Advanced Psychometrics Methods Workshop

Longitudinal Data Analysis

Richard N Jones ScD
jones@hsl.harvard.edu

UW Friday Harbor Laboratories  June 12, 2012

Objective

Resources to learn more about LDA and LDA using SEM methods

Introduce some of the concepts and terminology relevant to longitudinal data analysis (LDA)

with special emphasis on applications in cognitive aging research using structural equation modeling (SEM)

Emphasis issues and applications that might be useful with this year's FHL Psychometrics Workshop studies

Acknowledgements

- Funded in part by Grant 5R13AG030995-05 from the National Institute on Aging

- The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government

Friday Harbor Psychometrics Workshop

2012

Session Overview

1. Other resources
2. General Modeling Framework
3. Example
   - Alden's composite in four studies
   - Proc Mixed *vs.* Latent Growth Curve Modeling
4. Survey of applications and approaches
5. Considerations: FHL Psychometrics 2012
6. Questions and discussion

Other Resources

- **What is longitudinal data analysis?**
  - Singer JD Willett JB. Applied longitudinal data analysis: Modeling change and event occurrence. 2003, New York: OUP.
  - Worked examples at UCLA Academic Technology Services web site http://www.ats.ucla.edu/stat/examples/alda.htm
- **How do I do LDA, especially these SEM Approaches?**
  - Newsom J, Jones RN, Hofer S (Eds). Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health and Social Sciences. 2011. New York: Routledge.
  - Duncan TE, Duncan SC, Strycker LA. An introduction to latent variable growth curve modeling: concepts, issues and applications. Second ed. 2006, Mahwah, NJ: LEA, Inc.
- **Tell me more about the math behind latent curve methods**
  - Bollen KA, Curran PJ. Latent curve models: a structural equation perspective. 2006, Hoboken, NJ.: Wiley

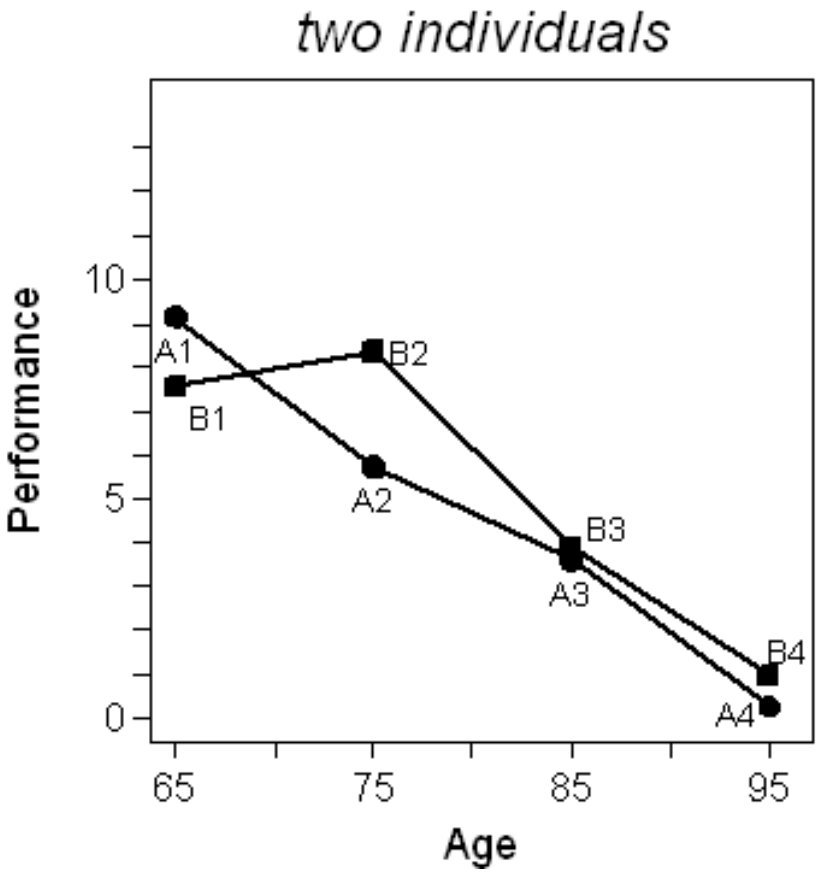
Other Resources

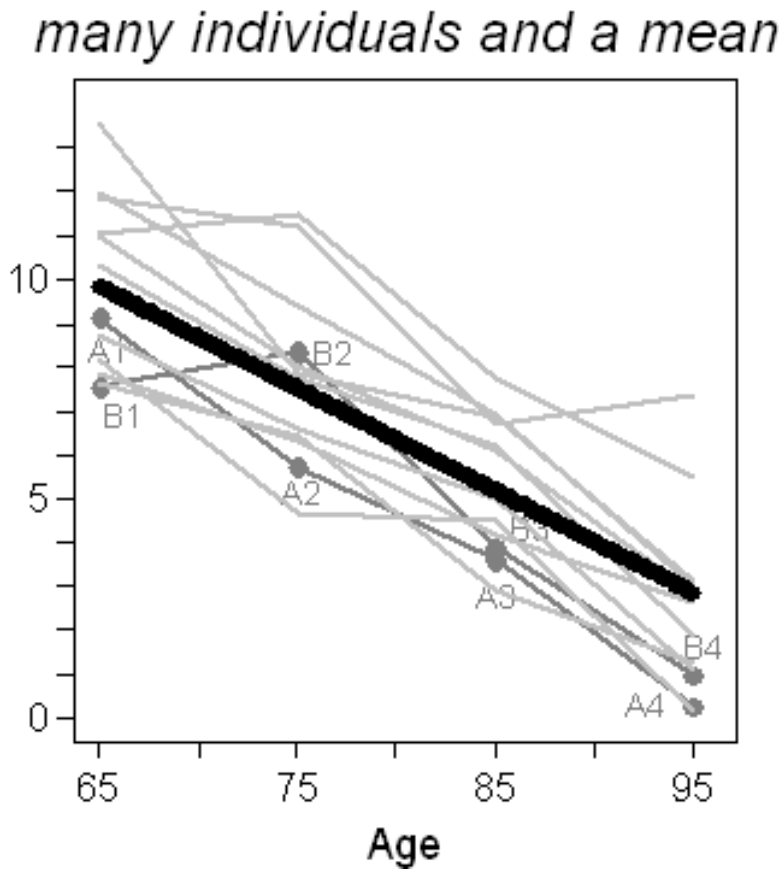
- **Latent Variable Methods Workshop** in Boston
  - Measurement models - even years (e.g., 2012)
  - Longitudinal data analysis - odd years (e.g., 2013)
  - For information
  - For LDA slides, examples (code and data) from 2011
- Other courses listed on the Mplus web site.

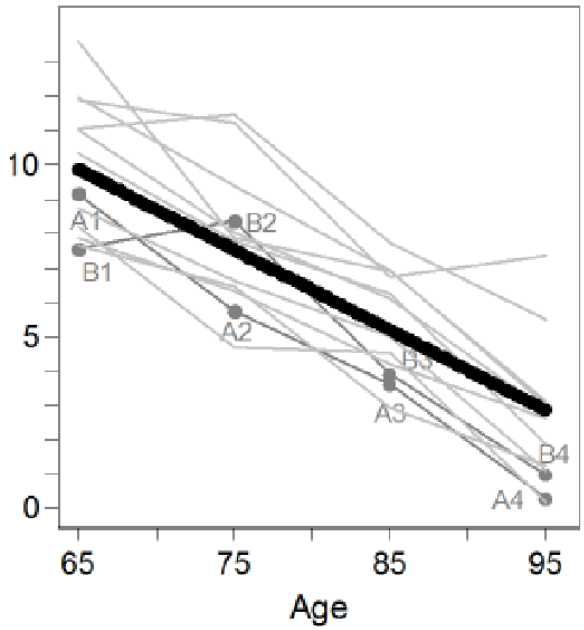
What is Longitudinal Data Analysis (LDA)?

