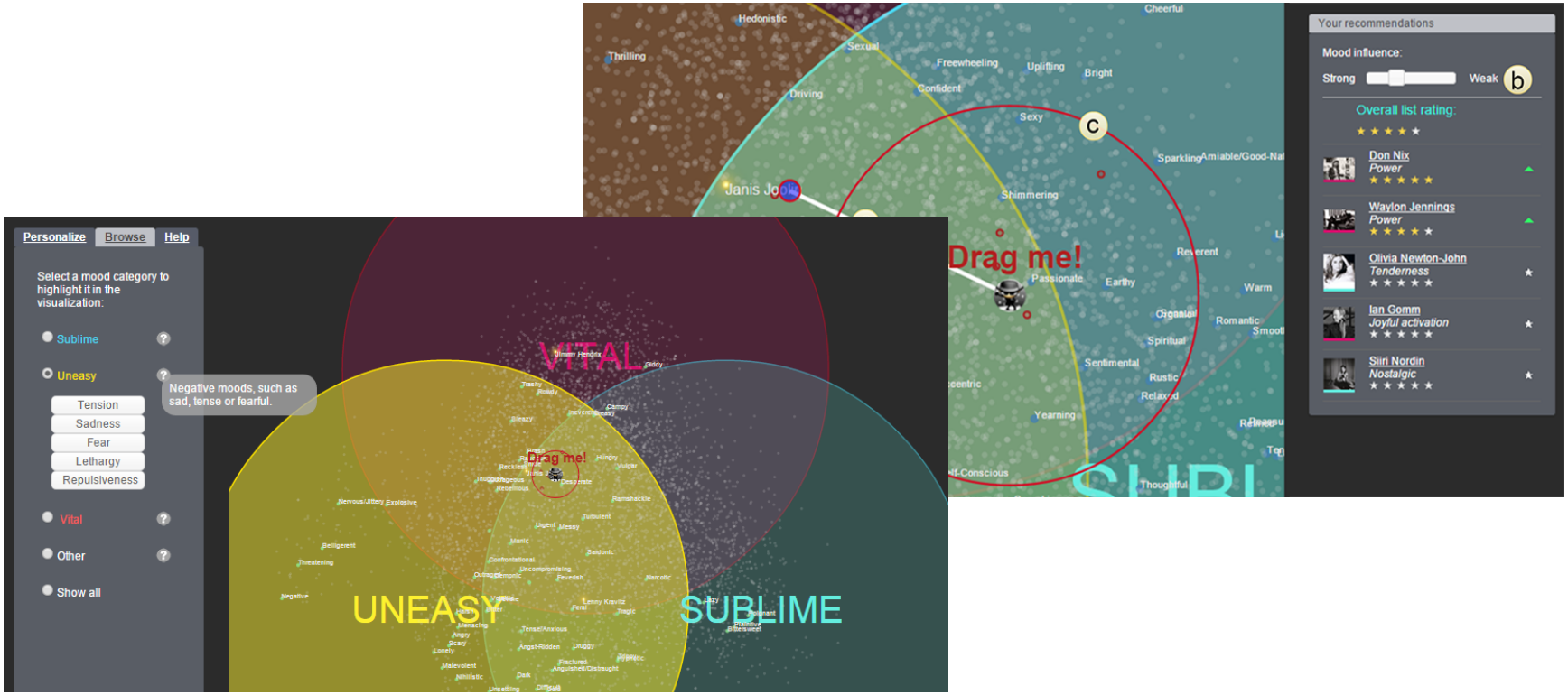
Factor Analysis & Structural Equation Models

CS185 Human Computer Interaction

MoodPlay Recommender (Andjelkovic et al, UMAP 2016)

