The Finale: Path Modeling and Structural Equation Modeling (SEM)

- Topics:
  - Path modeling: vocabulary, fit, and testing mediation
  - The Big Picture of SEM
  - What to do (and what NOT to do) when SEM breaks for you
    - Single indicator (ASU) models
    - Parceling indicators
    - Using single factor scores
    - Multiple plausible values of factor scores

Path Models: Pictures and Equations

- Path model: Multivariate models for predicting 2+ outcomes simultaneously for the same unit of analysis
- Most often expressed as a diagram using these conventions:
  - Boxes = observed variables; circles = latent variables (in SEM) or residual
  - One-headed arrow = regression (arrow points from predictor to outcome)
  - Two-headed arrow = residual covariance; intercepts typically not shown

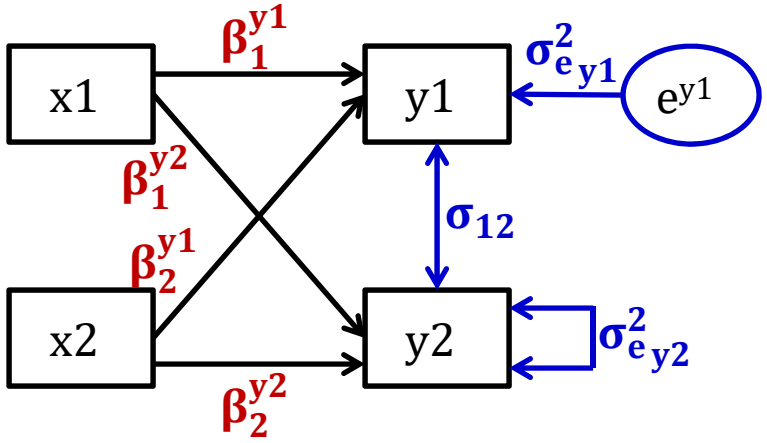
Diagram translates into these simultaneous regression models (in which superscripts denote the outcome of each parameter):

$$y1_i = \beta_0^{y1} + \beta_1^{y1}(x1_i) + \beta_2^{y1}(x2_i) + e_i^{y1}$$
$$y2_i = \beta_0^{y2} + \beta_1^{y2}(x1_i) + \beta_2^{y2}(x2_i) + e_i^{y2}$$

Unstructured R matrix for outcome variances and covariance(s):

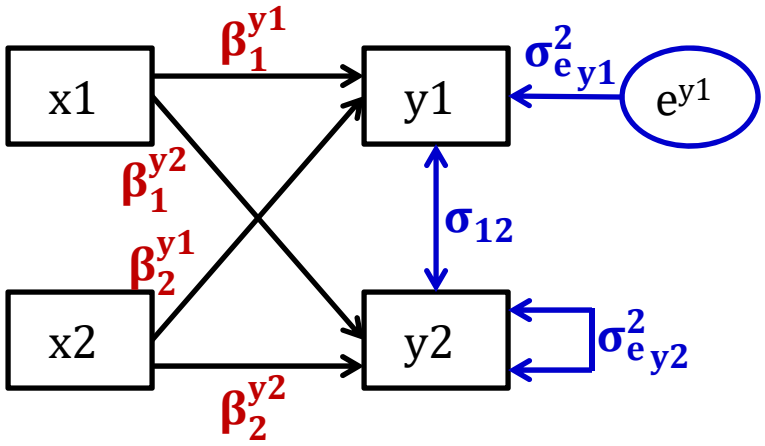
$$\begin{bmatrix} \sigma^2_{e_{y1}} & \sigma_{12} \\ \sigma_{12} & \sigma^2_{e_{y2}} \end{bmatrix}$$

The idea of residual variance is either expressed using a separate circle (e.g., for Y1) or a two-headed arrow into itself (e.g., for Y2).



Multivariate Regression via Path Models

- This example is really just two univariate regression models estimated simultaneously
  - β_1 and β_2 provide the unique effects of x1 and x2 for y1 and y2 outcomes
  - Can calculate R^2 for each outcome



- So why do both at once?
  - To test differences in effect size (e.g., does $\beta_1^{y1} = \beta_2^{y1}$?)
  - To test mediation and indirect effects, in which a variable is both a predictor and an outcome in the same analysis (stay tuned)

If these variables came from a dyad of two persons (1 and 2), this could be an example of an "actor–partner model"

- Arrows within same person = "actor effects"
- Arrows across different people = "partner effects"

2 Types of Path Model Solutions

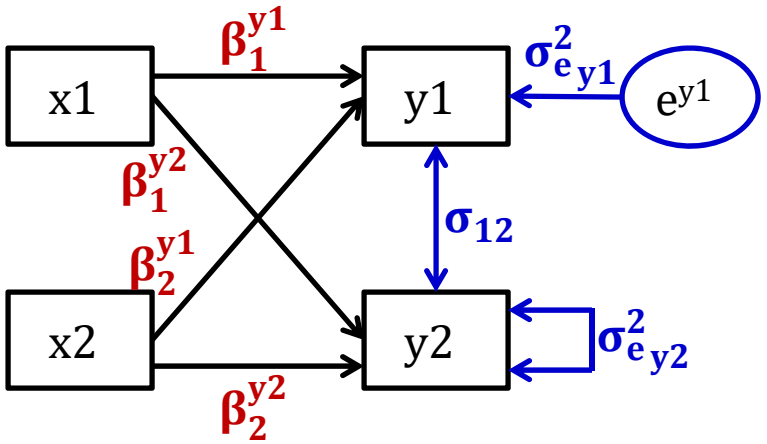
- Unstandardized → predicts scale-sensitive original variables:
  - **Regression Model:** $y1_i = \boldsymbol{\beta_0^{y1}} + \boldsymbol{\beta_1^{y1}}(x1_i) + \boldsymbol{\beta_2^{y1}}(x2_i) + \boldsymbol{e_i^{y1}}$
  - Useful for comparing across groups (whenever absolute values matter)
  - Model parameters predict the intercepts and covariance matrix
  - Variance of **y1** = **[variance explained by predictor fixed effects]** + $\boldsymbol{\sigma^2_{e_{y1}}}$

- Standardized → Solution using z-scored versions of variables:
  - Useful when comparing effects within a solution (are then on same scale)
  - Standardized model parameters predict the **variable correlation matrix**
  - Standardized slope = $[\boldsymbol{\beta_1^{y1}} * \mathbf{SD(x1)}]$ / $\mathbf{SD(y1)}$ = **unique correlation**
  - $\mathbf{R^2}$ for **y1** = $\mathbf{1 - standardized\ \sigma^2_{e_{y1}}}$

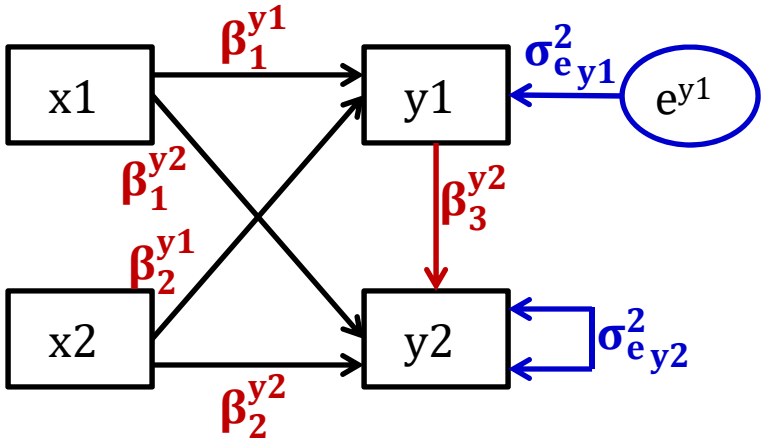
New (and Confusing) Terminology

- Predictors are known as **exogenous** variables (X-ogenous to me)
- Outcomes are known as **endogenous** variables (IN-dogenous to me)
- Variables that are both at once are called **endogenous** variables

<u>Our previous example model:</u>
2 exogenous variables (x1 and x2)
2 endogenous variables (y1 and y2)



<u>Our modified example model:</u>
2 exogenous variables (x1 and x2)
2 endogenous variables (y1 and y2)



New (and Confusing) Terminology

- What parameters get estimated for exogenous "predictor" and endogenous "outcome" variables differs importantly by program!
  - Only the intercepts, residual variances, and residual covariances of "outcome" variables are estimated as part of the likelihood…
- But this distinction is not as clear-cut as one might think…

- By default **in Mplus**, *truly* exogenous predictor variables cannot have missing data (the same as in any linear model)
  - Cases with missing predictors are **listwise deleted** out of the model (incomplete data are assumed missing completely at random)
  - Because predictors are not explicitly part of likelihood function
    - LL contains $\hat{y}_i$ for each person and σ_e^2 for each outcome
    - So LL can't be calculated without the predictors that create each $\hat{y}_i$
  - **But these exogenous predictors do not have distributions**…
    - Good when you want to include non-normally-distributed predictors!