- Analysis of data where observations are repeated or replicated
- More interesting if more than 2 observations (generally)
- Outcomes can be
  - absorbing events
    - ★ e.g., death, conversion to dementia
  - discrete but non-absorbing
    - ★ e.g., conversion to mild cognitive impairment
  - quantitatively distributed
    - ★ e.g., neuropsychological test performance
  - quantities not directly observed but measured indirectly
    - ★ e.g., latent variables

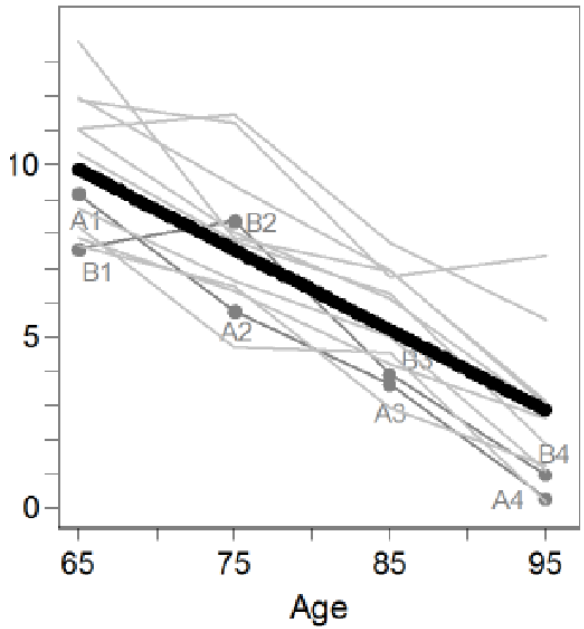
I purposefully do not say anthing about *time* or *age*. But before you conclude this talk is just about *repeated measures data analysis*: rest assured I'll talk plenty about time and age as I go on. But, the point is to emphsize that the *data analysis* techniques presented can be applied in any suitable repeated measures study, regardless of whether the repeated observations are sequenced along a time scale.


two individuals
Performance
10
5
0
A1
B1
B2
A2
B3
A3
B4
A4
65
75
85
95
Age


many individuals and a mean
10
5
0
A1
B1
B2
A2
B3
A3
B4
A4
65
75
85
95
Age



$$y_{ij} = b_{0i} + b_{1i}x_j + e_{ij}$$
$$b_{0i} = a_0 + \zeta_{0i}$$
$$b_{1i} = a_1 + \zeta_{1i}$$
$$i \in [1, N]$$
$$j \in [1, T]$$
$$e \sim N(0, \theta), COV(y, \theta) = 0$$
$$\zeta \sim N(0, \psi), COV(y, \zeta) = 0$$