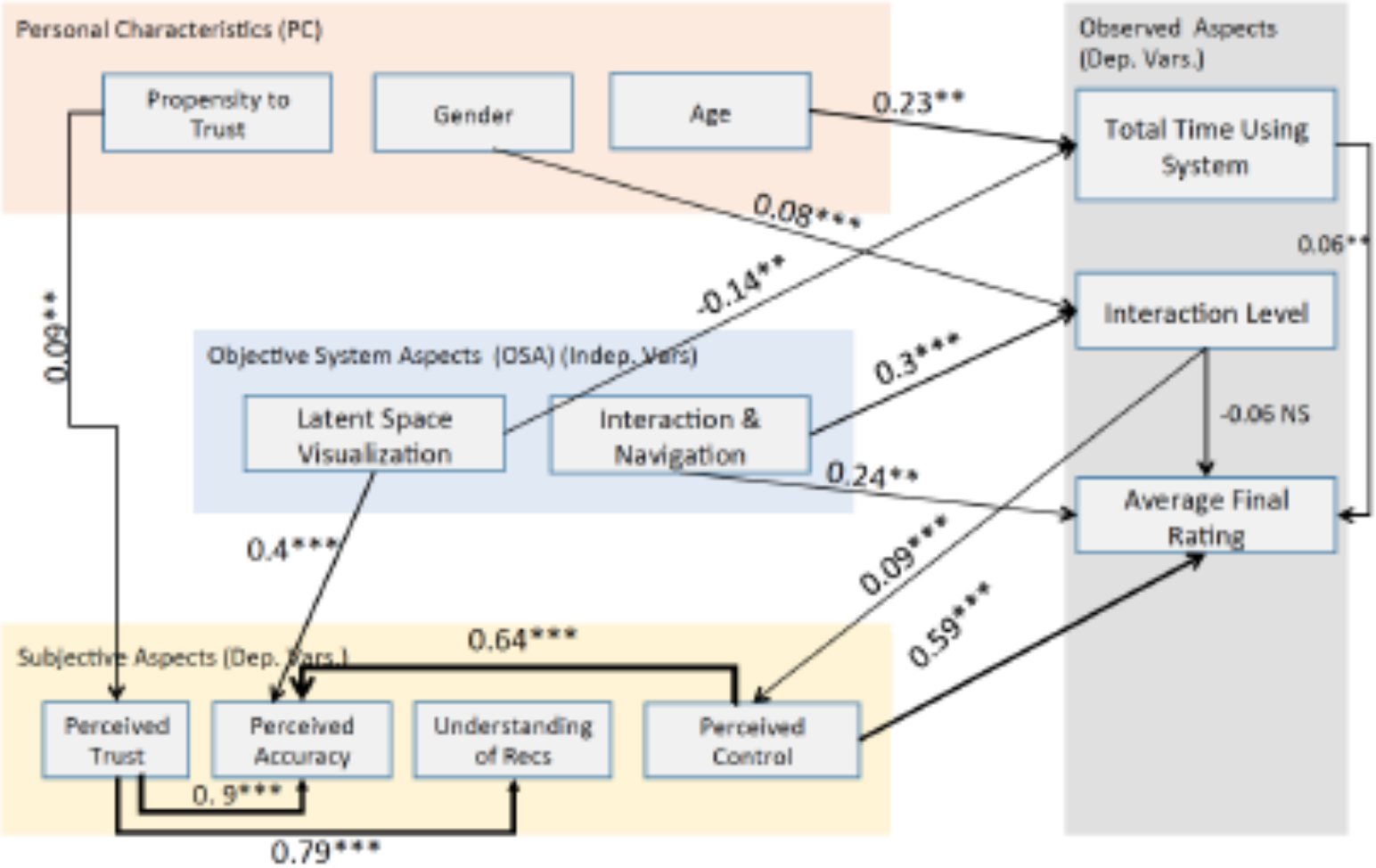


Online system available here:
http://ugallery.pythonanywhere.com/

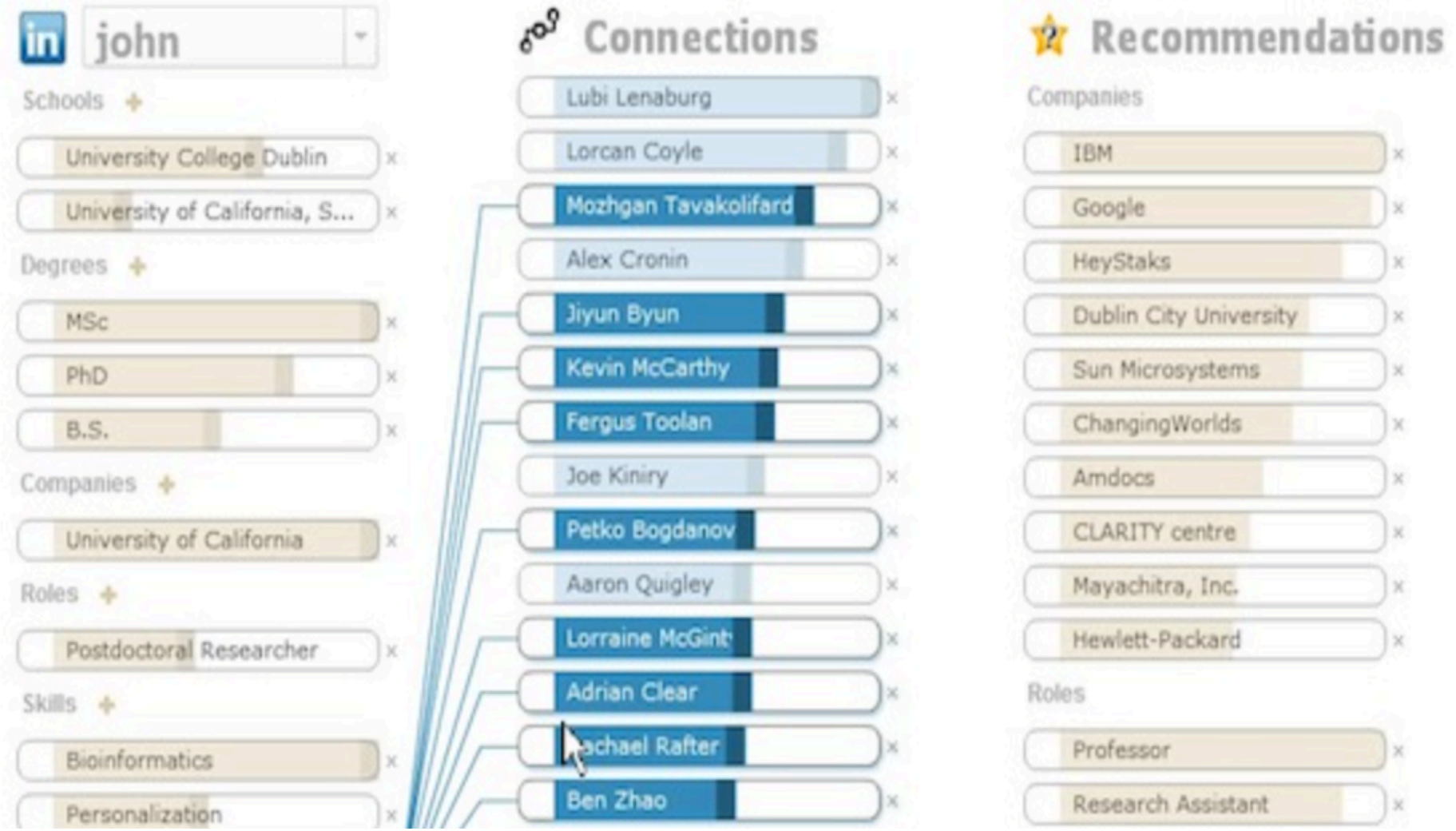
MoodPlay

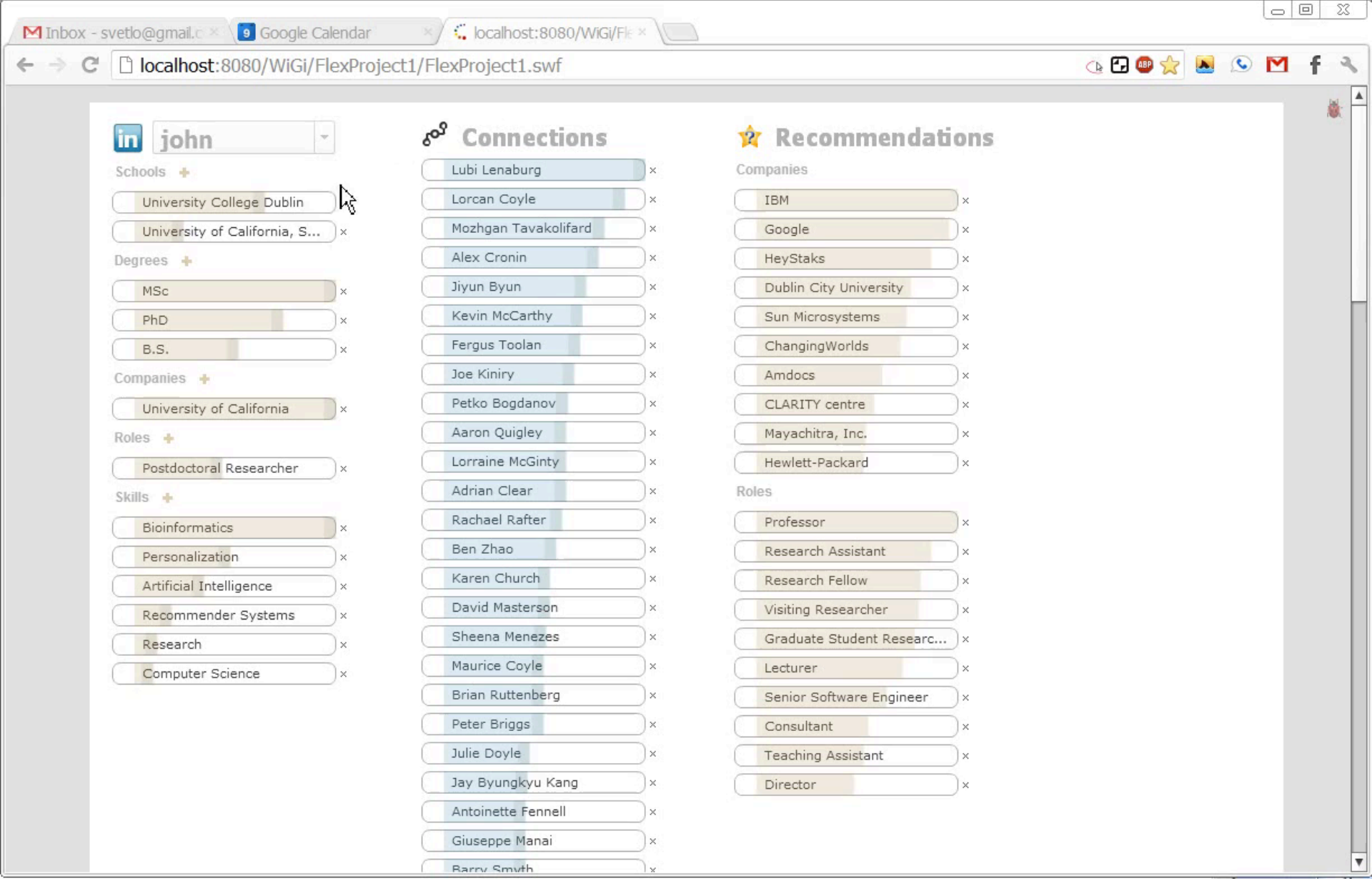
Music recommendation system that enables exploration and discovery of new artists through an interactive mood space.