"Predictors" as Endogenous Outcomes

- Mplus allows you to bring exogenous predictors into the likelihood → predictors then become "outcomes" in terms of their parameters (estimated means, variances, and covariances)
  - Even if nothing predicts the predictor (it's not really an outcome)
  - These predictors can then have missing data assuming missing at random (conditionally random given the rest of the model)
  - **These predictors then have distributional assumptions (usually MVN)**
  - Mplus will not let endogenous "predictors" have other distributions (so you will have to make them an outcome of something else to fix this)

- **Exogenous predictors are forced into the likelihood in STATA SEM and SAS CALIS** (and I have not been able to find how to force predictors out of the likelihood, but STATA GSEM may allow it)
  - STATA SEM "xconditional" computes their means, variances, and covariances from the observed data to save time given complete data (and searches for them as model parameters otherwise), but these values then go into the likelihood, which means exogenous predictors have assumed distributions

What Goes In (data used as input)

- Observed mean per variable
- Observed variance per variable
- Observed covariance between each pair of variables
- This is the data the model is trying to "fit"!

What Comes Out (estimated parameters)

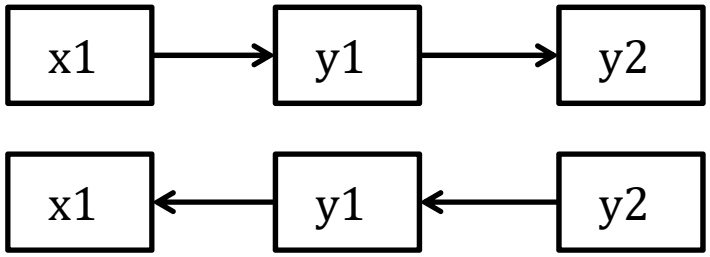
- Estimated intercept per variable (to *perfectly* re-create the observed variable means)
- Estimated residual variance per variable (to *perfectly* re-create the observed variances)
- Estimated regression path or covariance between each pair of variables (to predict their observed covariances)
  - If some are omitted, then observed covariances will not be perfectly reproduced → **room for misfit**

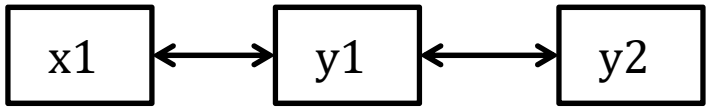
Model Identification (assuming all variables are in the likelihood)

- Identification: can the model parameters actually be "solved for"?
  - Requires that # of estimated parameters is ≤ # of possible parameters
  - # possible is sum of # means, variances, and covariances for v variables
    → shortcut formula = possible degrees of freedom = $(v[v + 1] / 2) + v$
- 3 possible model identification scenarios:
  - **Under-identified:** # estimated parameters > # possible → negative df
    - Model is not solvable (parameter estimates cannot be found); game over
  - **Just-identified:** # estimated parameters = # possible → 0 df
    - Model is solvable (is most common scenario perfectly reproduces original data)
    - Assessment of absolute model fit will NOT be relevant (which is good for path models)
  - **Over-identified:** # estimated parameters < # possible → positive df
    - Model is still solvable (and is more parsimonious description of original data)
    - Assessment of absolute model fit is then necessary (more relevant for latent variables)

Model Identification Examples
(in which each variable has an estimated mean/intercept and variance/residual variance)

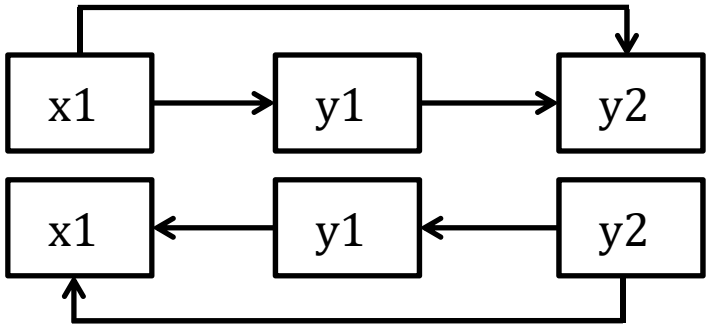
- Over-identified: have positive df leftover (estimated<possible)

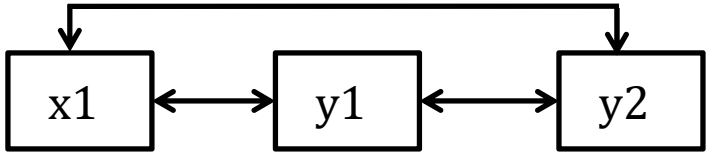




These 3 models all have equivalent fit with **df=1** (for the 1 missing direct relationship).

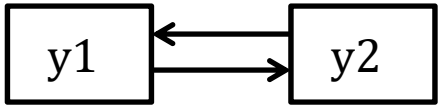
- Just-identified: have 0 df leftover (estimated=possible)





These 3 models all have equivalent fit with df=0 (for 0 missing direct relationships).

- Under-identified: have negative df (estimated>possible)



This model is trying to estimate 2 paths using only 1 covariance (can't be solved).

Model Evaluation: Steps 1, 2, and 3

1. **Assess global absolute model fit**
   - Recall that variable means and variances are perfectly predicted (just-identified) → *misfit comes from messed-up covariances*
   - χ^2 is sensitive to large sample size, so pick at least one global fit index from each class; hope they agree (e.g., CFI, RMSEA)
2. **Identify localized model strain**
   - Global model fit means that the observed and predicted covariance matrices aren't too far off on the whole… says nothing about the specific matrix elements (reproduction of each covariance)
   - Consider normalized residuals and modification indices to try and "fix" the model – add missing relationships that should be there
3. **Revise the model until it fits**

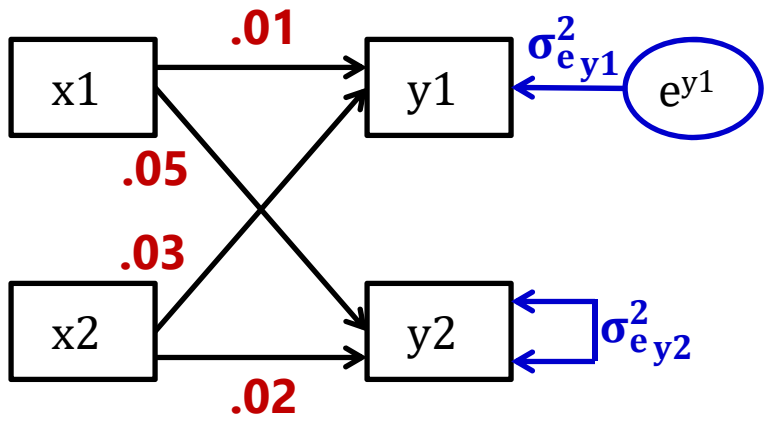
Good global and local fit? Great, but we're not done yet…

Step #4 in Model Evaluation

4. Inspect **parameter effect sizes** and significance
   - A good-fitting model does not necessarily imply a good model!
     - Can reproduce lack of covariance quite well and still not have anything useful – e.g., correlation of .2 → 4% shared variance?
     - **Effect size (R^2 for variance explained) is practical significance**

This example model may have "excellent fit" (testable because df=2) but no significant regression paths…

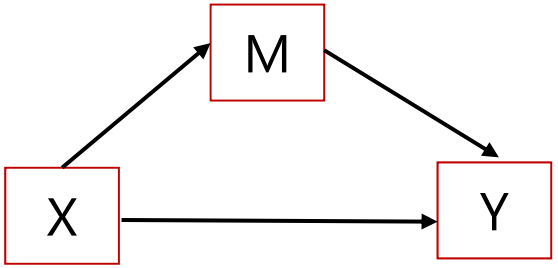
Why? Good absolute fit just means it has successfully reproduced the (non)relationships among these variables—not whether there are relationships worth reproducing!