$$y_{ij} = \eta_{1i} \times 1 \quad + \eta_{2i}\lambda_j + \epsilon_{ij}$$

$$y_{ij} = \eta_{1i} \quad + \eta_{2i}\lambda_j + \epsilon_{ij}$$

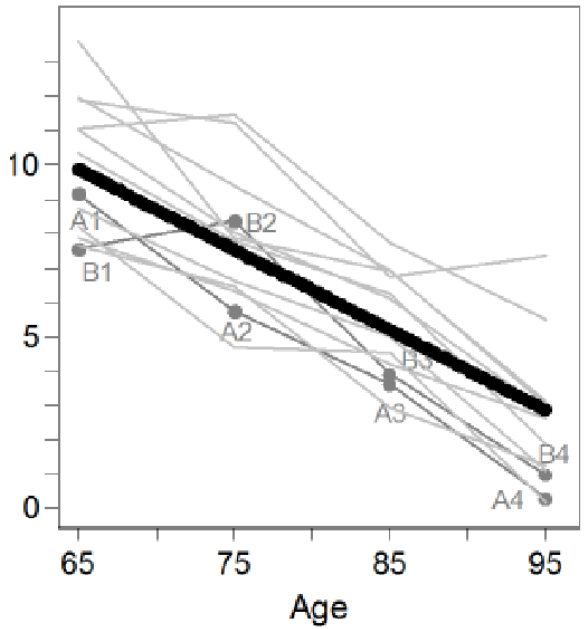
$$\eta_{1i} = \alpha_1 + \zeta_1$$

$$\eta_{2i} = \alpha_2 + \zeta_2$$

$$i \in [1, N]$$

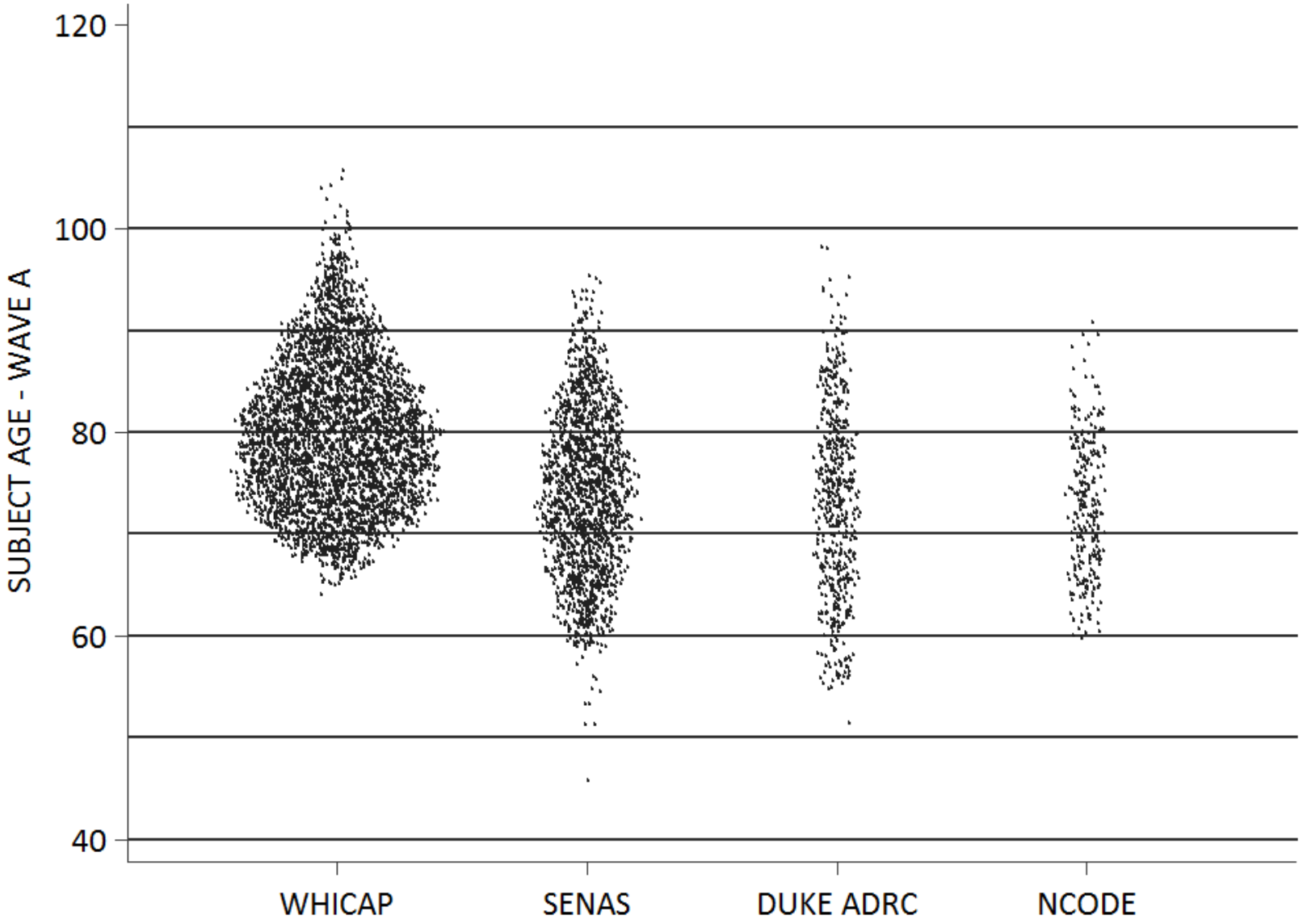
$$j \in [1, T]$$

$$\epsilon \sim N(0, \theta), \quad \zeta \sim N(0, \psi)$$

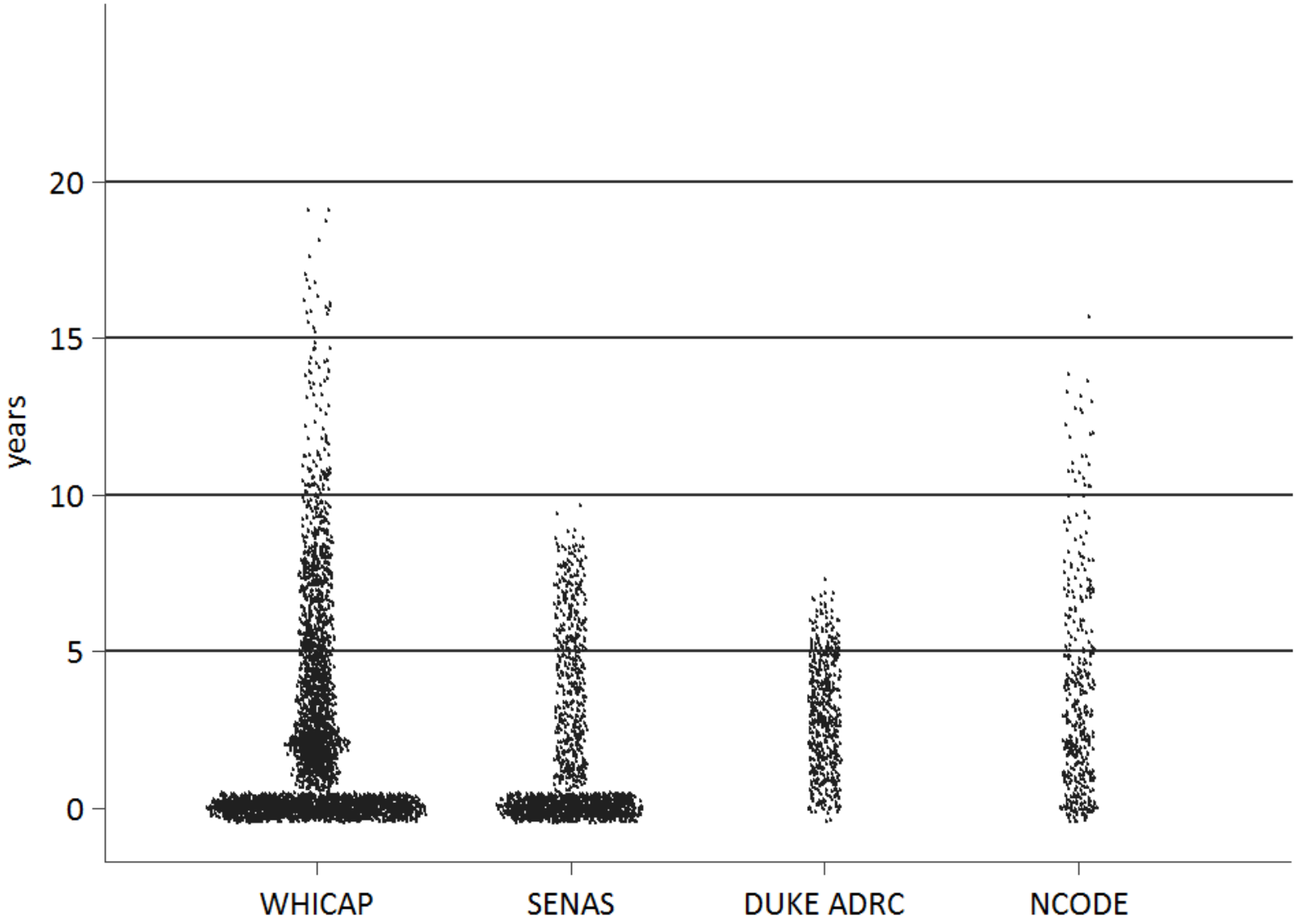


$$y_{ij} = \mathsf{i}_i + \mathsf{s}_i \lambda_j + e_{ij}$$
$$\mathsf{i}_i = \alpha(i) + \zeta(i)_i$$
$$\mathsf{s}_i = \alpha(s) + \zeta(s)_i$$
$$i \in [1, N]$$
$$j \in [1, T]$$
$$e \sim N(0, \theta), \zeta \sim N(0, \psi)$$

