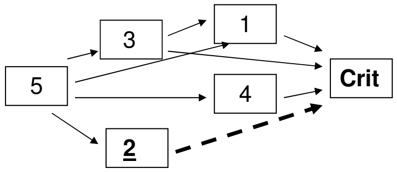
- non-recursive (bi-directional) models
- help decide among alternative structural models
- provide tests of causality (unless experimental data)

So… You have to convince yourself and your audience of the "reasonableness" of your structural model (the placing of the predictors), and then you can test hypotheses about which arrows amongst the variables have unique contributions.

Alternative ways to "think about" path analysis…

- to capture the "causal paths" among the predictors and to the criterion
- to capture the "temporal paths" among the predictors and to the criterion
- to distinguish "direct" and "indirect" paths of relationship
- to investigate "mediation effects"

… to distinguish "direct" and "indirect" paths of relationship…

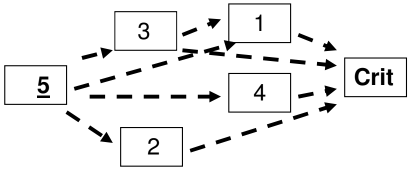


2 has a *direct effect* on Crit

- a "contributor" in both the regression and the path models

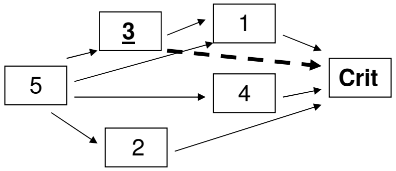
5 does not have a direct effect on Crit – but does have multiple *indirect effects*

- not "contributing" in the regression model could mistakenly lead us to conclude "5 doesn't matter in understanding Crit"



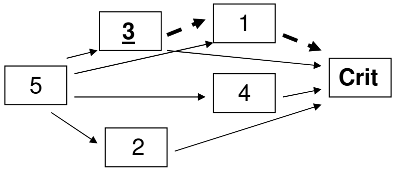
…to distinguish "direct" and "indirect" paths of relationship…, cont.

3 has a *direct* effect on Crit



3 also has an *indirect* effect on Crit

- there's more to the 3 → Crit relationship than was captured in the regression model



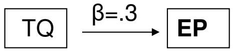
… to investigate "mediation effects"…

Mediation effects and analyses highlight the difference between bivariate and multivariate relationships between a variable and a criterion (collinearity & suppressor effects).

For example…

For Teaching Quality & Exam Performance → r = .30, p = .01

- for binary regression β = r,
  so we have the path model…

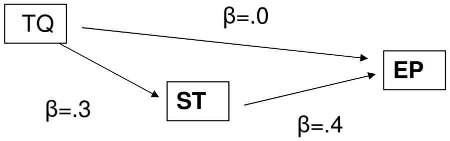


It occurs to one of the researchers that there just might be something else besides Teaching Quality related to (influencing, even) Exam Performance.

- The researcher decides that Study Time (ST) might be such a variable.
- Thinking temporally/causally, the researcher considers that Study Time "comes in between" Teaching and Testing.
- So the researcher builds a mediation model, getting the weights from a multiple regression with TQ and ST as predictors of EP

… to investigate "mediation effects"…

The resulting model looks like …



We might describe model as, "The apparent effect of Teaching Quality on Exam Performance (r=.30) is mediated by Study Time."

We might describe the combination of the bivariate analysis and the multiple regression from which the path coefficients were obtained as, "While Teaching Quality has a bivariate relationship with Exam Performance (r=.30), it does not contribute to a multiple regression model (β=.0) that also includes Study Time (β=.40).

Either analysis reminds us that the bivariate contribution of a given predictor might not "hold up" when we look at that relationship within a multivariate model!

Notice that TQ is "still important" because it seems to have something to do with study time – an indirect effect upon Exam Performance.

The “when” of variables and their place in the model …

When a variable is “measured” → when we collect the data:
- usually concurrent
- often postdictive (can be a problem – memory biases, etc.)
- sometimes predictive (hypothetical – can really be a problem)

When a variable is “manifested” → when the value of the variable came into being
- when it “comes into being for that participant”
- may or may not be before the measure was taken

E.g., State vs. Trait anxiety

- trait anxiety is intended to be “characterological,” “long term” and “context free” → earlier in model

- state anxiety is intended to be “short term” & “contextual” → depends when it was measured

Some caveats about the “when” of Path & Mediation Analyses…

1. The “Causal Ordering” must be theoretically supported → path analysis can’t “sort out” alternative arrangements -- it can only decide what paths of a specific arrangement can be dropped