Terminology: Mediation ≠ Moderation

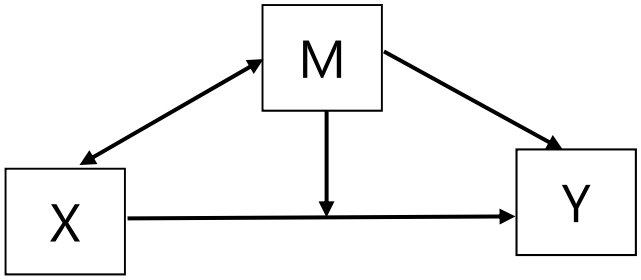
Mediational model (regression with better marketing):

- X **causes** M, M **causes** Y
- M is an outcome of X but a predictor of Y

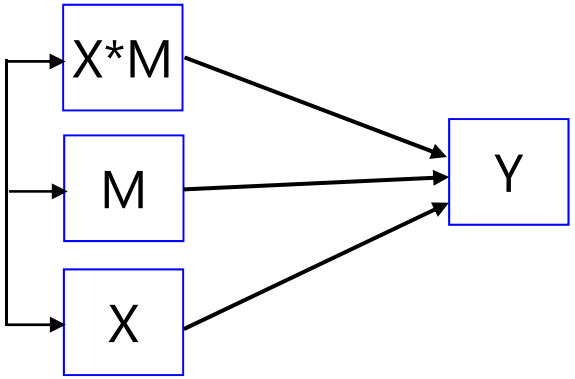


Moderator model:

- M adjusts the size of X→Y relationship
- M is a predictor of Y, and is **correlated** with X
- Moderation is represented by an **interaction** effect



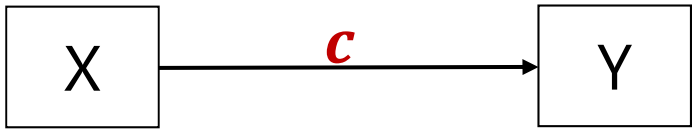
This figure does NOT depict an estimable model.

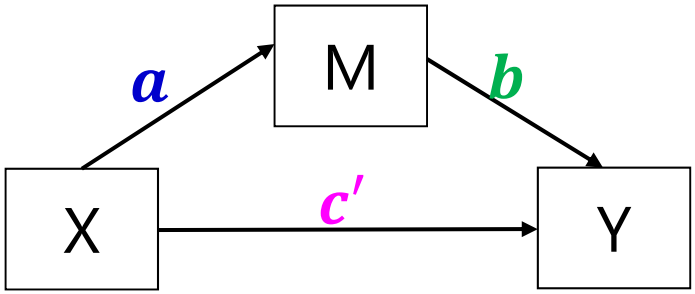


This is what is actually implied by above model.

Terminology: Mediation Effects

c = uncontrolled X to Y path
(Y regressed on X)





The big question in mediation:

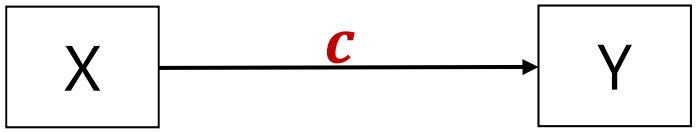
- Phrased as usual regression → *Is the effect of X predicting Y still significant after controlling for M?*
- Phrased as "mediation" → *Is the effect of X predicting Y significantly mediated by M?* **OR** *Is there a significant indirect effect of X through M in predicting Y?*
- Phrased either way, is $c \neq c'$?

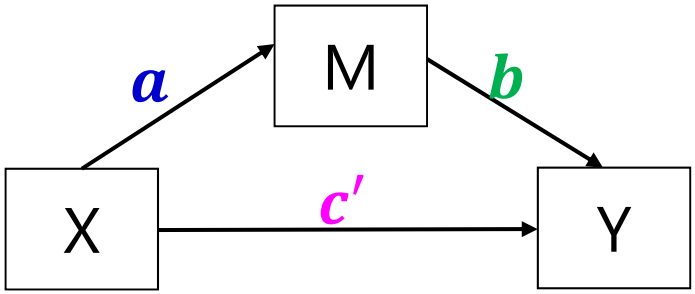
Direct Effects:

- a = X to M path (M on X;)
- b = M to Y path (Y on M;)
- c' = X to Y path controlled for M (Y on X;)
- $a * b$ = indirect effect of X to Y
- The estimates for $c - c'$ and $a * b$ will be equivalent in MVN observed variables (if same *N*)

Old versus New Rules for Mediation

c = uncontrolled X to Y path
(Y regressed on X)





- Baron & Kenny (1986, JPSP) rules were standard for a long time…
  - Simulation studies have found these rules to be way too conservative

- Old rule that can now be broken:
  - X must predict Y in the first place (c must be initially significant)
  - When not? Differential power for paths; suppressor effects of mediators
  - Mediation is really about whether $c \neq c'$, not whether each is significant

- Old rules that pry still hold:
  - X must predict M (a must be significant)
  - M must predict Y (b must be significant)

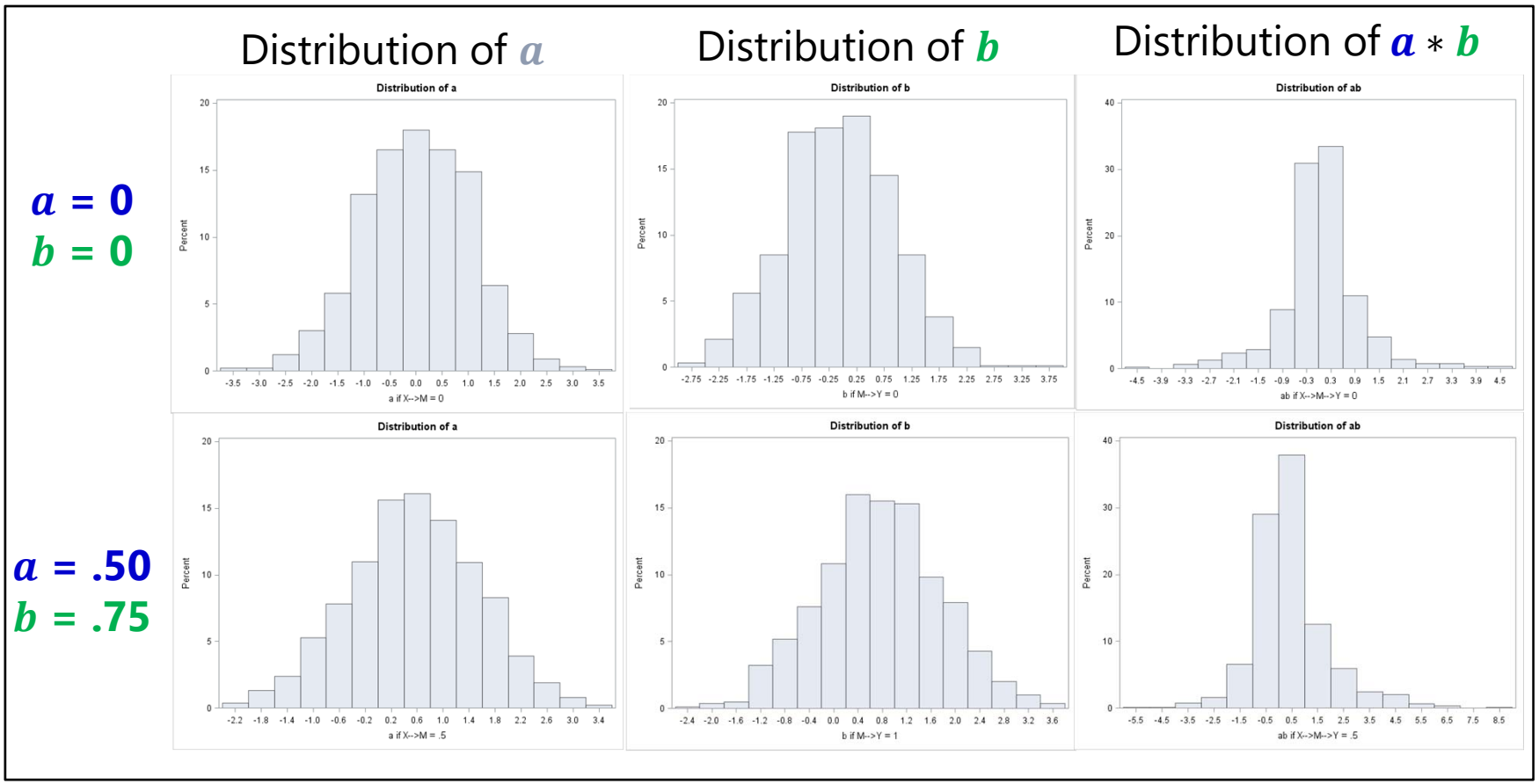
Testing Significance of Mediation

- Need to obtain a SE in order to test if $c - c' = 0$ or if $a * b = 0$
  - For $c - c'$ → "difference in coefficients SE"
  - For $a * b$ → "product of coefficients SE" → we'll start here