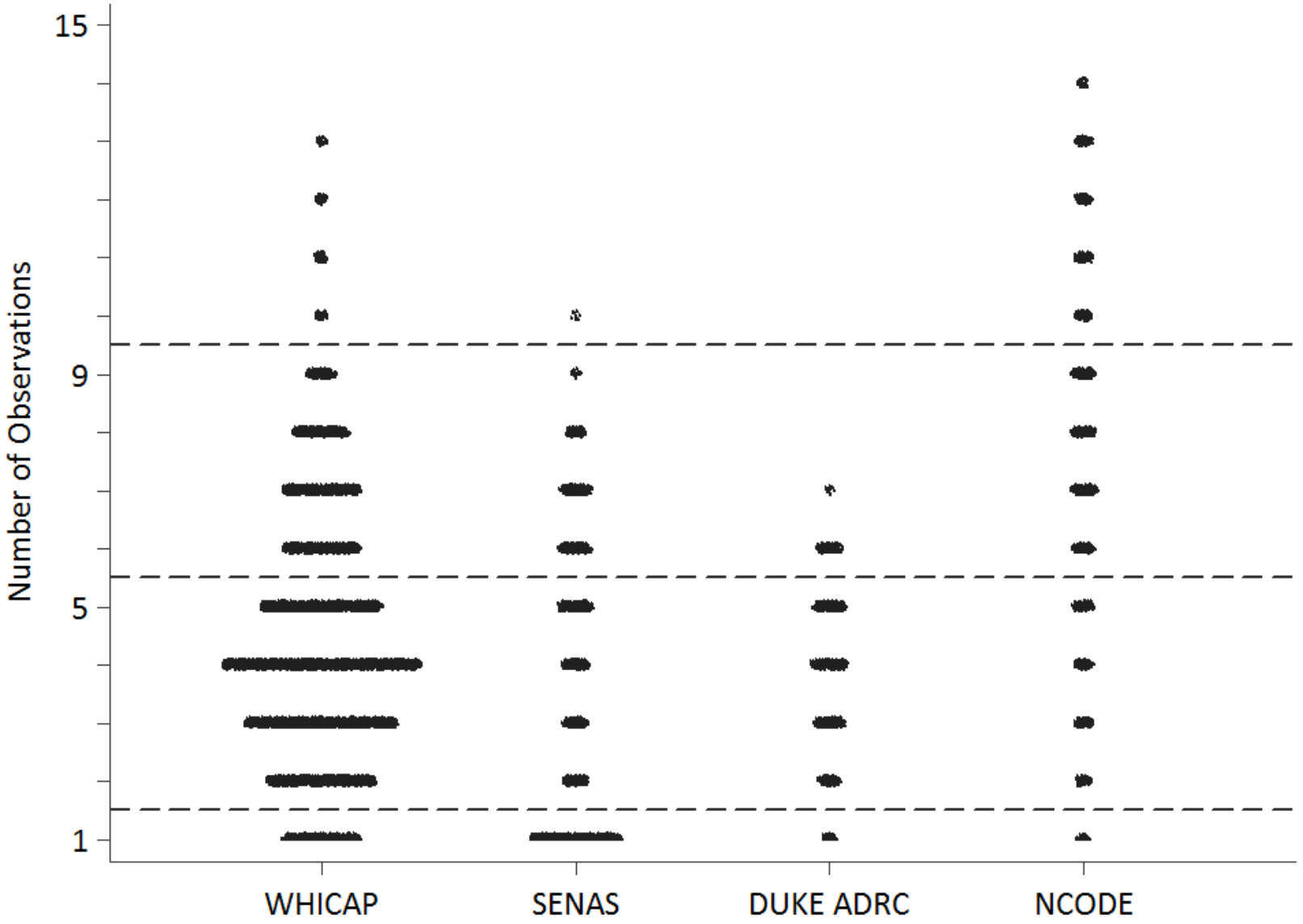
Age at baseline



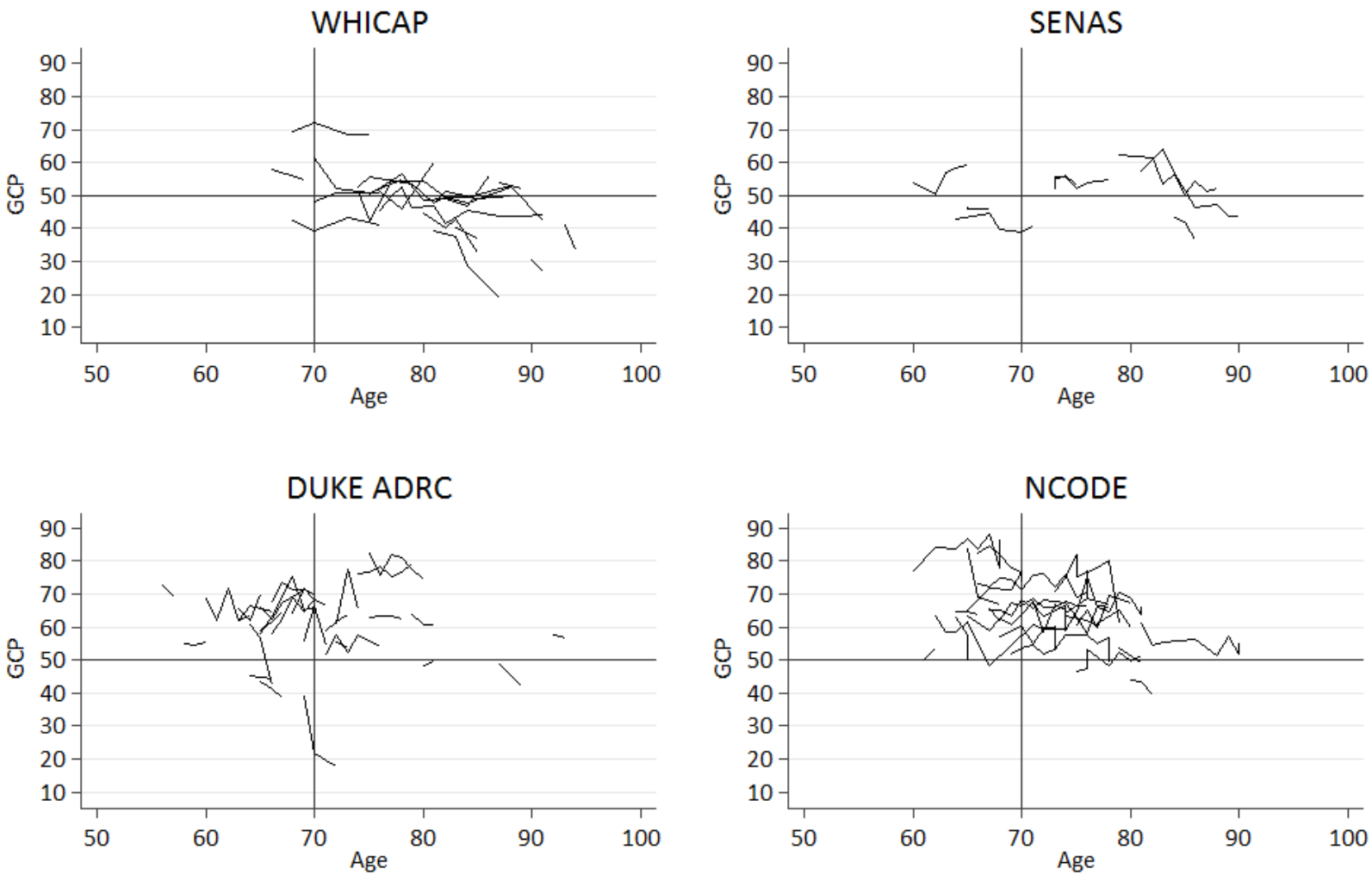
Maximum years of follow-up



Maximum Repeat Observations



Alden's GCP


WHICAP
SENAS
DUKE ADRC
NCODE
GCP
Age
10
20
30
40
50
60
70
80
90
50
60
70
80
90
100

example1.sas

```
PROC MIXED METHOD=ML NOCLPRINT NOINFO COVTEST ;
   CLASS id ;
   MODEL gcp = cage70 cage70sq /SOLUTION NOTEST ;
   RANDOM INTERCEPT /TYPE=UN SUB=id ;
RUN;
```

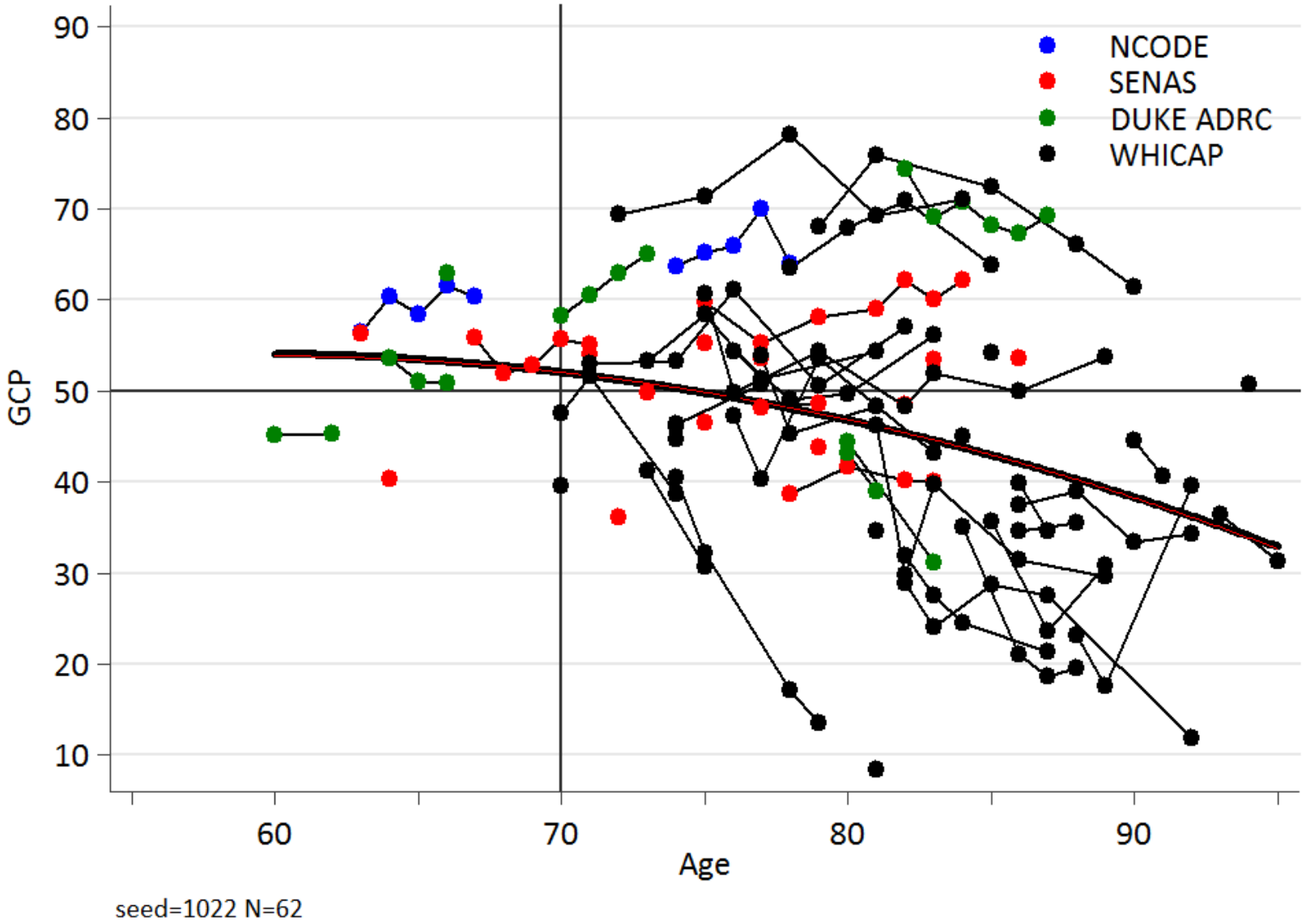
example1.inp

```
DATA:     FILE = __000001.dat ;
VARIABLE: NAMES = gcp1-gcp12 cage1-cage12 ;
          MISSING are all (-9999) ;
          TSCORES = cage1-cage12 ;
ANALYSIS: TYPE = random ;
          COVERAGE = 0 ;
MODEL:    i s q | gcp1-gcp12 at cage1-cage12 ;
          s@0; q@0; s with q@0; i with s@0 q@0 ;
          gcp1-gcp12*(e) ;
```