Structural Equation Analysis

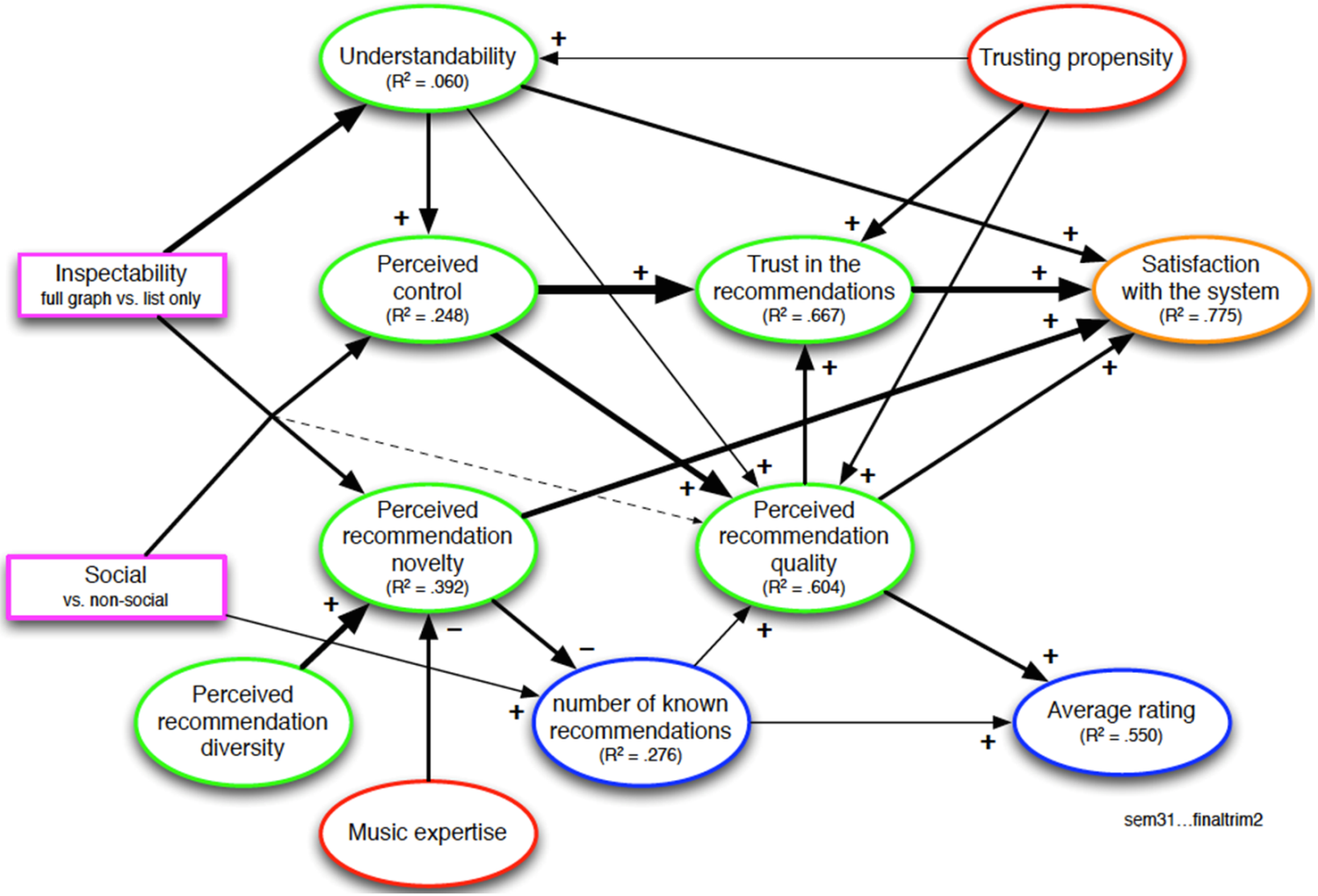


Career recommendations – LinkedVis




Inbox - svetlo@gmail.c
Google Calendar
localhost:8080/WiGi/Fl
localhost:8080/WiGi/FlexProject1/FlexProject1.swf
john
Schools
University College Dublin
University of California, S...
Degrees
MSc
PhD
B.S.
Companies
University of California
Roles
Postdoctoral Researcher
Skills
Bioinformatics
Personalization
Artificial Intelligence
Recommender Systems
Research
Computer Science
Connections
Lubi Lenaburg
Lorcan Coyle
Mozhgan Tavakolifard
Alex Cronin
Jiyun Byun
Kevin McCarthy
Fergus Toolan
Joe Kiniry
Petko Bogdanov
Aaron Quigley
Lorraine McGinty
Adrian Clear
Rachael Rafter
Ben Zhao
Karen Church
David Masterson
Sheena Menezes
Maurice Coyle
Brian Ruttenberg
Peter Briggs
Julie Doyle
Jay Byungkyu Kang
Antoinette Fennell
Giuseppe Manai
Barry Smyth
Recommendations
Companies
IBM
Google
HeyStaks
Dublin City University
Sun Microsystems
ChangingWorlds
Amdocs
CLARITY centre
Mayachitra, Inc.
Hewlett-Packard
Roles
Professor
Research Assistant
Research Fellow
Visiting Researcher
Graduate Student Researc...
Lecturer
Senior Software Engineer
Consultant
Teaching Assistant
Director

Experimentation and Results



Factor Analysis

- Factor analysis is an exploratory tool
  - Helps identify simple patterns that underlie complex multivariate data
    - Not about hypothesis testing
    - Rather, it is more like data mining
  - And also helps us understand some principles of SEM

Factor Analysis

- Things you can do with factor analysis:
  - 1. Examine factor loadings
    - Use them to interpret factors that are identified in the data
  - 2. Plot factor loadings
    - Vividly describe which variables "go together" (people score high on one tend to score high on another or vice versa)
  - 3. Compute factor scores
    - Estimate how individual cases score on underlying factors
    - How depressed is each case?
  - 4. Determine variation explained by factors
    - See which factors account for the major patterns in your data
  - 5. "Rotate" the factors
    - Modify them to enhance interpretability… Will discuss later.