2. Mediating variables must come after what they are mediating

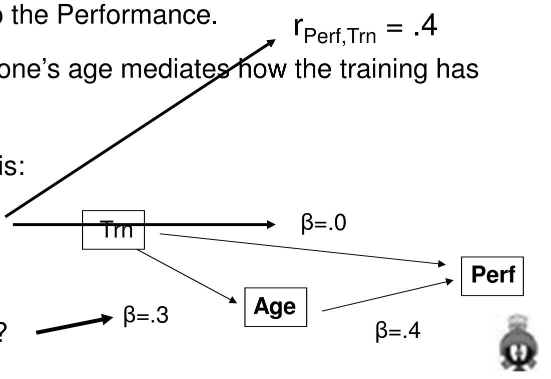
E.g. The Training is related to the Performance.

But the researcher thinks that one’s age mediates how the training has its effect…

So we run a mediation analysis:

Looks like a participant’s age mediates the training.

But it also looks like training causes a participant’s age ???



An example → “when” and “operational definition” matter!!!

Bivariate & Multivariate contributions – DV = Exam 1% grade

predictor→	Motiv	St. Time	GPA	% Pink
r(p)	.28(<.01)	.45 (<.01)	.46 (<.01)	.33(<.01)

All of these predictors have substantial correlations with Exam grades!!

β(p)	.32(.02)	**-.25(.04)**	.09(.51)	.48 (.01)

GPA does not have a significant regression weights – after taking the other variables into account, it has no unique contribution!

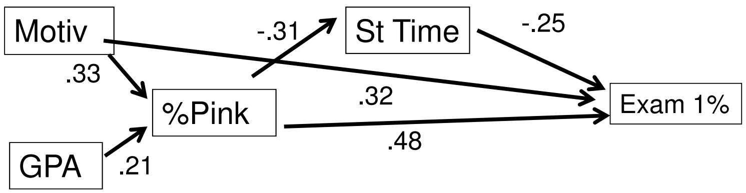
Exam study time has a significant regression weight, however, notice that it is part of a suppressor effect! After taking the other variables into account, those who study more for the test actually tend to do poorer on the exam.

%Pink does have a significant regression weight. Even after taking the other variables into account, those who do more MTAs do better on the exam.

Motivation does have a significant regression weight. After taking the other variables into account, those who are more motivated do better on the exam.

Notice that only two of the 4 predictors had the same “story” from the bivariate and multivariate analysis!!!! Let’s path this…

Path Analysis – allows us to look at how
relate to the direct effects" indirect effects criterion

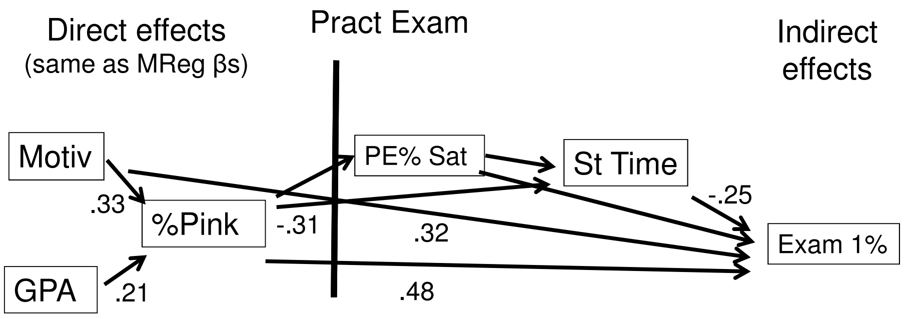


GPA → no direct effect – but indirect effects thru %pink & St Time

Motiv → direct effect – also indirect effects thru %pink & St Time

%Pink → direct effect – also indirect effect thru St Time

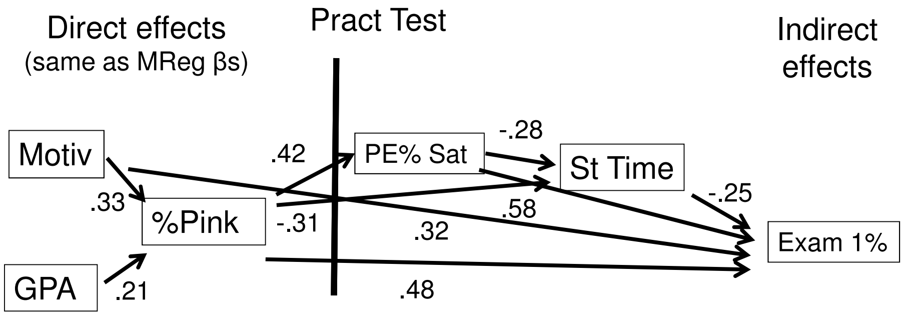
-β for St Time? Fewer %Pink predicts more St Time, predicts poorer Exam performance?
Let's add a variable and specify operationalizations!