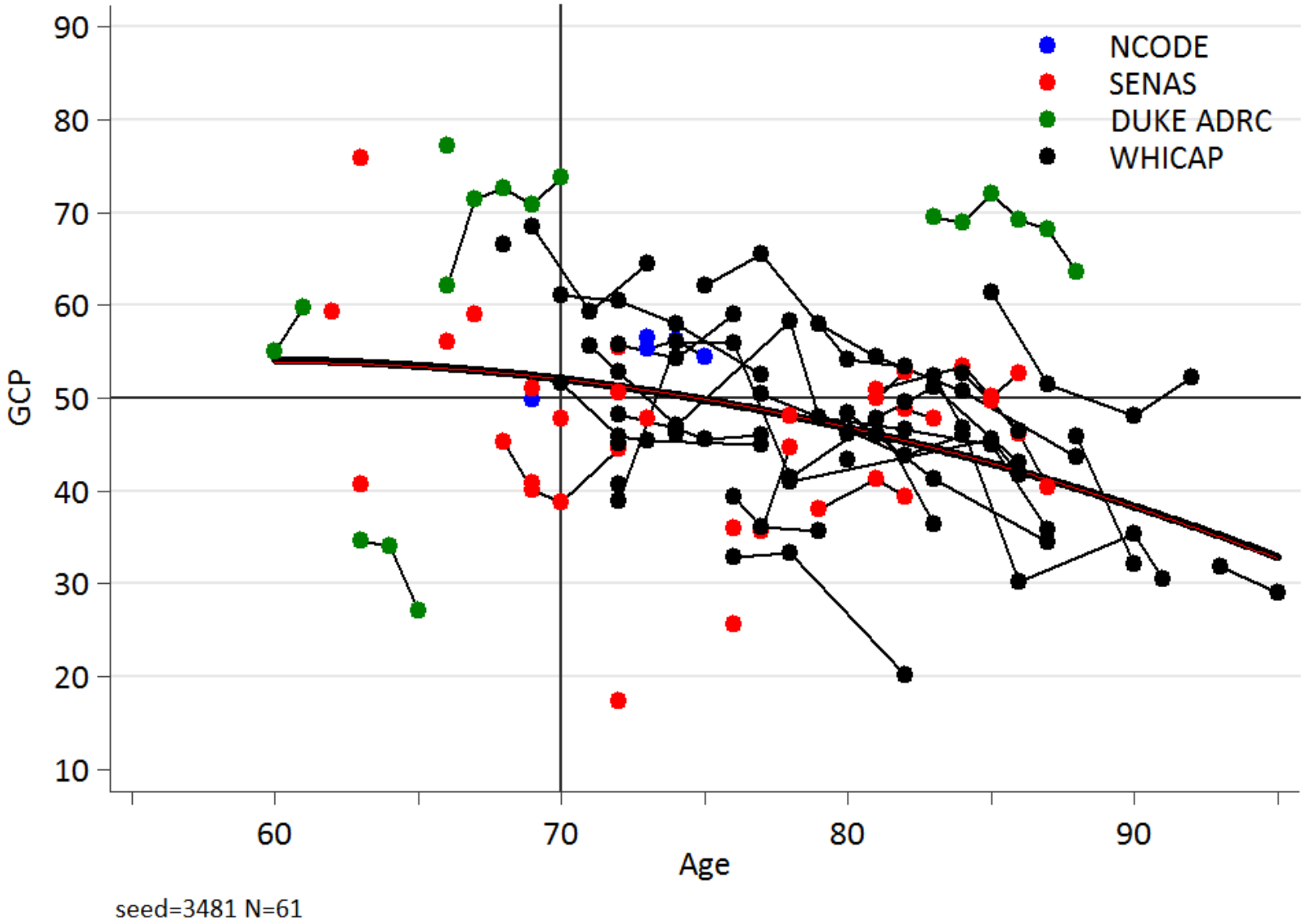
example1alt.inp

```
DATA:     FILE = ex1alt.dat ;
VARIABLE: NAMES = gcp1-gcp36 ;
MISSING are all (-9999) ;
ANALYSIS: COVERAGE = 0 ;
MODEL:    i s q | gcp1@0 gcp2@1 gcp3@2
gcp4@3 gcp5@4 gcp6@5
gcp7@6 gcp8@7 gcp9@8
gcp10@9 gcp11@10 gcp12@11
gcp13@12 gcp14@13 gcp15@14
gcp16@15 gcp17@16 gcp18@17
gcp19@18 gcp20@19 gcp21@20
gcp22@21 gcp23@22 gcp24@23
gcp25@24 gcp26@25 gcp27@26
gcp28@27 gcp29@28 gcp30@29
gcp31@30 gcp32@31 gcp33@32
gcp34@33 gcp35@34 gcp36@35
s@0; q@0; s with q@0; i with s@0 q@0 ;
gcp1-gcp12*(e) ;
```

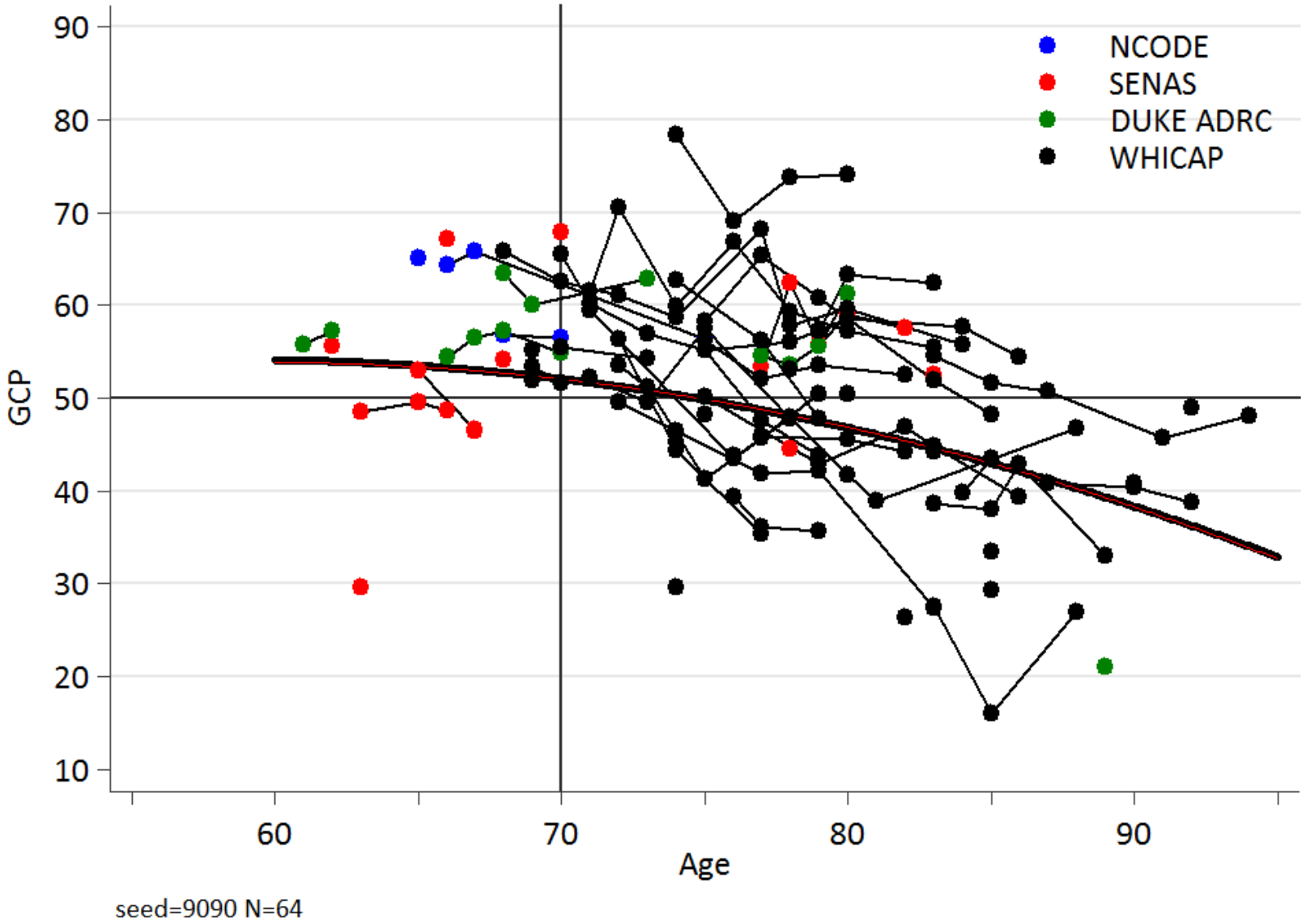
Fitted Values (seed 1022)