EFA: Civic Participation

- Factor loadings describe patterns in data
  - A powerful exploratory tool

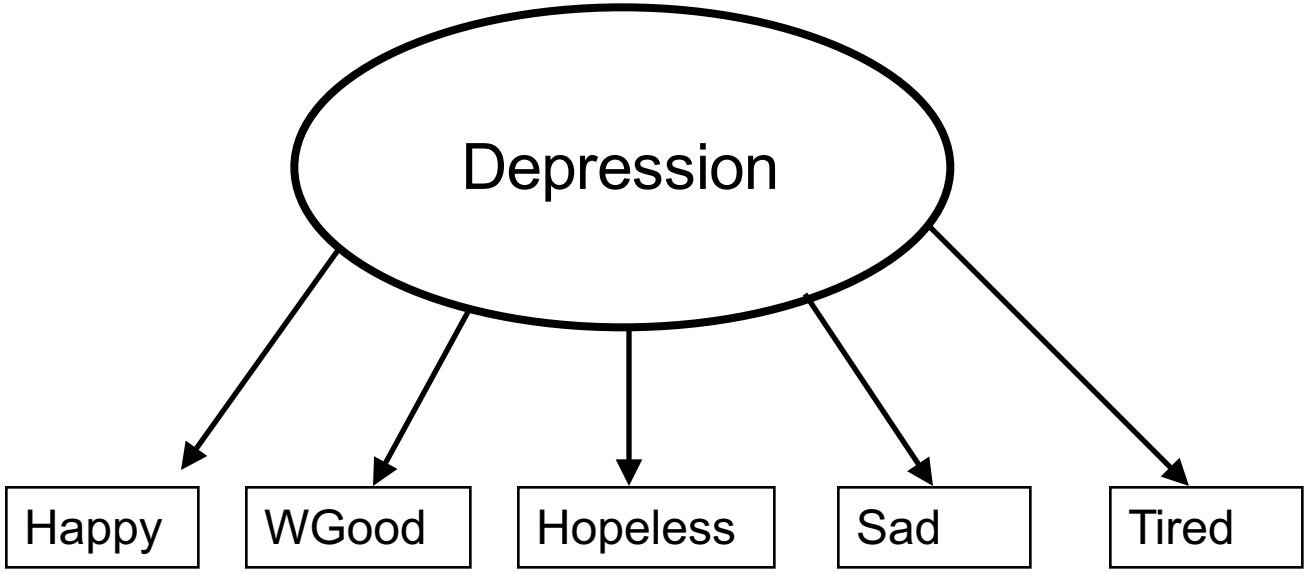
Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
member	0.8061	0.0974	0.0139	0.3405
volunteer	0.8055	0.0377	-0.0087	0.3497
petition	0.0615	0.3130	-0.1456	0.8771
boycott	0.1504	0.5724	0.0165	0.6494
demonstrate	0.1358	0.5614	0.0671	0.6619
strike	0.0371	0.3536	0.2421	0.8150
occupybldg	-0.0030	0.2439	0.2501	0.8780

Here, we see a clearer pattern… Factors 1 & 2 are more distinct.
Factor 1 = civic membership; factor 2 = protest/social mvmts, etc…

Confirmatory Factor Analysis

- Factor analysis is purely exploratory
  - It is data mining, not a model
  - However, it is based on the idea that factors – which are unobserved – give rise to (i.e., cause) variation on observed variables

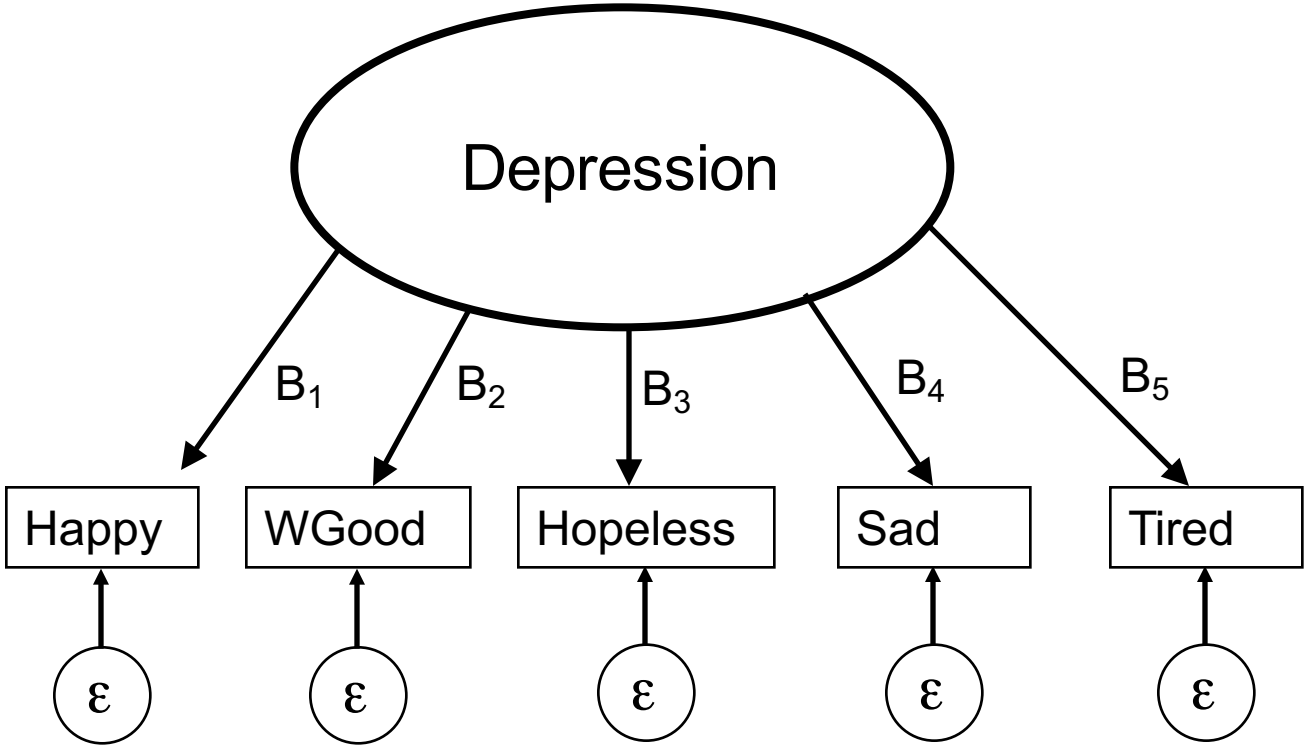


Confirmatory Factor Analysis

- Idea: Let's imagine that depression is a **latent variable**
  - i.e., a variable we can't directly measure… but gives rise to observed patterns in things we can observe
  - Note: No observed variable perfectly measures the latent variable
    - Each observable variable is a measure… but there is error
    - Observed variables aren't perfectly correlated with latent variable (even though they are "caused" by it)…