Students took a Practice Exam a week before the Exam, got the scores back the same day and gave a 1-10 rating of their "Practice Exam% Satisfaction".

Study Time was defined as how much they studied between when then took the practice exam and the exam

Let's rerun the model with the Practice Exam% Satisfaction score included!!



Things make more sense now!!!

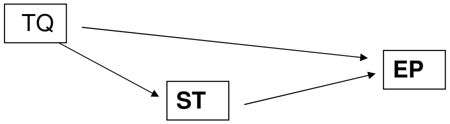
higher %Pink → higher PE%Sat → lower Stime → higher Exam%

Explaining the two "surprising" negative weights…

higher %Pink → lower Stime → higher Exam%

Some of the ways to improve a path analysis

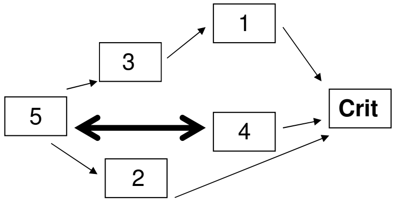
For a given model, consider these 4 things….



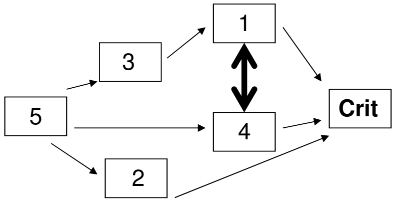
1. Antecedents to the current model
   - Variables that “come before” or “cause” the variables in the model
2. Effects of the current model
   - Variables that “come after” or “are caused by” the variables in the model
3. Intermediate causes
   - Variables that “come in between” the current causes and effects.
4. Non-linear variations of the model
   - Curvilinear & interaction effects of & among the variables

About non-recursive (bi-directional) models

Sometimes we want to consider whether two things that “happen sequentially” might have “iterative causation” – so we want to put in a back-and-forth arrow



Sometimes we want to consider whether two things that “happen at the same time” might have “reciprocal causation” – so we want to put in a sideways arrow



Neither of these can be “handled” by path analysis.

However, this isn’t really a problem because both are a misrepresentation of the involved causal paths! The real way to represent both of these is …

The things to remember are that:

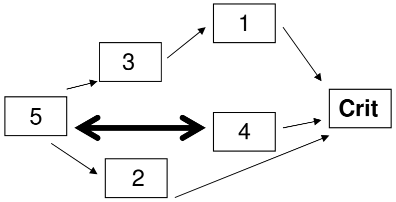
1. “cause takes time” or “cause is not immediate”
   - even the fastest chemical reactions take time
   - behavioral causes take an appreciable amount of time
2. Something must “be” to “cause something else to be”
   - a variable has to be manifested as an effect of some cause before it can itself be the cause of another effect
   - Cause comes before effect → not at the same time

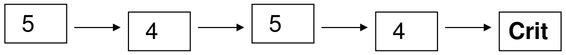
When you put these ideas together, then both “sideways” and “back-and-forth” arrows don’t make sense and are not an appropriate portrayal of the causations being represented.

The causal path has to take these two ideas into account…

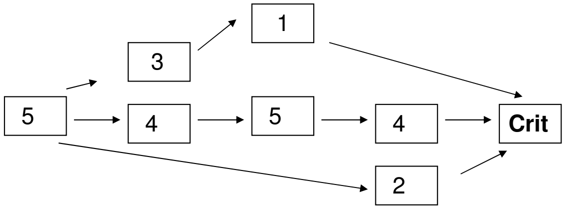
About non-recursive (bi-directional) models

If “5” causes “4”, then “4” changes “5”, which changes “4” again, all before the criterion is caused, we need to represent that we have 2 “4s” and 2 “5s” in a hypothesized sequence.



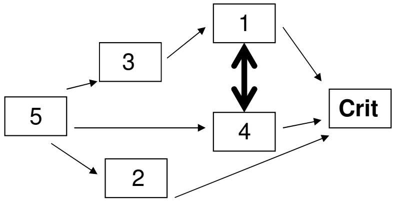


We also have to decide when1, 2 & 3 enter into the model, temporally &/or causally. Say …

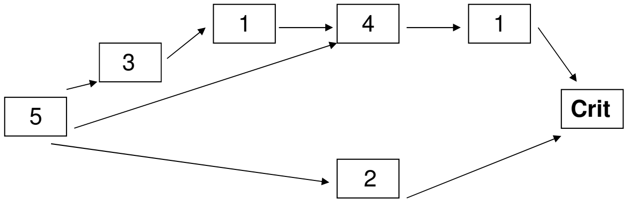


About non-recursive (bi-directional) models, cont…

When applying these ideas to “sideways arrows” we need to remember that the cause comes before the effect.