Fitted Values (seed 3481)



Fitted Values (seed 9090)

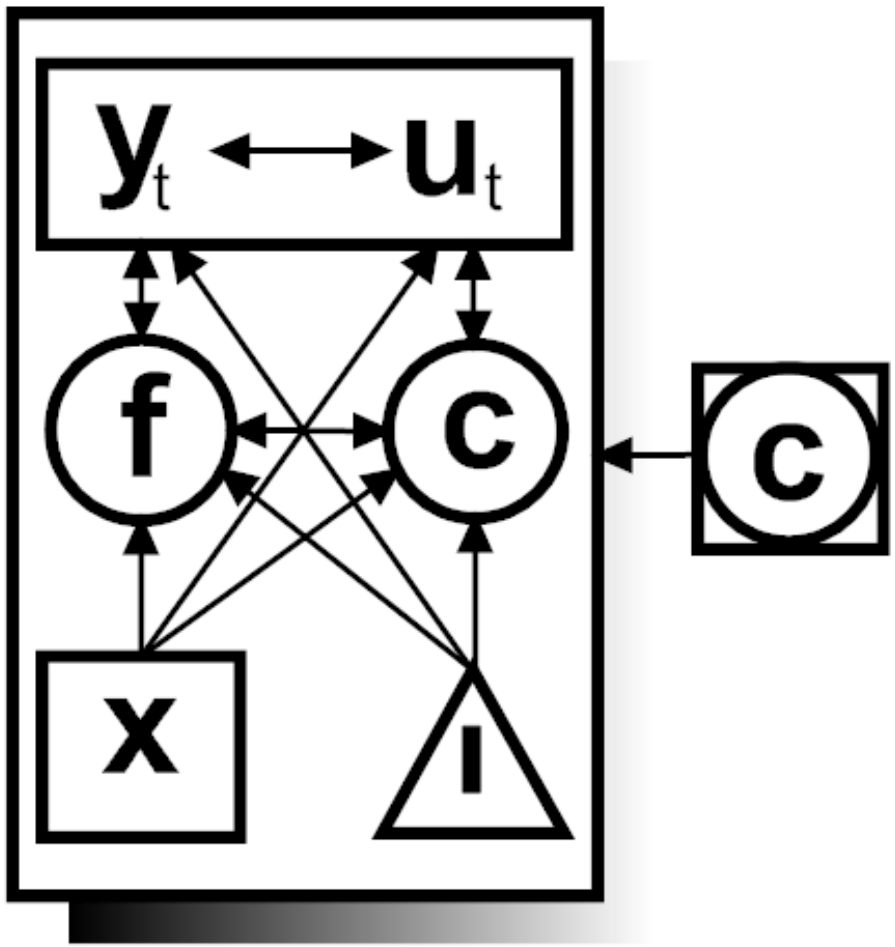


Proc Mixed			Mplus (using tscores)		
Parameter	**Est**	**SE**	**Parameter**	**Est**	**SE**
Intercept	51.968	0.145	Mean i	52.008	0.149
cage70	-0.324	0.019	Mean s	-0.309	0.024
cage70sq	-0.017	0.001	Mean q	-0.019	0.001
un(1,1)	105.980	1.909	Variance(i)	104.182	1.976
Residual	20.069	0.252	Resid Var	19.042	0.399

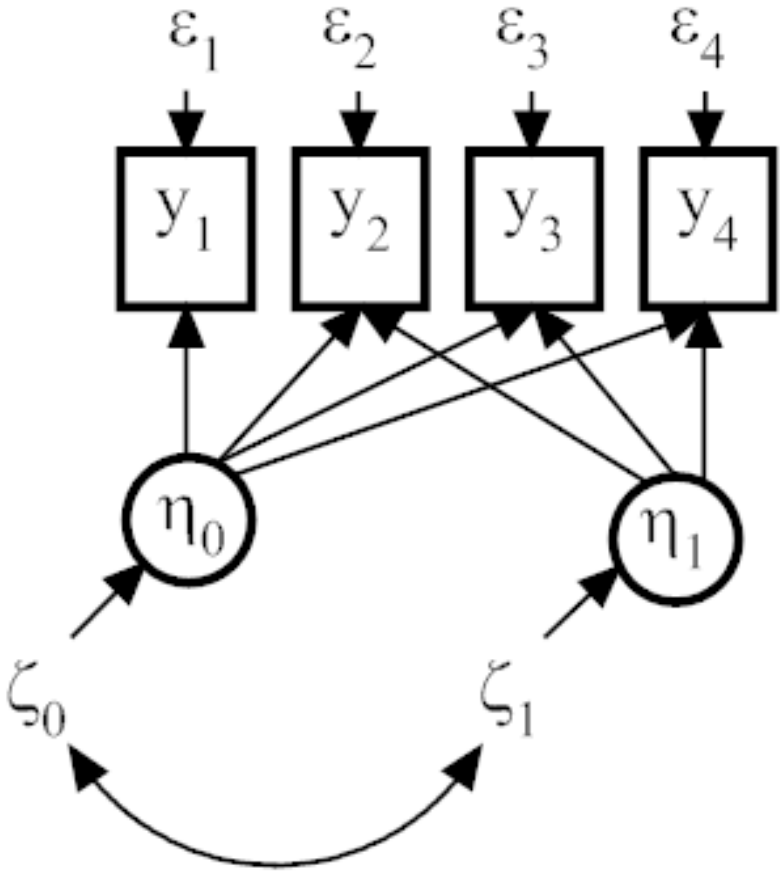
Resid var = residual variance.

Advantages of LDA in SEM

- There aren't any if the analysis question is relatively straight-forward.
  - If interested in change over time, including group differences in change over time, conventional random effects or mixed effect modeling is a better choice than SEM-based approaches
  - Why? More readers (and article reviewers) will know what you are talking about
- SEM based LDA can be extended in a variety of ways to more easily address
  - Challenging design issues
  - Complex but substantively important
  - Joint models 2+ processes
  - Sequential processes
  - Missing data modeling
  - Sub-populations
  - Multi-level models
- These, and other, model extensions can be combined