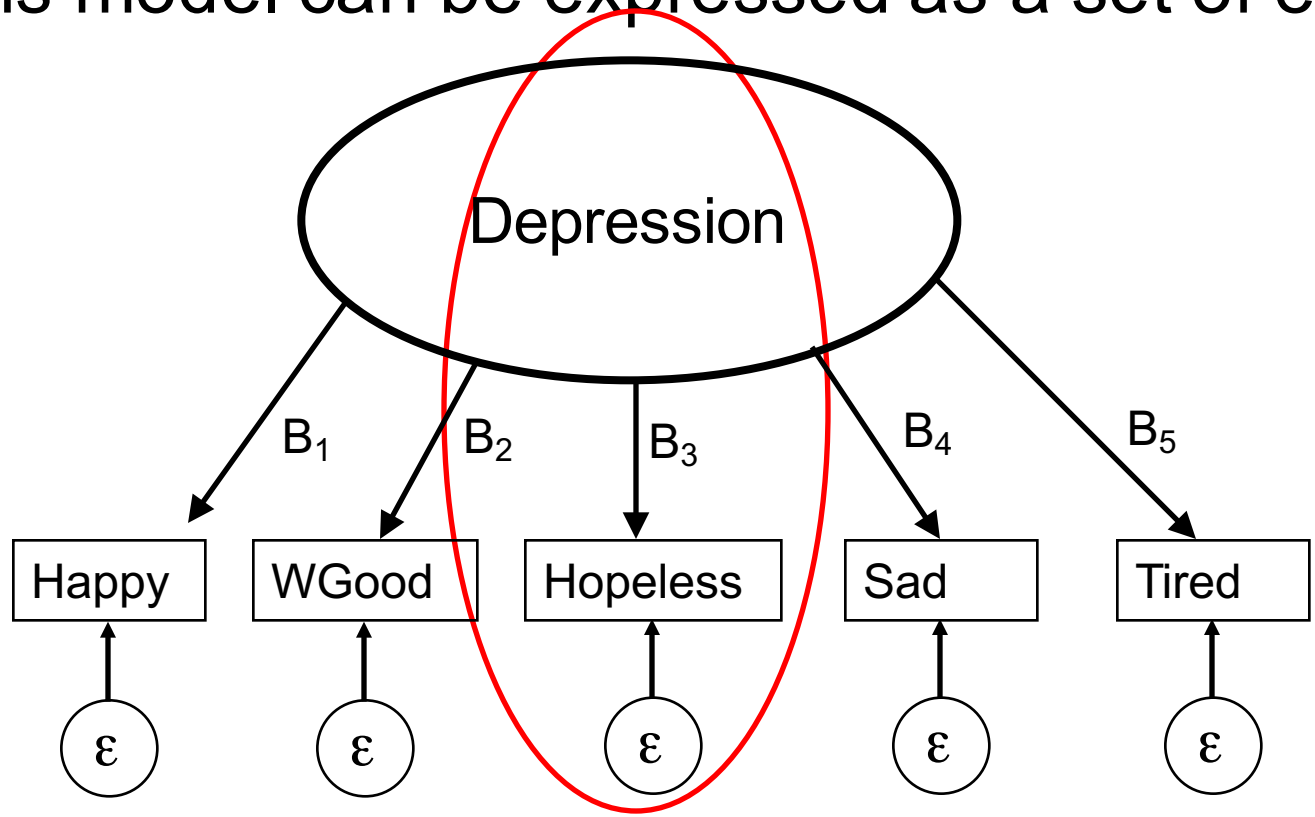
Confirmatory Factor Analysis

- This forms the basis for a kind of model:



Confirmatory Factor Analysis

- This model can be expressed as a set of equations:



- Hopeless = B_3Depression + ε

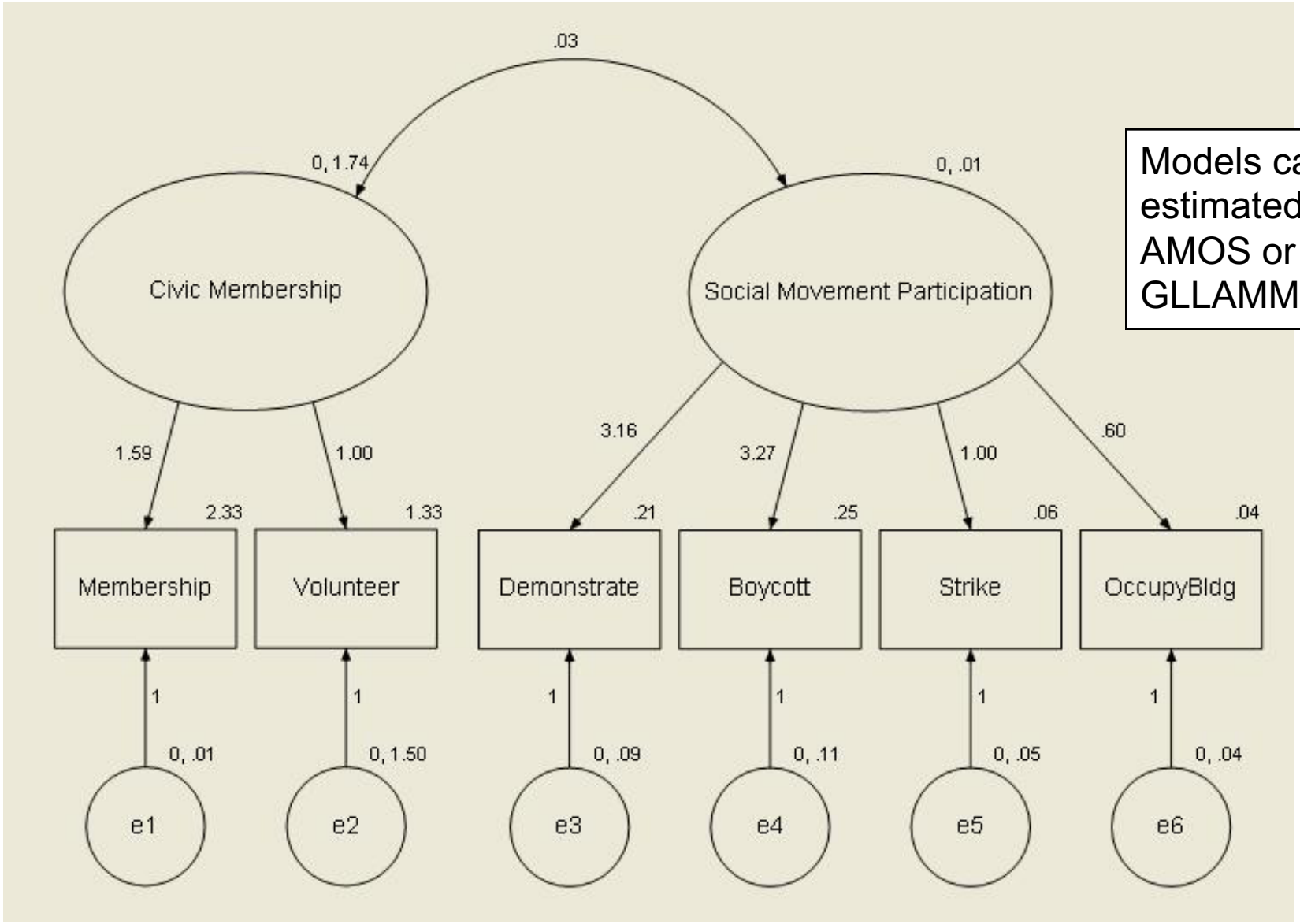
Confirmatory Factor Analysis

- Full set of Equations:
  - $\text{Happy} = B_1\text{Depression} + \varepsilon$
  - $\text{WorldGood} = B_2\text{Depression} + \varepsilon$
  - $\text{Hopeless} = B_3\text{Depression} + \varepsilon$
  - $\text{Sad} = B_4\text{Depression} + \varepsilon$
  - $\text{Tired} = B_5\text{Depression} + \varepsilon$