To do that, we have to decide (& defend) which comes first – often the hardest part) and then add in the second causation, etc.… As well as sort out where the other variables fall temporally &/or causally. Perhaps …

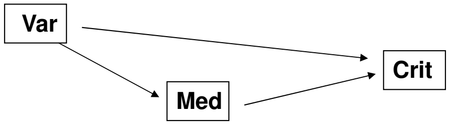


Mediation Analyses

The basic mediation analysis is a 3-variable path analysis.

A correlation shows that “var” is related to the “crit” .

But we wonder if we have the “whole story” – is it really that variable that causes Crit ???



So, we run a regression analysis w/ Var & Med as preds of Crit. Then we compare two estimates of the Var – Crit relationship

- $r_{Crit,Var}$ from the bivarate model &
- β_{Var} from the multivariate model

If β_{Var} = .00 → complete mediation

If $.00 < \beta Var < r_{Crit,Var}$ → partial mediation

If $\beta Var = r_{Crit,Var}$ → no mediation

The Sobol test is used to evaluate the $r_{Crit,Var} - \beta_{Var}$ difference

Model “Identification” & Testing

Just-identified model

- number of path coefficients to be estimated equals the number of independent correlations → $(k*(k-1)) / 2$
- “full model” with all recursive paths

Over-identified model

- more correlations than path coefficients
- because one or more path coefficients are set to zero

Under-identified model

- more math coefficients to be estimated than independent correlations
- “can’t be uniquely estimated”
- full model with nonrecursive paths

Testing Causal Models

Theory Trimming

- fancy phrase for “deleting non-contributing paths”
- identify paths with nonsignificant contributions (non significant β in the relevant regression model) and call them “zero”

Concerns & Challenges

- usual problems of *post-hoc* procedures – must support model…
  - based on literature review
  - “test” model on a new sample
- problem is compounded in path analysis (relative to a single regression model) because testing of contributions within a single regression is not a test of the contribution of that path to the model
  - it is possible to find that deleting one or variables that do not contribute to a particular multiple regression does degrade the fit of the path model to the data

Testing Over-identified models

When we hypothesize that certain path coefficients are zero (that certain direct effects don’t contribute to the model) the resulting model is over-identified and can be compared to the fit of …

- the related just-identified (full) model
- other related over-identified models in which it is “nested”

It is really important to remember that you can not deduce that one path model (the arrangement of “layers” and “variables”) is better than another from these tests!! These tests only examine the contribution of specific variables within a specific model to that model, they do not test “the model”

- By analogy …
- we know we can’t talk about which multiple regression model is better based on which one has the bigger R^2 change when we drop a particular predictor from each
- we can’t say which path model is better based on which one changes most when certain paths are set to zero

Testing Over-identified models

Testing H0: “The Reduced model fits the data as well as the Full model”

1. Calculate the variance accounted for by the full model

$$R^2_{full} = 1 – \Pi(1-R^2_{Fi}) = 1 – (1-R^2_{F1})*(1-R^2_{F2})*(1-R^2_{F3})…$$

where R^{2Fi} is the R^2 from each regression used to get the coefficients of the full model (all with all predictors included)

2. Calculate the variance accounted for by the reduced model

$$R^2_{reduced} = 1 – \Pi(1-R^2_{Ri}) = 1 – (1-R^2_{R1})*(1-R^2_{R2})*(1-R^2_{R3})…$$

where R^2_{Ri} is the R^2 from each regression used to get the coefficients of the reduced model (at least one of which has had one or more predictors excluded; i.e., that predictor’s path set to .00)

Testing Over-identified models

3. Calculate W – the summary statistic of model-fit difference

$$W = -(N – d) \log_e \left(\frac{1 - R^2_{full}}{1 - R^2_{reduced}} \right)$$

N = sample size

d = # deleted paths

4. Obtained the W_{crit} value

$$W_{crit} = X^2_{crit} \text{ for df} = d$$

5. Test the H0:

If $W > W_{crit}$, reject H0: that Full = Reduced and conclude

- the full model fits the data better than the Reduced model
- one or more of the deleted paths contributes to the model

If $W < W_{crit}$, retain H0: that Full = Reduced and conclude

- the Reduced model fits the data as well as the Full model
- the deleted paths did not contribute to the model