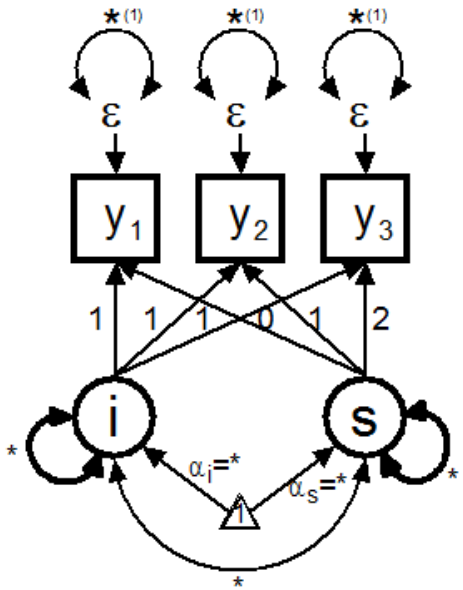

y_t
u_t
f
c
c
x
I

Simple Unconditional LGCM



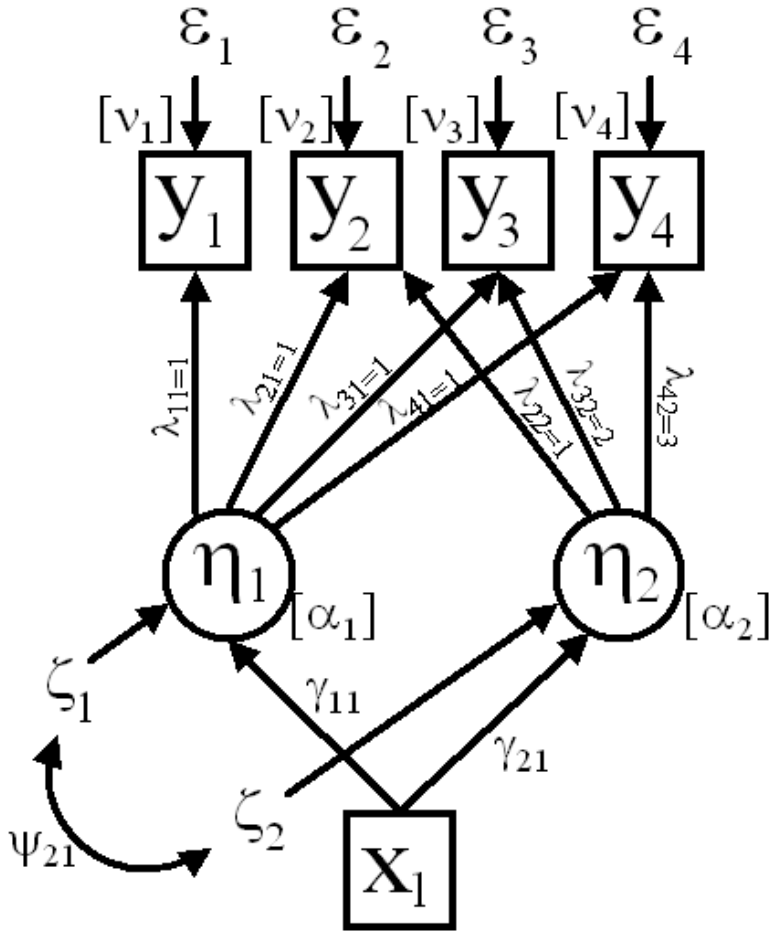
From Singer and Willet (2004) and UCLA/ATS

Model B (Table 4.1)



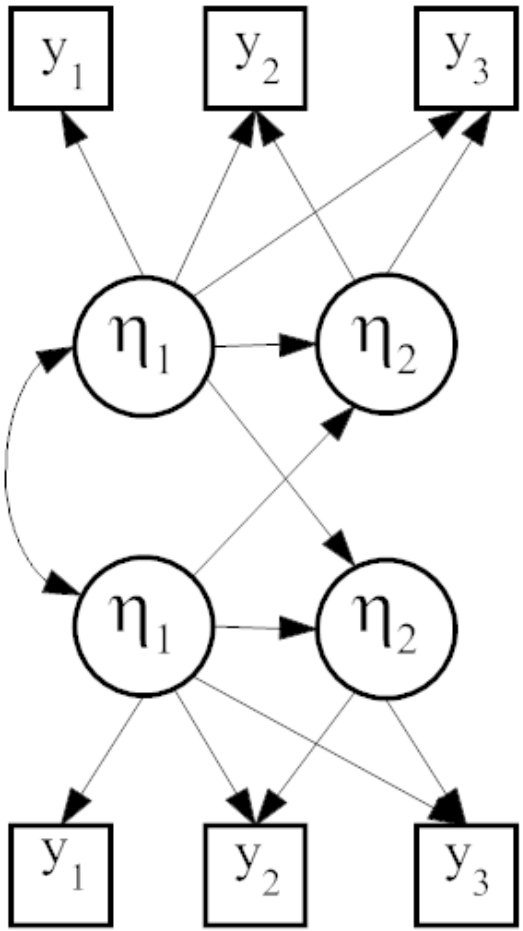

```
* SAS Example from http://www.ats.ucla.edu/stat/examples/alda.htm
*-------------------------------------------------------------------------------
* SAS
proc mixed data="c:\alda\alcohol1_pp" method=ml noclprint noinfo covtest;
  title2 "Model B";
  class id;
  model alcuse = age_14/solution notest;
  random intercept age_14/type=un sub=id;

*-------------------------------------------------------------------------------

* Mplus (short hand)
DATA:     FILE = C:\work\Shows\SHORTC~1\2009\data\swch4.dat ;
VARIABLE: NAMES = alcuse1 alcuse2 alcuse3 ;
ANALYSIS: ESTIMATOR = mlr ;
MODEL:    i s | alcuse1@0 alcuse2@1 alcuse3@2 ;
            alcuse1-alcuse3 (1) ;

*-------------------------------------------------------------------------------

* Mplus (long hand)
DATA:     FILE = C:\work\Shows\SHORTC~1\2009\data\swch4.dat ;
VARIABLE: NAMES = alcuse1 alcuse2 alcuse3 ;
ANALYSIS: ESTIMATOR = mlr ;
MODEL:    i by alcuse1-alcuse3@1 ;
          s by alcuse1@0 alcuse2@1 alcuse3@2 ;
          [alcuse1-alcuse3@0] ;
          [i* s*] ;
          alcuse1-alcuse3 (1) ;
```