Confirmatory Factor Analysis

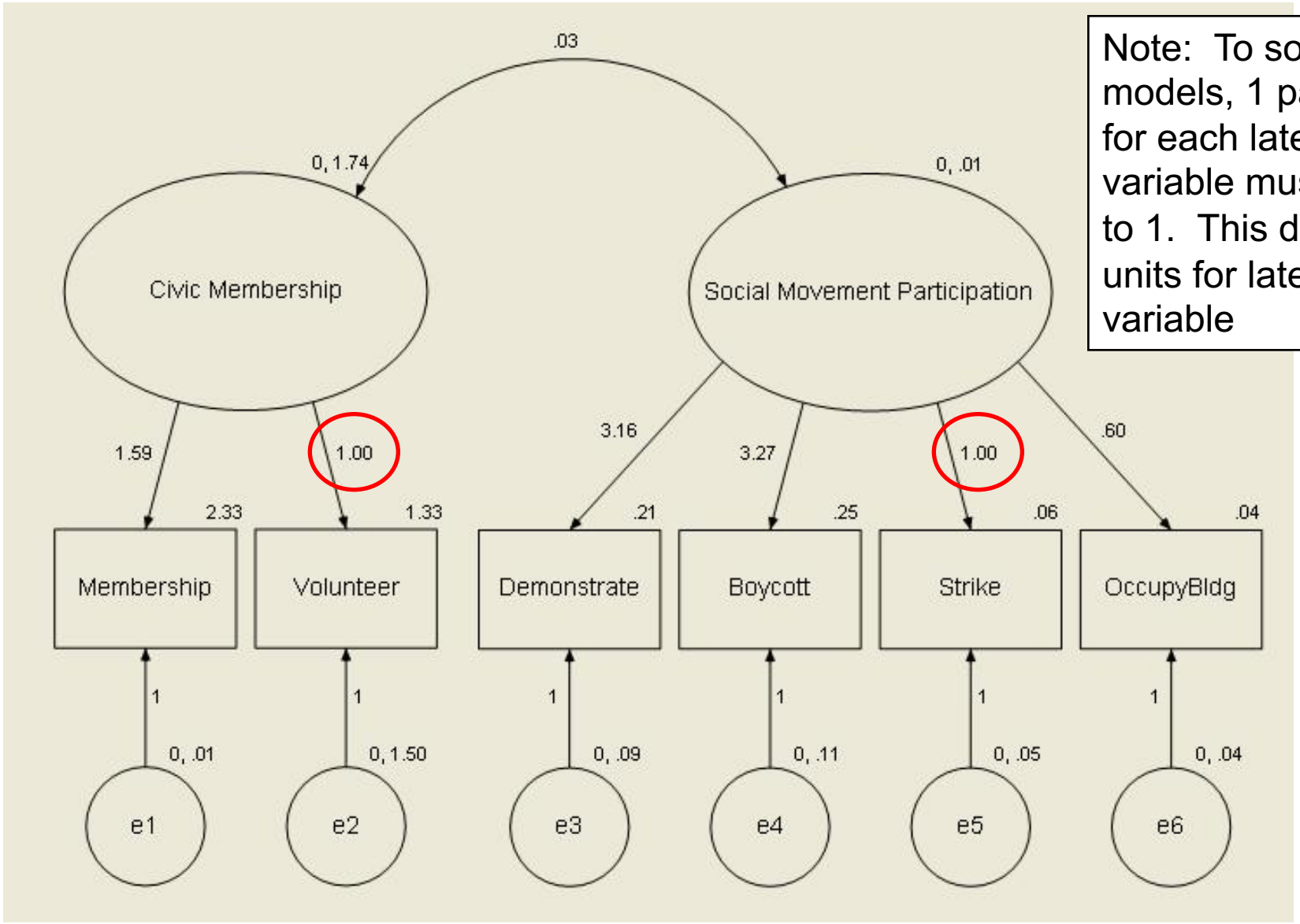
- Idea: We can model real data based on those presumed relationships…
  - Estimate slope coefficients for each arrow
    - How do latent variables affect observed variables?
  - Examine overall model fit
    - How much does our theoretically-informed view of the world map onto observed data?
    - If model fits well, our concept of "depression" (and measurement strategy) are likely to be good
  - "Confirmatory" implies that we aren't just "exploring"
    - Different from "exploratory factor analysis"…
    - Rather than data mining, we're testing a theoretically-informed model.

CFA: Civic Engagement



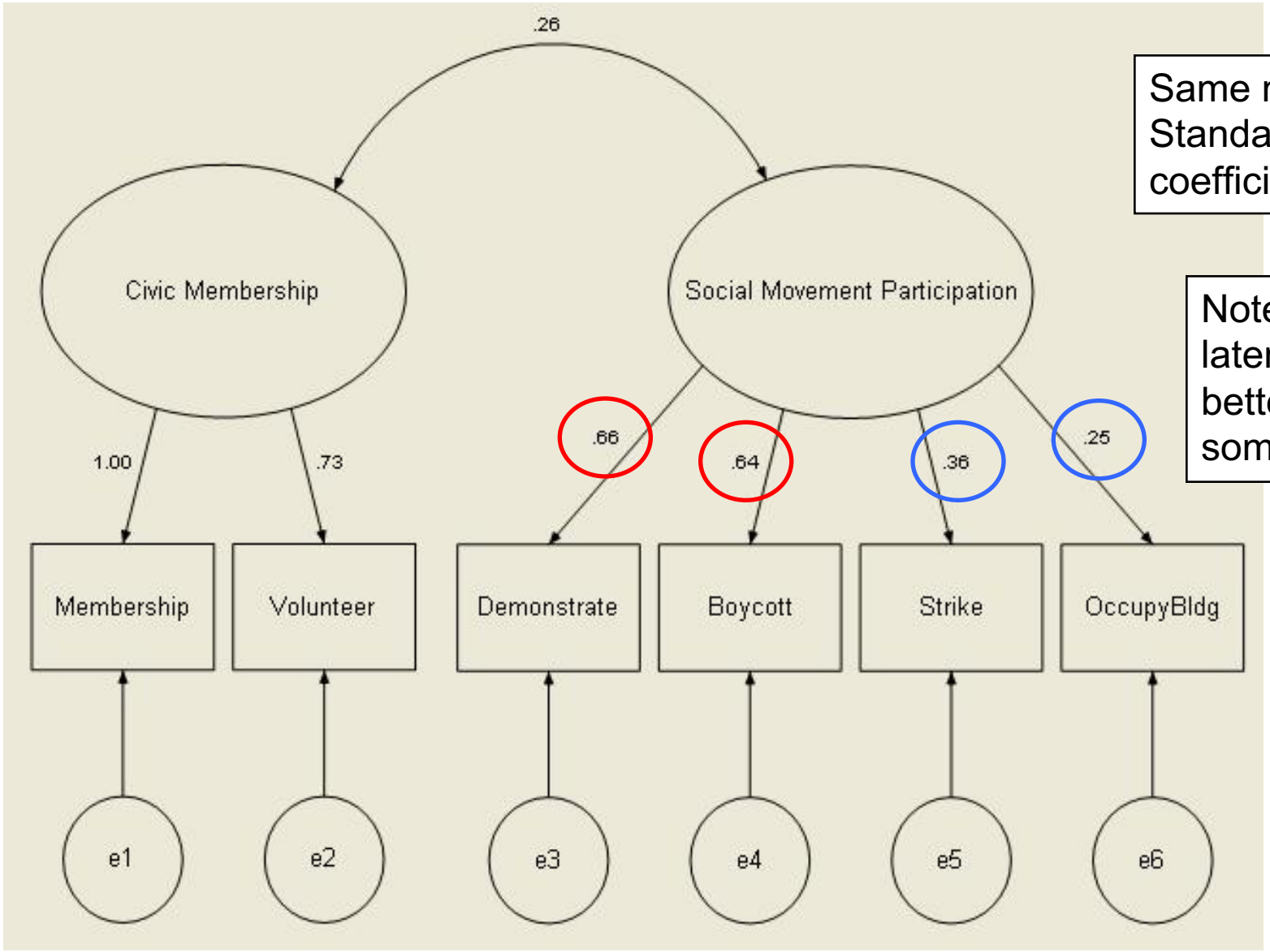
Models can be estimated with AMOS or GLLAMM