- Use "multivariate delta method" (second-derivative approximation shown here) to get SE for product of two random variables $a * b$
  - $SE_{a*b} = \sqrt{a^2 SE_b^2 + b^2 SE_a^2 + SE_a^2 SE_b^2}$
  - An equivalent formula to calculate SE_{a*b} that may have less rounding error because it avoids squaring a and b is $SE_{a*b} = \frac{ab\sqrt{t_a^2 + t_b^2 + 1}}{t_a t_b}$
  - This is known as the "Sobel test" and can be calculated by hand using the results of a simultaneous path model or separate regression models, also provided through MODEL INDIRECT/CONSTRAINT in Mplus, NLCOM in STATA SEM, or TESTFUNC in SAS PROC CALIS

Testing Significance of Mediation

- One problem: we *shouldn't* use this SE for usual significance test
  - So, nope: $t_{indirect} = \frac{a*b}{SE_{a*b}}$ or $95\%\ CI = a * b \pm 1.96 * SE_{a*b}$
  - Why? Although the estimates for a and b will be normally distributed, the estimate of their product won't be, especially if a and b are near 0



Testing Significance of Mediation

- So what do we do? Another idea based on same premise:
  - For $\boldsymbol{a} * \boldsymbol{b}$ → find "distribution of the product SE" → $\boldsymbol{z_a} * \boldsymbol{z_b} = \frac{\boldsymbol{a}}{\boldsymbol{SE_a}} * \frac{\boldsymbol{b}}{\boldsymbol{SE_b}}$ in which the sampling distribution does not have a tractable form, but tables of critical values have been derived through simulation for the single mediator case (but may not generalize to complex models)
  - Implemented in PRODCLIN program for use with SAS, SPSS, and R

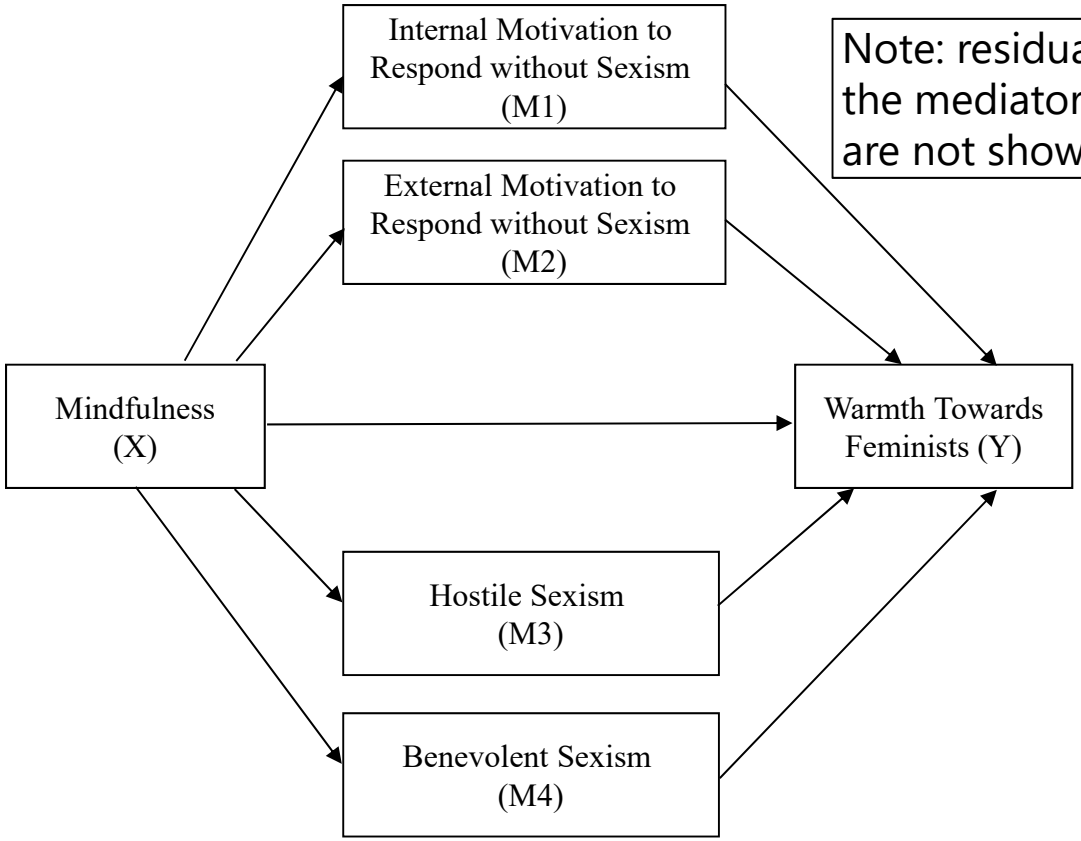
- A better solution: **bootstrap the data** to find the empirical SE and asymmetric CI for the indirect effect
  - Bootstrap = draw n samples with replacement from your **data**, re-estimate mediation model and calculate $\boldsymbol{a} * \boldsymbol{b}$ for each bootstrap sample
  - Point estimate of $\boldsymbol{a} * \boldsymbol{b}$ is mean or median over n bootstrap samples
  - SE_{a*b} is standard deviation of estimated $\boldsymbol{a} * \boldsymbol{b}$ over n bootstrap samples
  - 95% CI can be computed as estimates at the 2.5 and 97.5 percentiles
  - Typically at least 500 or 1000 n bootstrap samples are used

Testing Significance of Mediation

- There are multiple kinds of bootstrap CIs possible in testing the significance of the $a * b$ indirect effect within MVN data
  - Regular bootstrap CI = "**percentile**" (as just described)
    - In Mplus, OUTPUT: CINTERVAL(bootstrap); in STATA SEM, vce(bootstrap)
  - **Bias-corrected bootstrap** CI = shifts CIs so median is sample estimate *** *Supposed to be best one*
    - In Mplus, OUTPUT: CINTERVAL(BCbootstrap); not sure about STATA SEM
  - Accelerated bootstrap CI = ???
    - Not given in Mplus (as far as I know); not sure about STATA SEM

- For not simply MVN data (i.e., non-normal mediators or outcomes, multilevel data), a different bootstrap approach can be used as a separate step using any program's output
  - *Parametric*, *Monte Carlo*, or *empirical-M* bootstrap → Draw repeatedly from a and b parameter distributions instead of the data, then compute point estimates, SEs, and CIs from those distributions
  - See http://www.quantpsy.org/medn.htm for online calculators

Our Mediation Example 9a



Note: residual covariances among the mediators were estimated but are not shown for diagram clarity.

Figure 1 from: Gervais, S. J. & Hoffman, L. (2013). Just think about it: Mindfulness, sexism, and prejudice towards feminists. *Sex Roles, 68*(5), 283-295.

Research questions:

(1) To what extent do these four mediators account for the relationship between mindfulness and warmth towards feminists?

(2) How do these direct and indirect effects differ by gender?