CFA: Civic Engagement



Note: To solve models, 1 parameter for each latent variable must be set to 1. This defines units for latent variable

CFA: Civic Engagment



Same model: Standardized coefficients…

Note that the latent variable better predicts some vars…

CFA: Text Output

Slopes			Estimate	S.E.	C.R.	P
Volunteer	<---	Civic Membership	1.000			
Member	<---	Civic Membership	1.588	.211	7.517	***
Strike	<---	Social Movement Participation	1.000			
Boycott	<---	Social Movement Participation	3.270	.386	8.473	***
Demonstrate	<---	Social Movement Participation	3.165	.376	8.406	***
OccupyBldg	<---	Social Movement Participation	.596	.105	5.694	***

Intercepts	Estimate	S.E.	C.R.	P
Volunteer	1.333	.052	25.639	***
Member	2.328	.060	38.506	***
Strike	.058	.007	8.517	***
Boycott	.248	.013	19.691	***
Demonstrate	.207	.012	17.552	***
OccupyBldg	.041	.006	7.031	***

CFA: Model Fit

- So, did the model fit?
  - Many strategies to assess fit: Chi-square; "fit indices"
  - Ex: Chi-square test
    - Large Chi-square indicates that data deviate from model expectations
      - e.g., when used to test independence in a crosstab table…
    - If model "fits" well, chi-square will be NON-significant
      - However, this is a sensitive test… if N is large, the model almost always yields a significant Chi-square…

Result (Default model): Civic Participation
N = 1,200
Chi-square = 28.379
Degrees of freedom = 8
Probability level = .000

Low p-value indicates significant difference between model and observed data
(not uncommon for large N model)

Model Fit: NFI

- Another way to assess fit: NFI
  - Also called the Bentler-Bonett index

$$NFI = \frac{\chi^2 null - \chi^2 full}{\chi^2 null}$$

- Where X^2 null is chi-square of null model (independence)
- X^2 full is chi-square of model of interest
- NFI ranges from 0 to 1
- NFI > .9 = OK, NFI > .95 is good.

Model Fit: CFI

- Comparative Fit Index: CFI

$$CFI = \frac{(\chi^2 null - df) - (\chi^2 full - df)}{(\chi^2 null - df)}$$

This image cannot currently be displayed.

  - CFI ranges from 0 to 1
  - CFI > .9 = OK, CFI > .95 is good.

Model Fit: RMSEA

- Root Mean Square Error of Approximation

$$RMSEA = \sqrt{\frac{(\chi^2 / df) - 1}{df - 1}}$$

This image cannot currently be displayed.

  - RMSEA of 0 = perfect fit
  - RMSEA < .05 = good fit
  - RMSEA > .1 = poor fit.

CFA: Civic Engagement

- Model Fit Summary:
  - Results greatly edited… many fit indices reported…

Model	RMSEA	NFI	CFI
Default model	.046	.979	.985
Saturated model		1.000	1.000
Independence model	.231	.000	.000

This image cannot currently be displayed.

Fit indices look pretty good. Not perfect, but OK.

Why Use CFA

- 1. If CFA model fits well, it strongly supports theory underlying the model
  - Poor fitting CFA implies that the latent variables are not empirically present
    - Or don't relate to observed variables in the way we specified
- 2. CFA can be used to compare models
  - Are "petitions" part of "civic membership" or "social movements"? Or both?
  - We can use CFA to assess fit of various models
    - And settle debates about how measures relate to latent variables.

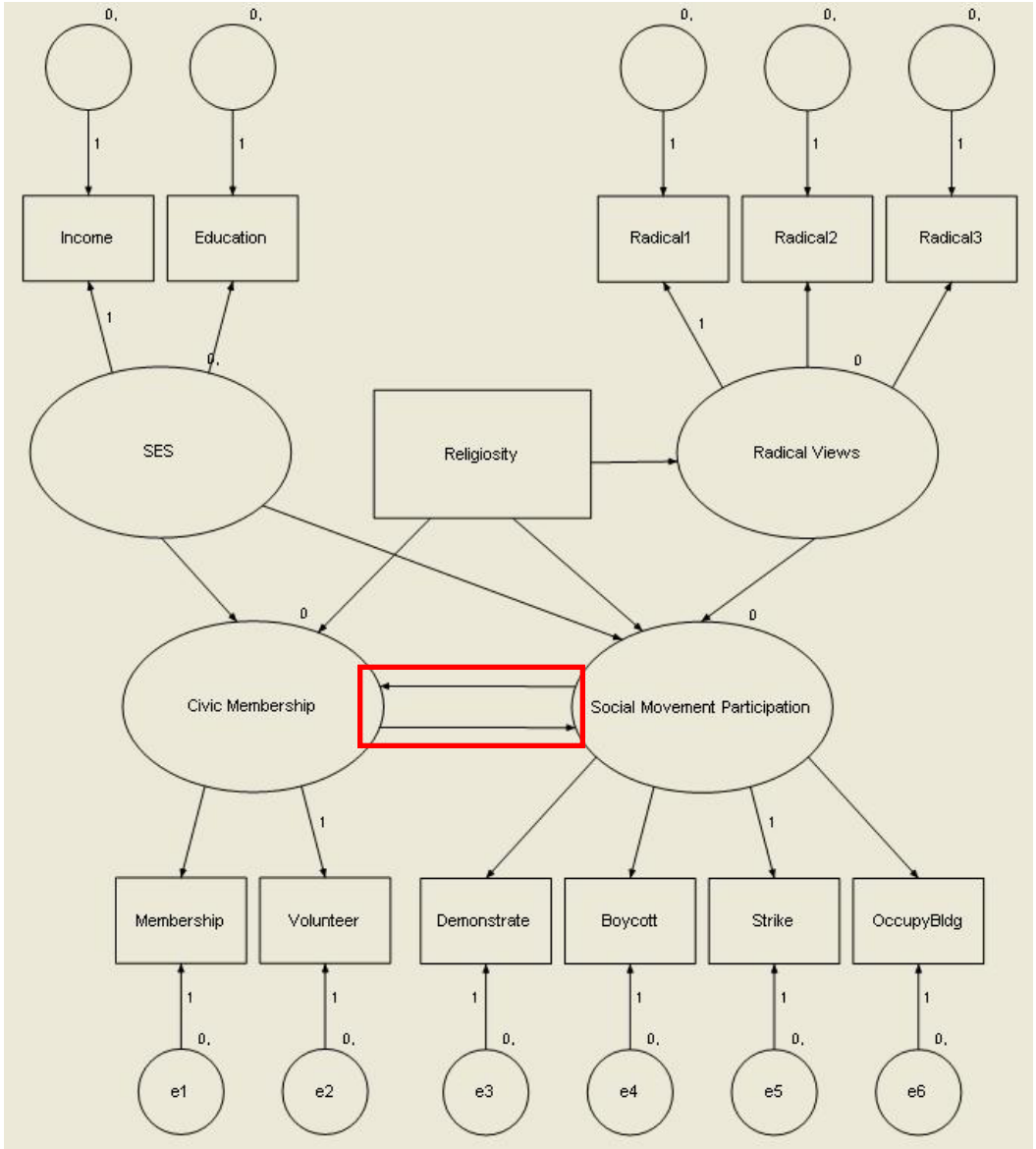
Why Use CFA

- 3. CFA can be used to test applicability of models to different groups
  - Does model for US apply to other countries? Or just to those similar to US (e.g., canada)?
  - Men vs. women… Are patterns of civic life the same?