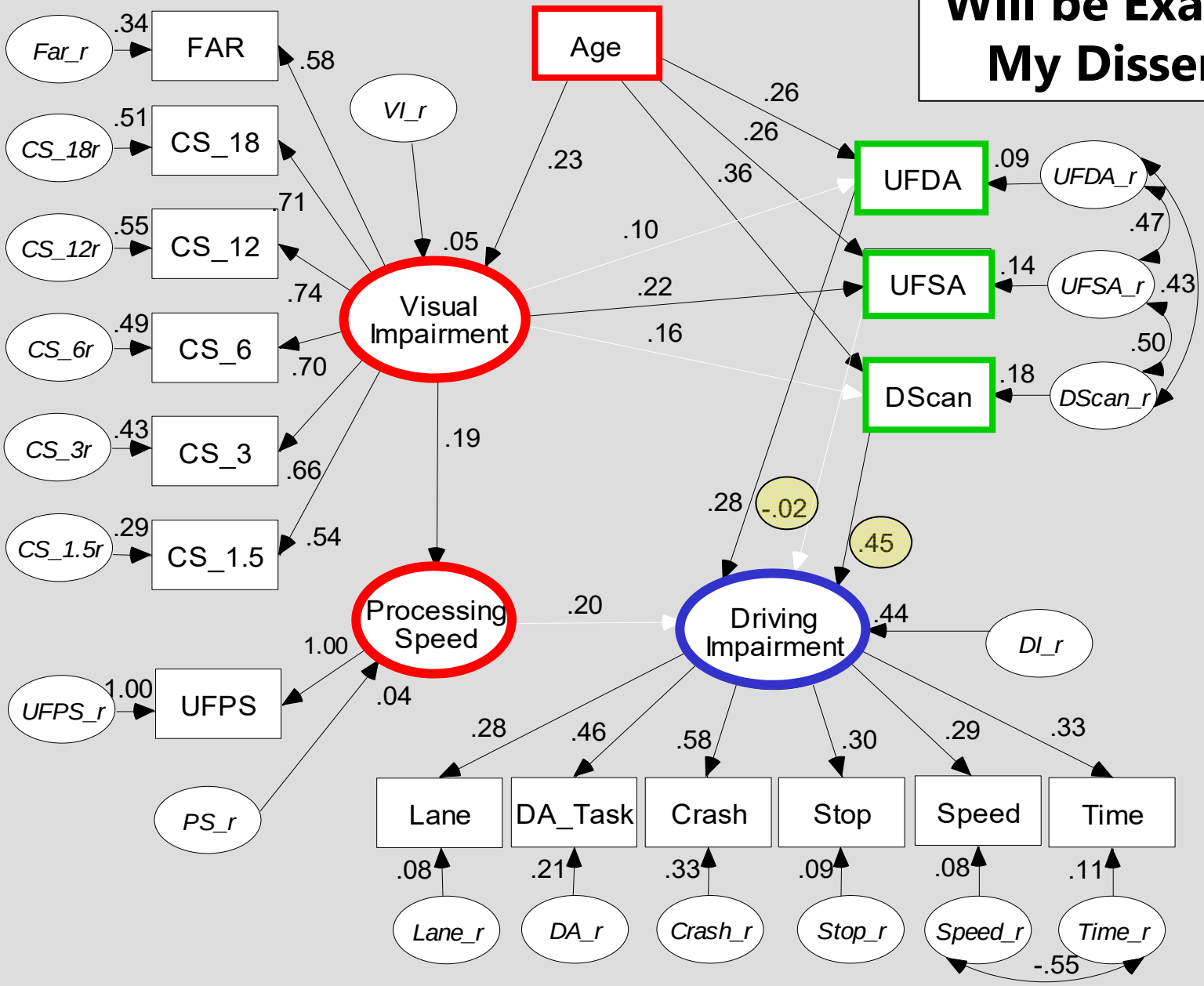
Mediation with Non-Normal Variables

- All the path models shown so far have assumed every variable in the likelihood* is multivariate normal
  - * In the likelihood → is predicted by something or has an estimated mean, variance, or covariance (i.e., the missing data trick called "I used FIML")
  - In reality, one may have non-normal (NN) mediators or outcomes…

- Estimation gets tricky, because there is no closed-form ML anymore
  - NN outcomes → fit link function to Y, requires numeric integration
    - Becomes exponentially more complex with more non-normal variables
  - NN mediators → fit link function M, but estimation is even trickier
    - In Mplus, requires Monte Carlo integration (re-sampling approach)

- Interpretation gets tricky, because the paths are of different kinds
  - For example, X → M → binary Y: X → regular M, M → logit Y
  - For example, X → binary M → Y: X → logit M, regular M → Y
  - Oh, and there are no standard absolute model fit statistics in ML (no observed covariance matrix to compare the model predictions to)

Will be Example 9c: My Dissertation



Example 9b: Hoffman & McDowd (2010, *Psychology and Aging*)

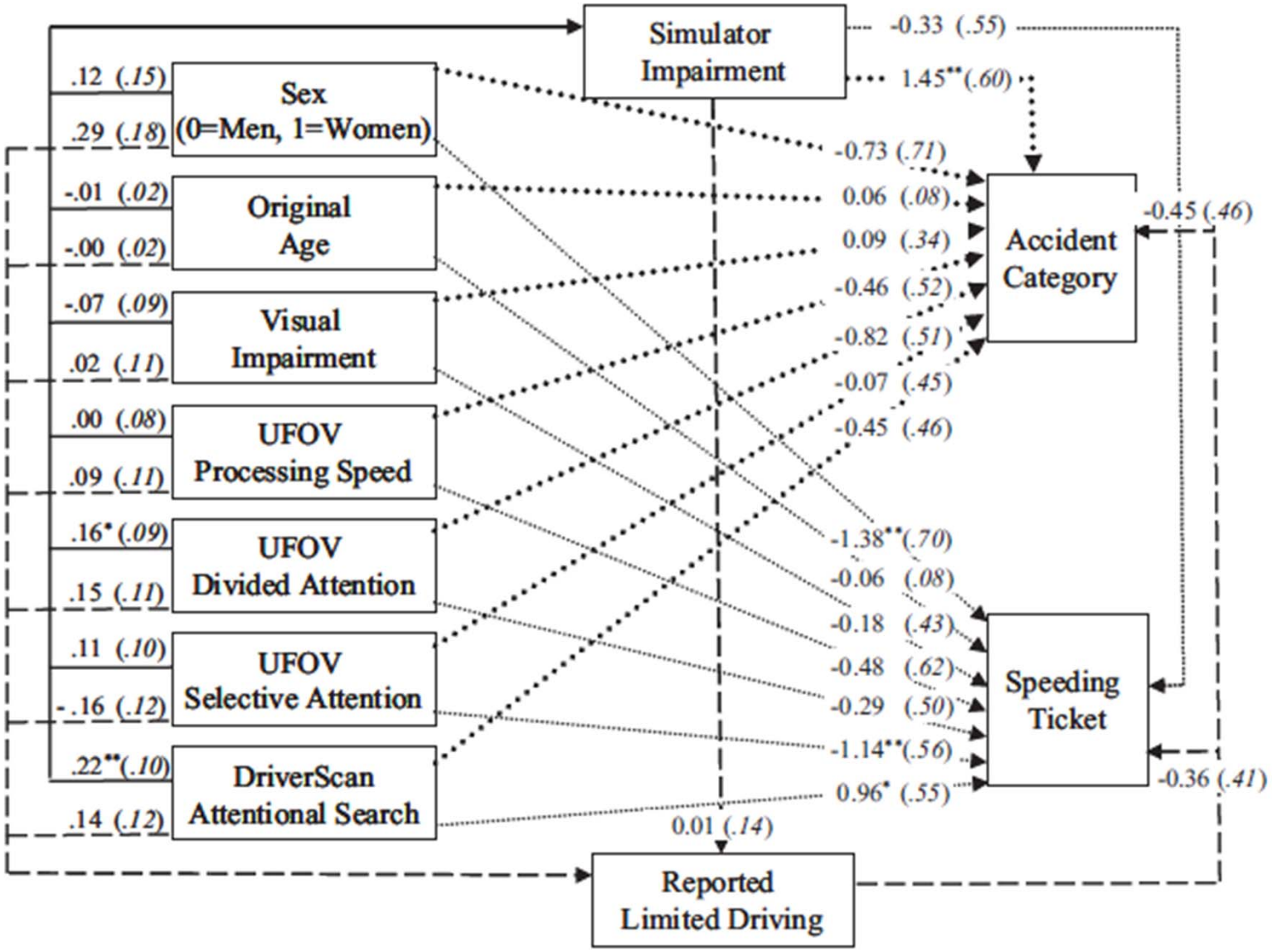
- Follow-up data from 114/152 persons from dissertation sample
  - 91 reported no accident since then, 9 reported no-fault accident
  - 14 reported at least partially-at-fault accident
  - 14 reported a speeding ticket
  - Tendency to limit driving (mean of 4 Likert items on 1-5 scale, 0 = 2)
  - Only 3 persons no longer drove

- No differences were found between completers/non-completers in sex, age, visual impairment, UFOV, DriverScan, or simulator impairment

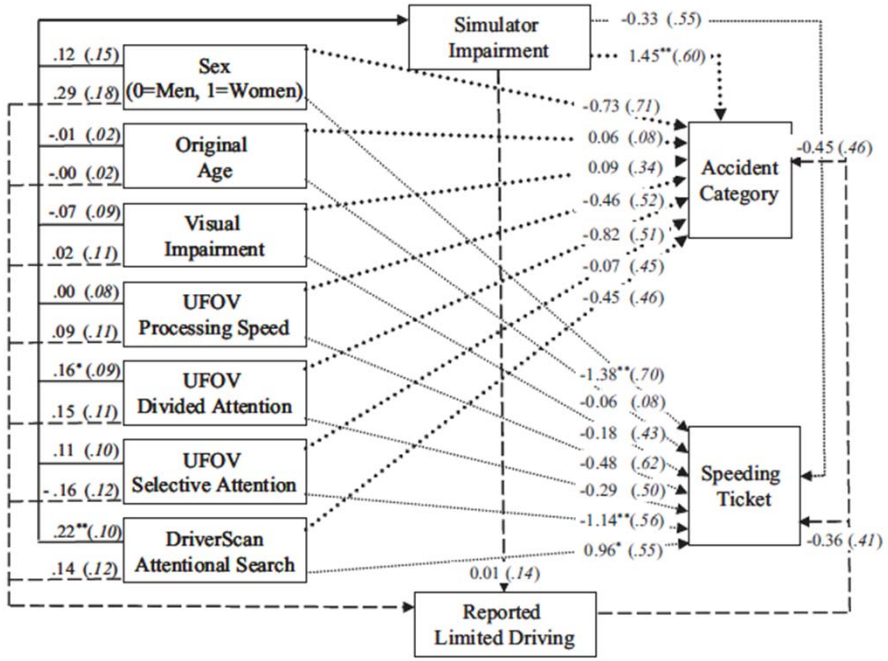
- Model: Predict accidents and speeding tickets (binary outcomes)

- Original analysis used ML with MonteCarlo Integration
  - I'll use MLR to demonstrate here → MVN then assumed for continuous mediators of simulator driving impairment and limiting driving

Path Model Predicting Binary Driving Outcomes



Mplus Code for Direct and Indirect Effects



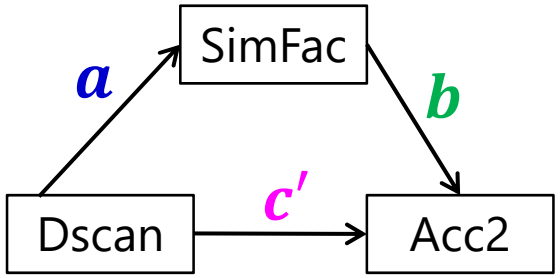

```
TITLE:   Path Analysis Dissertation Follow-up
DATA:    FILE = driver.dat;
VARIABLE:
! List of variables in data file
NAMES = PartID sex age75 cs_1_5 cs_3 cs_6
cs_12 cs_18 far near zufov1 zufov2 zufov3
Dscan lane da_task crash stop speed time
simfac part visfac attfac limit4 ticket2
speed2 follow attr nacc2 jacc2 acc2;
! Variables to be analyzed in this model
USEVARIABLE = sex age75 visfac zufov1 zufov2
zufov3 Dscan simfac limit4 speed2 acc2;
! Missing data identifier
MISSING = .;
! Categorical outcomes
CATEGORICAL = acc2 speed2;
ANALYSIS: ! Estimation options
ESTIMATOR = MLR; INTEGRATION = MONTECARLO;
OUTPUT: STDYX;
```

```
MODEL: ! With labels for specific paths in order of list
simfac ON sex age75 visfac zufov1 zufov2 zufov3 Dscan                (sim1-sim7);
limit4 ON sex age75 visfac zufov1 zufov2 zufov3 Dscan simfac         (lim1-lim8);
acc2   ON sex age75 visfac zufov1 zufov2 zufov3 Dscan simfac limit4  (acc1-acc9);
speed2 ON sex age75 visfac zufov1 zufov2 zufov3 Dscan simfac limit4  (spd1-spd9);

MODEL CONSTRAINT:        ! Like ESTIMATE in SAS
NEW(DStoAcc);            ! List names of estimated effects on NEW
DStoAcc = sim7 * acc8;   ! Indirect effect of Dscan --> Sim --> Acc
```

Partial Mplus Output (for Direct and Indirect Effects)

```
MODEL FIT INFORMATION
Number of Free Parameters                       39
Loglikelihood
          H0 Value                          -356.400
          H0 Scaling Correction Factor         1.0066
            for MLR
Information Criteria
          Akaike (AIC)                       790.799
          Bayesian (BIC)                     907.953
          Sample-Size Adjusted BIC           784.529
            (n* = (n + 2) / 24)
```



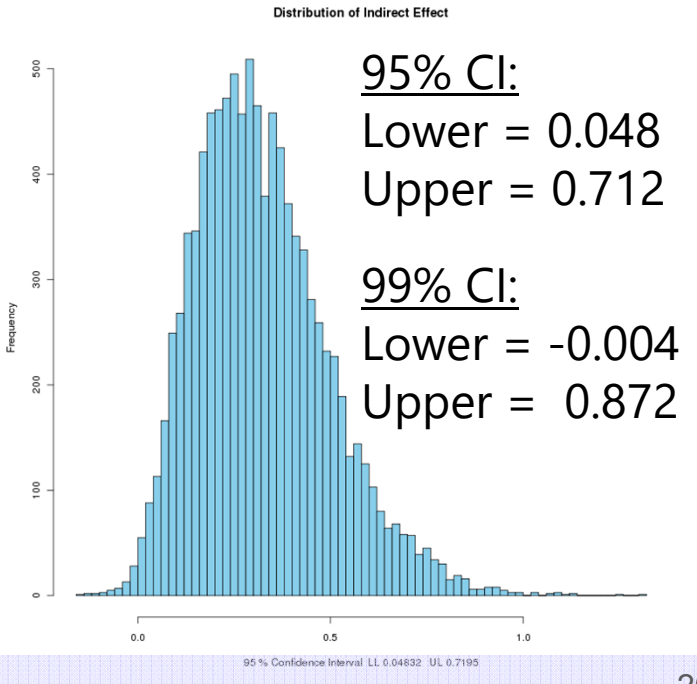
Then used Monte Carlo resampling to assess empirical distribution of indirect effect via this web utility: http://www.quantpsy.org/medn.htm

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value
SIMFAC     ON
   DSCAN             0.216      0.081      2.661      0.008

ACC2       ON
   DSCAN            -0.477      0.320     -1.491      0.136
   SIMFAC            1.497      0.532      2.813      0.005