SEM

- Next step: Structural Equation Models (SEM) with Latent Variables
  - Once we've identified latent variables, it makes sense to analyze them!
  - We can develop models in which we estimate slopes relating latent variables…
  - This is particularly useful when we are interested in latent concepts that are difficult to measure with any single variable.

SEM: Civic Engagement



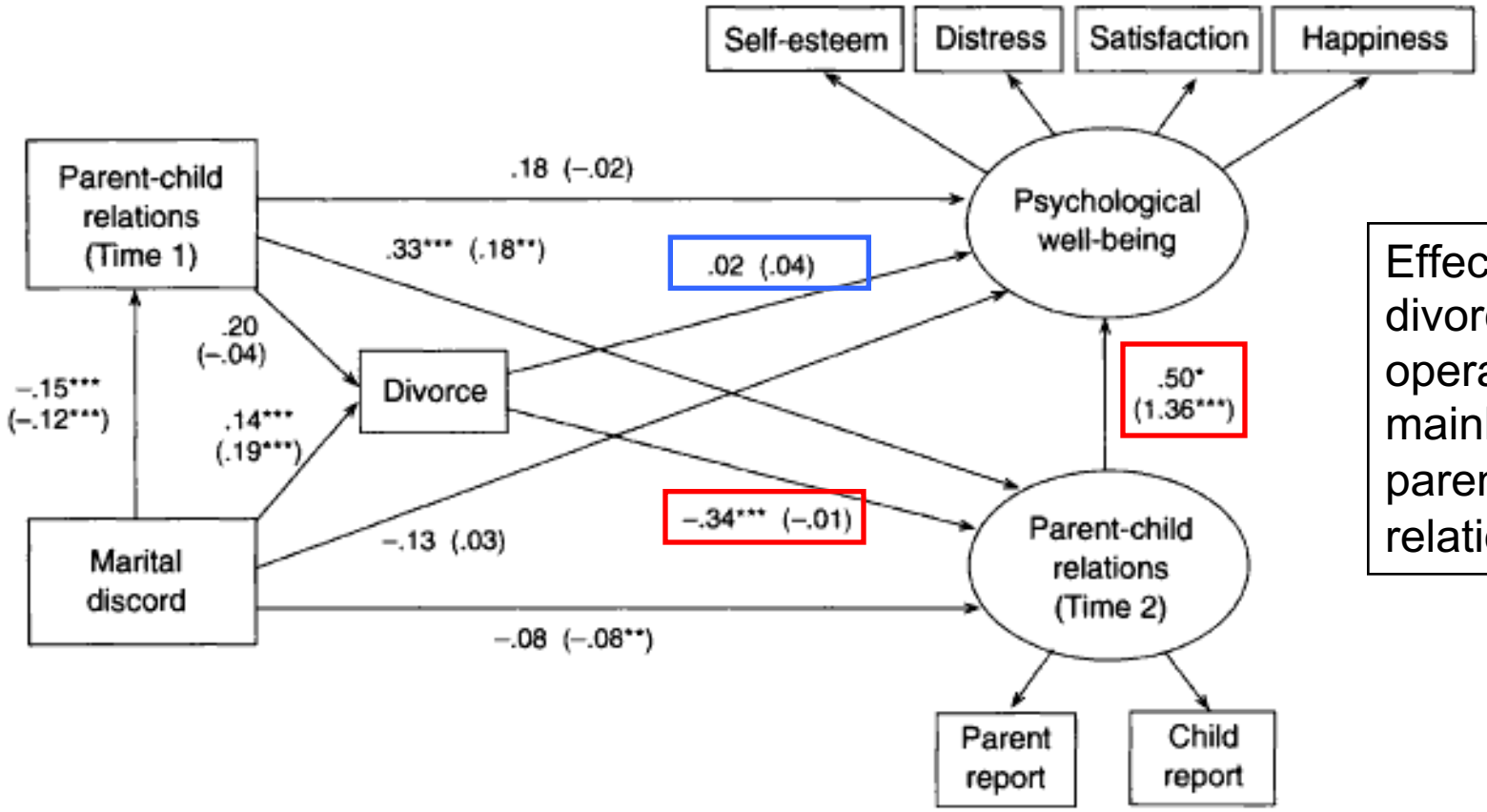
Note that both latent and observed variables can be used to predict outcomes

Also, under some conditions you can estimate non-recursive models (paths in both directions)

SEM: Divorce & Well-being

- Example 2: Amato, Paul R and Julia M. Sobelowski. 2001. The Effects of Divorce and Marital Discord on Adult Children's Psychological Wellbeing. *American Sociological Review*, 66,6:900-921.
  - What is effect of divorce on (adult) children's well-being
  - Answer: Divorce mainly has effects by harming parent/child relationship.

SEM: Divorce & Well-being



Effect of divorce operates mainly via parent-child relations

Figure 3. Unstandardized Coefficients from the Structural Equation Model Showing Links between Marital Discord, Divorce, and Parent-Child Relations at Time 1 and Time 2, and Offspring's Psychological Well-Being

Note: Numbers in parentheses are *b* coefficients for mothers. For fathers, $\chi^2 = 39.8$, d.f. = 34, GFI = .97, CFI = .98, and RMSEA = .03. For mothers, $\chi^2 = 62.2$, d.f. = 34, GFI = .98, CFI = .95, RMSEA = .05.

$^{*}p < .05$ $^{**}p < .01$ $^{***}p < .001$ (two-tailed tests)

Why Use SEM?

- 1. Very useful when you are concerned about measurement error
  - Use of multiple measures for each latent variable can yield robust analyses, despite weakness of each measure
- 2. Similar to path models (discussed in lab), but allows latent variables
  - You can model the relationship between many latent & observed variables at the same time

Why Use SEM?

- 3. Additional information afforded by multiple measures can permit solution of "non-recursive" models
  - i.e., models where two variables have a reciprocal relationship
  - Ex: Self-Esteem ←→ School Achievement
  - If models are well specified, SEM may help tease out complex issues of causality.

Why Use SEM?

- 3. (cont'd) Non-recursive models…
  - Issue: Identification
    - A big topic – can't be covered sufficiently today
    - Obviously, we can't estimate every causal path between vars…
      - Even if we imagine the theoretical possibility of a relationship
    - "Identification" refers to a model that is solveable
    - Models with too many paths = not identified
      - You must simplify the model to allow a solution.

Why Use SEM?

- 4. A powerful tool for formalizing complex theoretical relationships
  - And testing those theories
  - Indeed, many refer to SEM as “causal modeling”
    - The theorist specifies causal paths based on theory, tests those paths…

Problems with SEM

- 1. Model specification issues are even more complex than regression models
  - You are often dealing with MANY paths
  - If any part of the model is mis-specified, it will affect other parts of the model
  - Results are often unstable…
    - Adding a path between two variables can change results a LOT
    - It is easy to produce any desired result by tweaking paths…
- Perhaps not a panacea for determining causality after all…

Problems with SEM

- 2. SEM hasn't been adapted to address many limitations of linear models
  - Generally can't do non-linear models (ex: Poisson)
    - Though software keeps improving. Newest version of AMOS can handle ordered categorical data
  - Not designed to easily handle grouped data
    - E.g., Multi-level models
- 3. Still requires specialized software
  - LISREL, AMOS, EQS
  - Cumbersome – not user friendly.