New/Additional Parameters
   DSTOACC           0.323      0.160      2.026      0.043
```



95% CI:
Lower = 0.048
Upper = 0.712

99% CI:
Lower = -0.004
Upper = 0.872

Path Models and Mediation: Summary

- Path models are a very useful way to examine many different multivariate hypotheses simultaneously:
  - Unique direct and indirect effects ("mediation")
  - Differences in effect size (via model constraints)
  - Relationships among mediators or outcomes (direct and indirect effects)

- Good fit is a pre-requisite to actually interpreting the model results, but good fit does *not* mean it is a good model
  - Good fit = model reproduces the covariance matrix of the likelihood variables (but it does not indicate how big or small those relationships are)
  - However – when all possible relationships among variables are estimated (either as covariances or direct regressions), fit is perfect and irrelevant
    - Also known as "multivariate regression" with an "unstructured R matrix"

- Watch out for assumptions about exogenous predictor variables
  - Are their means, variances, and covariances part of the likelihood? Then they have an assumed distribution (usually MVN), which may not make any sense!

Structural Equation Modeling (SEM)

- The term "SEM" gets used to describe many different models, but fundamentally, **SEM consists of two parts**:
  - **Measurement model for each latent variable**
    - "CFA" if indicators are continuous and "normal enough"
    - "IFA" if indicators are binary, ordinal, or nominal
    - "?name?" if indicators require some other link function (e.g., counts)
    - Factors/thetas/traits are assumed to be multivariate normal
  - **Path analysis (regressions) amongst the latent variables**
    - And amongst other observed variables that are not used as part of the measurement model for those latent variables
    - Other observed variables can be of whatever kind, so long as the observed outcomes have their distributions modeled properly
      - e.g., a binary predictor variable does not require a logit, but a **binary outcome variable** does (so then it's on the CATEGORICAL statement)
      - THERE IS NO SUCH THING AS A CLASS STATEMENT IN MPLUS (I'm sorry), so you have to create manual contrasts to include categorical predictors

SEM: Model Identification

- SEM integrates both measurement and path models, so the identification rules for SEM borrow from both
  - Measurement models for each latent variables must be locally identified → each factor has its own scale (mean, variance)
  - The path model must be identified (solvable)

- A necessary (but not sufficient) way of ensuring identification is the t-rule (i.e., a counting rule that I never use in SEM)
  - Number of estimated ("free") parameters must be less than the total number of means + variances/covariances of **all** observed variables (v) in the analysis: Total possible df = $\frac{v*(v+1)}{2} + v$
  - Practical tip: don't count, just look at your model, and see if it seems logical (e.g., don't have a directed path AND a covariance between two variables), make sure all latent factors are locally identified, and beware of negative factor loadings (then factors won't know which way to go)

SEM: Predictors vs. Outcomes

- New terminology for use in SEM:
  - Predictor variables are called "**exogenous**" (arrows go out of it only)
  - Outcome variables are called "**endogenous**" (arrows go into it)
  - If a variable is *both* a predictor and an outcome, it is "endogenous"

- Some SEM books claim that when using ML, that *all* variables should have a multivariate normal distribution (MVN), but this is NOT true in Mplus for three reasons:
  - You can use ML with link functions and other distributions (e.g., CATEGORICAL tells it to use Bernoulli or Multinomial instead as needed)
  - Exogenous variables are not part of the ML function unless you make them (by referring to their means, variances, or covariances in syntax)
  - Only the *residuals* of endogenous variables are assumed MVN
  - MLR can help with continuous but overly kurtotic endogenous variables

SEM: Predictors vs. Outcomes

- The important distinction is whether each observed variable is **part of the maximum likelihood function or not**
  - Are its means/intercepts, variances/residual variances, or covariances/residual covariances being estimated? Then yes, it is
  - Are *just* its paths predicting endogenous variables being estimated? Then no, it is NOT part of the likelihood
- **Upside** of putting exogenous variables in the likelihood?
  - Predictors can have missing data (assuming missing at random)
- **Downside** of putting exogenous variables in the likelihood?
  - Distributional assumptions then apply, although Mplus gets cranky when exogenous variables are added to CATEGORICAL
    - A silly work-around is to make it a perfect single indicator of a latent factor, that way it becomes an "outcome" officially, but this may cause other problems
  - Covariances amongst "predictors" then contribute to fit…

SEM: What goes into model fit

- Back in CFA/IFA, misfit was almost always due to covariances
  - If each indicator has its own **intercept or thresholds**, then the indicator **means or response frequencies** will be predicted perfectly
  - If each indicator has its own **residual variance**, then the indicator **total variances** will be predicted perfectly
  - **Factor loadings** are supposed to predict covariances among indicators, so once you have 4+ indicators in a model → **potential for misfit**

- The same is true in SEM, but with a catch, because only some covariances "count" towards model fit
  - Covariances amongst variables in the likelihood COUNT
  - Covariances for "predictors" (NOT in the likelihood) with "outcomes" (in the likelihood) COUNT
  - Covariances amongst "predictors" (NOT in the likelihood) do NOT count

SEM: What to do first?

- **Because SEM is composed of two distinct parts…**
  - Measurement model that identifies latent variables
  - Structural model for relations involving those latent variables

- … **you should build these models <u>sequentially</u>**
  - Start by ensuring each over-identified factor fits adequately
  - When possible, then combine all factors of interest and other observed variables in the same model, estimating all possible relations among them (this "saturated" model is the best-fitting structural model)
  - Then modify the structural model to answer your questions, and see if the simpler model is NOT worse than the saturated structural model

- Because the measurement model will dominate model fit, informative tests of the structural model need to focus THERE

SEM: What to do if I can't do it?

- A simultaneous estimation of measurement and structural models in SEM is the gold standard, but may not work for you

- SEM is likely to break (i.e., not converge, give crazy SEs) when:
  - Sample sizes are small (few persons relative to # estimated parameters)
  - Many estimated parameters (especially with few persons)
  - Some outcomes are non-normal (link functions are involved)
  - Many latent variables are included (especially with link functions)
  - Latent factors are not well-identified (2 indicators is not enough)
  - Latent variable interactions are included (which require numeric integration → repeated rectangling of the latent trait distributions)

- What to do then? Alternatives range from ok to terrible…

First try a simpler measurement model

- One way to save estimated parameters—when possible to do so without hurting model fit too much—is to **fit constrained measurement models** (i.e., make the parcels a real structure)
- For example, for a factor with 12 original indicators:
  - Total possible DF for actual 12 indicators = $\frac{12(12+1)}{2} + 12 = 90$
  - Used DF for **full one-factor** model = $12\lambda+12\mu+12\sigma_e^2$ = **36**
  - Used DF for **tau-equivalent** (Rasch) factor model = $1\lambda+12\mu+12\sigma_e^2$ = **25**
    - **It is more difficult to estimate more loadings than more μ or σ_e^2**
  - Used DF for **parallel items** factor model: $1\lambda+12\mu+1\sigma_e^2$ = **14**
  - Used DF for an **"empty means" parallel items** model: $1\lambda+1\mu+1\sigma_e^2$ = **3**
  - If not all loadings/residual variances/intercepts can be constrained across items, perhaps at least some of them can?
  - Mplus allows you to test intermediate possibilities, not just all or nothing with respect to each indicator gets its own parameter(s)

3 Problems with SEM Alternatives*

1. Assuming **unidimensionality** and **tau-equivalence** (equal discrimination) of indicators within a single sum score
   - If these do not hold, the validity of the factor is questionable

2. Assuming **perfect reliability** of observed variables
   - If reliability is not perfect, then the estimates of its relationships with other variables will be downwardly-biased (weaker than they should be)

3. Assuming each person's **trait estimate is perfectly known**
   - If zero variability of a person's trait estimate is assumed, then the SEs for its relationships with other variables will be downwardly-based (so effects will look more precise and more significant than they should be)
   - This happens whenever we use only 1 observed trait value per person, **because a trait is essentially a missing value of a predictor variable**

* *Thanks to Jonathan Templin for helping me enumerate these problems*

Option 1: Single-Indicator Models

- If you have determined that a single latent factor fits a set of indicators, an option is a "single-indicator" (ASU) factor model

- Assuming perfect reliability (Omega=1) would look like this:
  - **`Factor BY subscale@1; subscale@0; Factor*;`**

- Better: Use **Omega reliability** as estimated from *your* data:
  - Omega: $\boldsymbol{\omega = \mathbf{Var(F)} * (\boldsymbol{\Sigma\lambda})^2 / [\mathbf{Var(F)} * (\boldsymbol{\Sigma\lambda})^2 + \boldsymbol{\Sigma}\,\mathbf{Var(e)} + \mathbf{2\Sigma(e\,cov)}]}$
  - **`Factor BY subscale@1; subscale* (Reliable); Factor*;`**
  - **`MODEL CONSTRAINT: Reliable =`** $\boldsymbol{(1-\omega) * \mathbf{Var(subscale)};}$
  - Subscale residual variance is then the "unreliable variance" only
  - Note: this is not possible if using IRT/IFA factors (reliability varies over trait)

- Either way, the factor can be "centered" by fixing its mean = 0:
  - **`[subscale*]; [Factor@0];`**

Option 1: Single-Indicator Models

- Problems with using **a sum score** in a single indicator model (or as an observed variable in an analysis more generally):

1. Assuming unidimensionality and tau-equivalence (equal discrimination) of indicators within a single sum score
   - **YEP, this is a definitely a problem.**

2. Assuming perfect reliability of observed variables
   - **This is a problem unless correcting for the omega reliability of the sum score (only possible when using CFA).**

3. Assuming each person's trait estimate is perfectly known
   - **YEP, this is a definitely a problem when using only one number to represent the trait level of each person.**

Option 2: Parceling Indicators

- **Parceling = ASU for only some of the indicators**
- For example, for a factor with 12 original indicators:
  - ParcelA = i1+i2+i3+i4, ParcelB = i5+i6+i7+i8, ParcelC = i9+i10+i11+i12
  - **`Factor BY ParcelA* ParcelB* ParcelC*; Factor@1; [Factor@0];`**

- **Guess what happens to model fit???**
  - Total possible DF for actual 12 indicators = $\frac{12(12+1)}{2} + 12 = 90$
  - Estimated DF for actual 12 indicators = $12\lambda+12\mu+12\sigma_e^2 = 36$
  - Remaining DF leftover = 90 – 36 = **54 = lots of room for misfit**
  - Total possible DF for 3 "parcels" = $\frac{3(3+1)}{2} + 3 = 9$
  - Estimated DF for 3 "parcels" = $3\lambda+3\mu+3\sigma_e^2 = 9$
  - Remaining DF leftover = 9 – 9 = **0 = fit is "perfect" (just-identified)**

Option 2: Parceling Indicators

- Contrary to what others may say… **PARCELING IS TOTALLY CHEATING AND YOU SHOULD NOT DO IT**

- That being said, here's how to parcel responsibly if you must:
  - Recognize that **parceling assumes tau-equivalence** (equal loadings) of the indicators within each parcel, so **test that ahead of time**
  - If tau-equivalence (a Rasch-type model) holds, then you aren't losing information (or cheating model fit) by combining the item responses
  - **Be honest** that parceling is an intermediate choice between:
    - ASU completely (single-indicator model for a construct)
    - ASU sort of (parceling only some of the indicators together)
    - An actual indicator-specific measurement model that reflects *all* the data
  - Recognize that different combinations of indicators to parcels can create very different results (especially for "subscales" of subscales), and **do NOT use parcels as a way to "control for" or HIDE misfit**

Option 2: Parceling Indicators

- Problems with **using parcels** rather than the original indicators (aside from an invalid assessment of model fit):

1. Assuming unidimensionality and tau-equivalence (equal discrimination) of indicators within a single parcel
   - **YEP, this is a definitely a problem (unless verified ahead of time).**

2. Assuming perfect reliability of observed variables
   - **The parcel is not assumed completely reliable, but the reliability across parcels is likely to be too optimistic (hidden error within).**

3. Assuming each person's trait estimate is perfectly known
   - **This is not a problem if the latent variable is retained in the model, but we are assuming perfectly known parcel-level scores.**

Option 3: Can I just use the factor scores?

- **In a word, NO. (Try not to, at least.)**
- Factor score = random effect = mean of a person's *unobserved* latent variable distribution given the observed responses
- Because this is a latent variable, each factor score really has a **distribution of possible values** for each person
  - Factor scores are estimated from a multivariate normal prior distribution, and thus will be **shrunken** (pushed to normal) given low reliability
  - There is likely much uncertainty per person, especially for few indicators
    - Although factor scores (thetas) are routinely used in IRT, it's because they are usually based on *dozens* of items per factor (→ small SE)

- Btw, you CANNOT create factor scores by using the loadings as such:
  - $\mathbf{F} = \boldsymbol{\lambda}_{11}\mathbf{y}_1 + \boldsymbol{\lambda}_{21}\mathbf{y}_2 + \boldsymbol{\lambda}_{21}\mathbf{y}_3$ … → Is a COMPONENT model, not a FACTOR model

Option 3: Single Factor Scores

- A factor score is an **observed variable** (just like a sum score is)
- Assuming perfect factor score reliability would look like this:
  - **`Factor BY fscore@1; fscore@0; Factor*;`**

- Better: In CFA (but not IRT/IFA in which reliability varies across the trait), you can use **factor score reliability** estimated from *your* data (true trait differences relative to total trait variance):
  - Factor score reliability $\boldsymbol{\rho} = \frac{\boldsymbol{\sigma_F^2}}{\boldsymbol{\sigma_F^2 + SE_F^2}}$

    $\boldsymbol{\sigma_F^2}$ = **factor** variance (not factor scores)
    $\boldsymbol{SE_F^2}$ = error variance of factor scores
  - **`Factor BY fscore@1; fscore* (Reliable); Factor*;`**
  - **`MODEL CONSTRAINT:`**
    `Reliable =` $\boldsymbol{(1-\rho)*(\sigma_{FS}^2 + SE_F^2);}$

    $\boldsymbol{\sigma_{FS}^2}$ = variance of factor **scores**
    $\boldsymbol{SE_F^2}$ = error variance of factor scores
  - Note this is NOT the same thing as Omega reliability for sum scores, and it's still not possible to do if using IRT/IFA factors (reliability varies over trait)

- Either way, the factor can be "centered" by fixing its mean = 0:
  - **`[fscore*]; [Factor@0];`**

Example: Estimating Reliability

```
! Model 4 -- Fully Z-Scored 2-Factor Model with all parameters labeled for reference
    SitP BY Sit2*  Sit4*  Sit6*  (L1-L3);      ! SitP loadings (all free)
    SitN BY Sit1r* Sit3r* Sit5r* (L4-L6);      ! SitN loadings (all free)
    [Sit2*  Sit4*  Sit6*]   (I1-I3);           ! SitP intercepts (all free)
    [Sit1r* Sit3r* Sit5r*]  (I4-I6);           ! SitN intercepts (all free)
    Sit2*   Sit4*  Sit6*    (E1-E3);           ! SitP residual variances (all free)
    Sit1r*  Sit3r* Sit5r*   (E4-E6);           ! SitN residual variances (all free)
    SitP@1 (VarP); SitN@1   (VarN);            ! Factor variances (fixed=1)
    SitP WITH SitN*  (FactCov);                ! Factor covariance (free)
    [SitP@0 SitN@0]  (MeanP MeanN);            ! Factor means (fixed=0)
```

```
MODEL CONSTRAINT:       ! Calculate omega model-based reliability per factor
    NEW(OmegaP OmegaN); ! Using 1 as placeholder for factor variances
    OmegaP = (1*(L1+L2+L3)**2) / ((1*(L1+L2+L3)**2) + (E1+E2+E3));
    OmegaN = (1*(L4+L5+L6)**2) / ((1*(L4+L5+L6)**2) + (E4+E5+E6));
```

Omega Reliability for *Sum* Scores

New/Additional Parameters

OMEGAP	**0.744**	**0.020**	**37.956**	**0.000**
OMEGAN	**0.775**	**0.014**	**56.803**	**0.000**

```
SAMPLE STATISTICS FOR ESTIMATED FACTOR SCORES
     SAMPLE STATISTICS
           Means
              SITP          SITP_SE       SITN          SITN_SE
              ________      ________      ________      ________
 1              0.000         0.472         0.000         0.418
           Covariances
              SITP          SITP_SE       SITN          SITN_SE
              ________      ________      ________      ________
 SITP           0.777
 SITP_SE        0.000         0.000
 SITN           0.533         0.000         0.825
 SITN_SE        0.000         0.000         0.000         0.000
```

***Factor Score* Reliability (proportion of <u>true</u> individual differences)**

SitP: $\rho = \frac{1}{1+.472^2} = .818$

SitN: $\rho = \frac{1}{1+.418^2} = .851$

Option 3: Single Factor Scores

- Problems with a **single factor score** as an observed variable:

1. Assuming unidimensionality and tau-equivalence (equal discrimination) of indicators within a single sum score
   - **These should be tested first. Unidimensionality should hold, but tau-equivalence doesn't have to (then just let the loadings vary).**

2. Assuming perfect reliability of observed variables
   - **This is not a problem, but factor score unreliability may still create downward bias for relationships with the factor score.**

3. Assuming each person's trait estimate is perfectly known
   - **YEP, this is a definitely a problem when using only one number to represent the trait level of each person.**

Option 4: Multiple Plausible Values

- Using a single factor score instead of a sum score can fix:
  - Assuming (without testing) unidimensionality and tau-equivalence
  - Assuming perfect reliability (can correct using factor score reliability)

- But **uncertainty in the factor scores** is still a problem…
- A potential solution: ***Multiple*** **plausible factor score values**
  - An intermediate option between full SEM and single trait estimates
  - Generate x draws from a person's factor score *distribution*, save those draws to separate datasets, analyze each dataset, then combine results using procedures and rules for multiple imputation of missing data
  - That way the uncertainty of factor scores per person is still represented, along with the factor model parameters that distinguish the indicators
  - This option CAN be used if using IRT/IFA
  - Mplus now provides this using a 4-step process

Plausible Values, Step by Step

- **Step 1: Estimate factor model** using ML/MLR, save syntax for estimated parameters as start values (use OUTPUT: SVALUES to save typing)
- **Step 2: Feed in estimated parameters** as fixed parameters (replace all ***** with **@**), re-estimate model using ESTIMATOR=BAYES to generate the factor score draws for each person and save to separate data sets
  - Could do BAYES estimation for all of it, but if you have been using ML/MLR, you should use those parameters instead of letting it find new ones
- **Step 3: Merge separate datasets together** to create x complete datasets for analysis (see my SAS macro as part of Example 10 to make this easier)
- **Step 4:** Tell Mplus to estimate your model **using the factor scores as observed variables on each of the x datasets**, and to combine the results (TYPE = IMPUTATION)
  - Will be easier and go faster than analyses of the original latent variables, but still preserves the uncertainty in the factor score estimates per person, along with the factor model from which those factor scores were derived

SEM: My Big Picture

- **SEM is great *when you can do it***
  - Provides a means to make almost any idea an empirical question
  - Measurement models create latent constructs (= random effects)
  - Structural models test relations among those constructs
  - Do not let your measurement model swamp structural relations tests by looking only at global fit: consider what the baseline model should be

- **SEM is not a panacea for everything**
  - IT WILL BREAK when your models get too complicated (or realistic)
  - You may have named your factors, but it doesn't mean you are right!
  - Distributional assumptions matter, but so do linear model assumptions (nonlinear measurement and structural models may be needed)
  - Factor scores are not real things (and neither are sum scores), so make sure to represent their uncertainty in any SEM alternative