Conditional LGCM



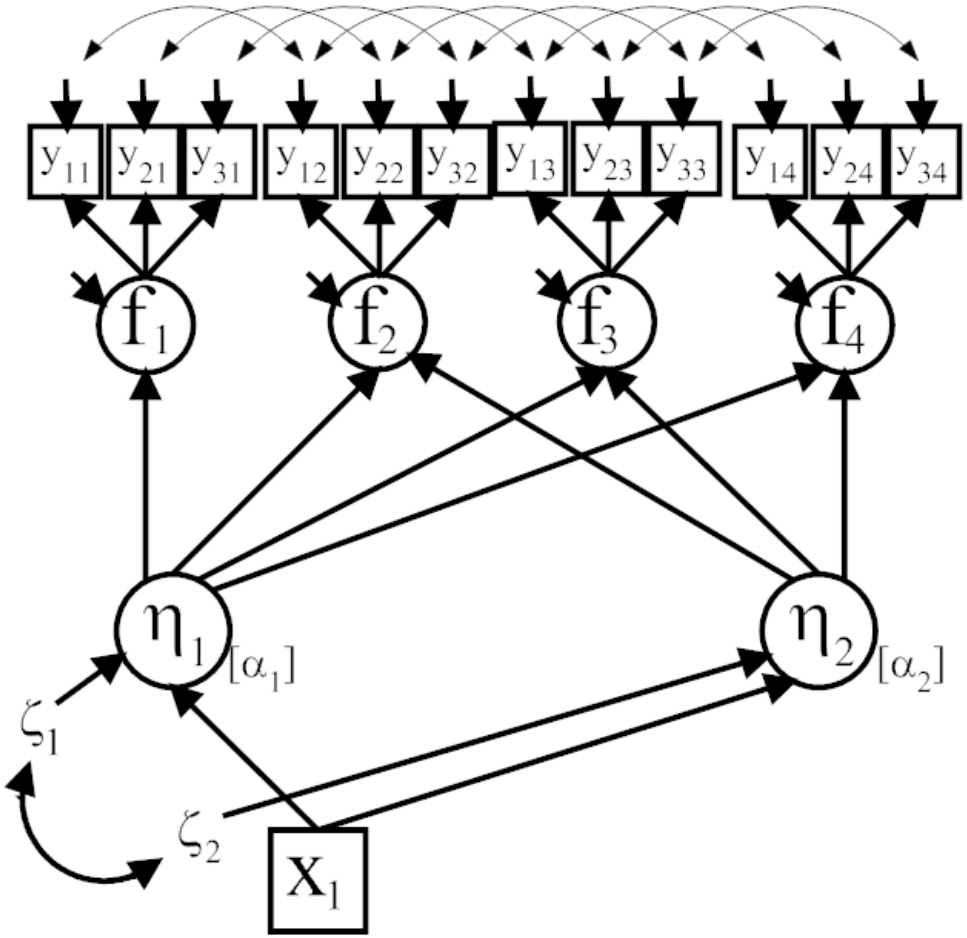
Parallel Process LGCM

Two Process Changing at the Same Time: Examine Covariation



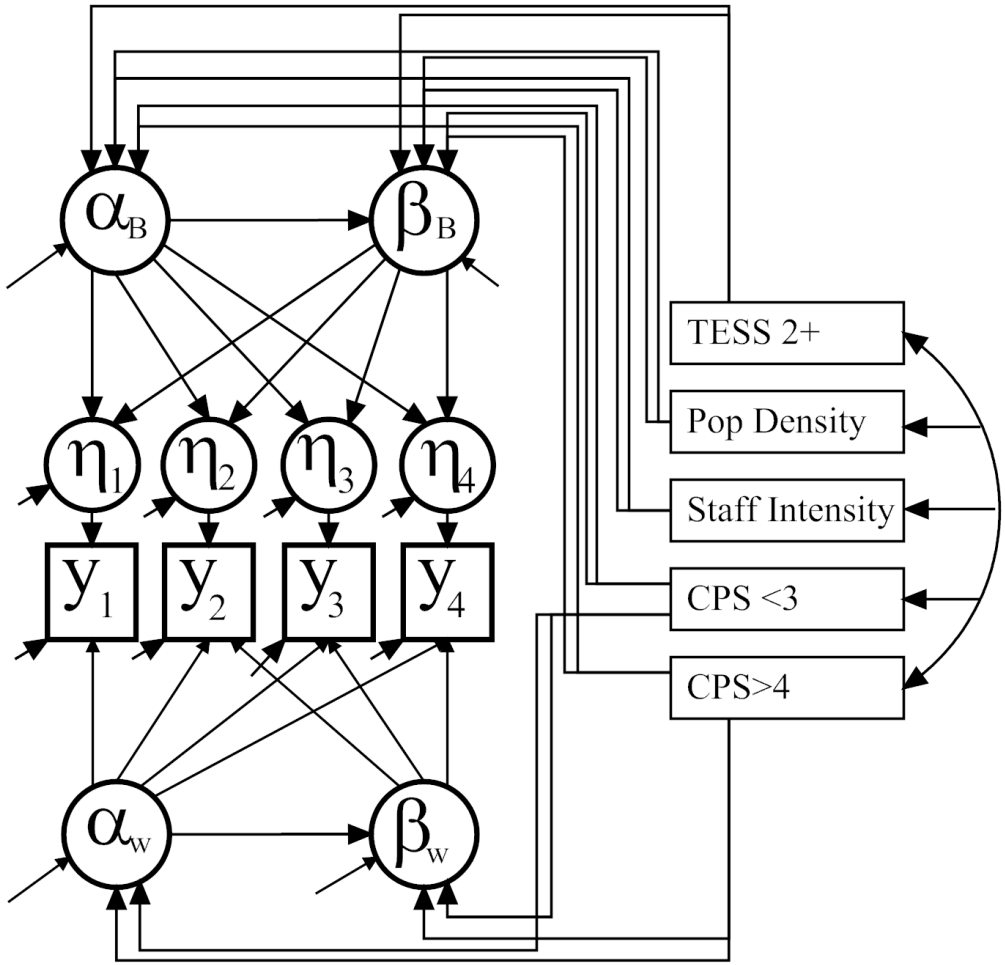
Multiple Indicator LGCM

Analyze change in a latent variable by explicitly modeling it's measurement at multiple occasions (allow for DIF, missing items, other noninvariance issues)



Multilevel Models

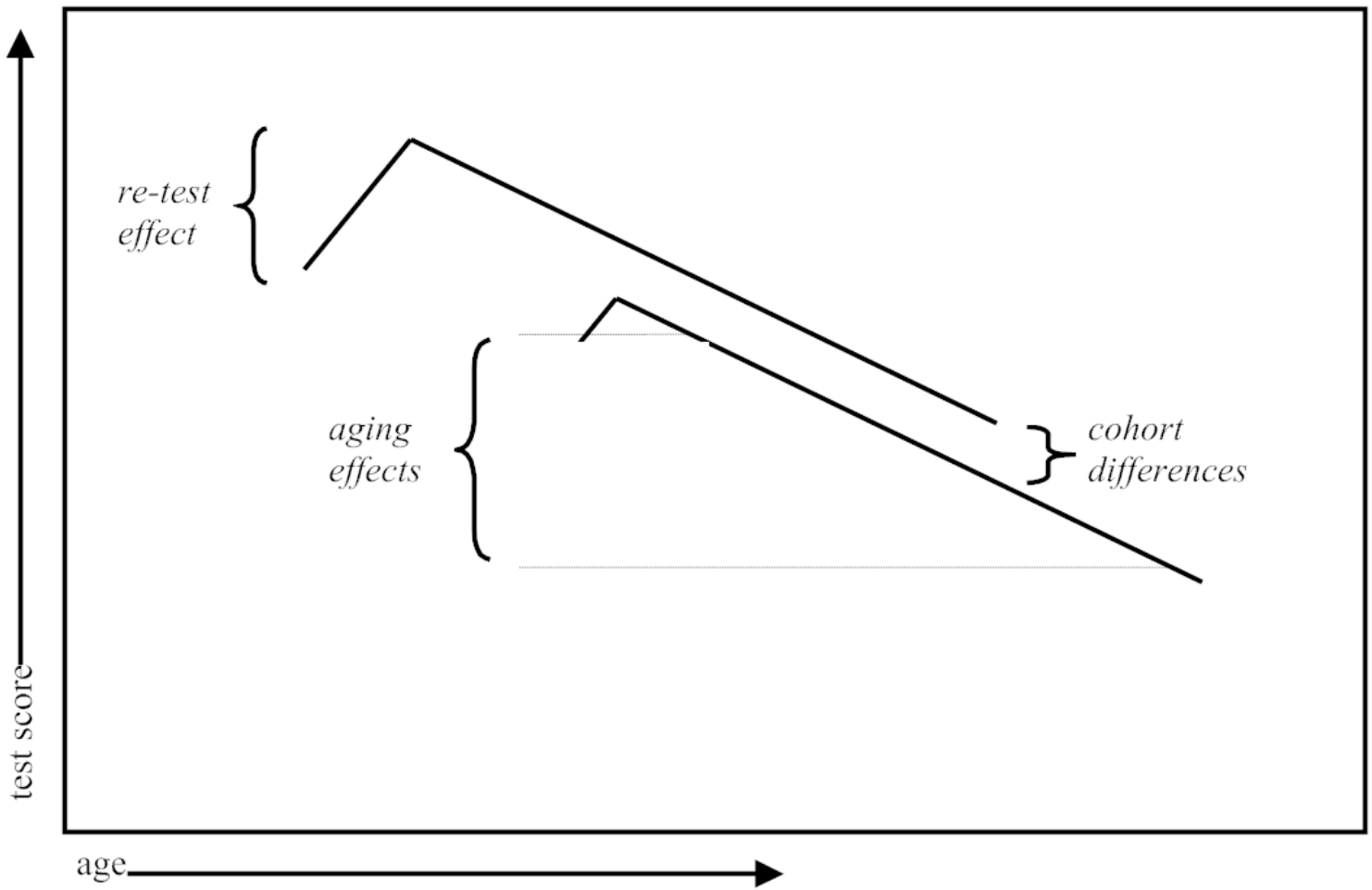
When observations are clustered.[1]

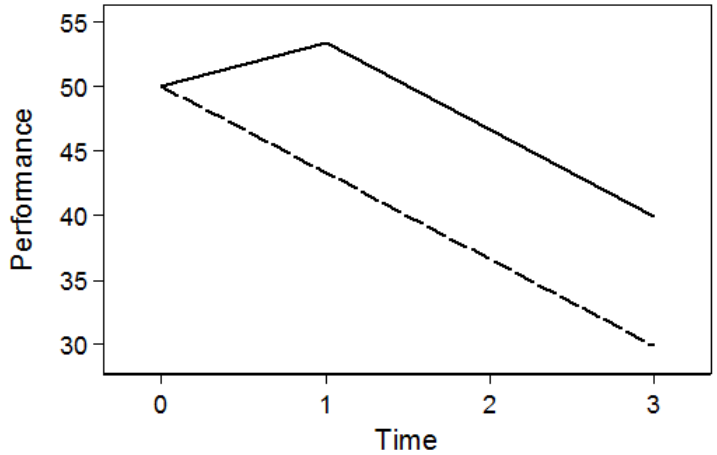


[1] There is a another meaning of multilevel models discussed later

Retest Effects

A real problem with repeat neuropsychological test adminstration.



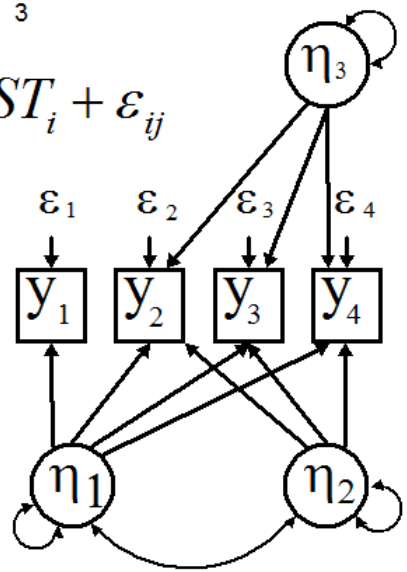


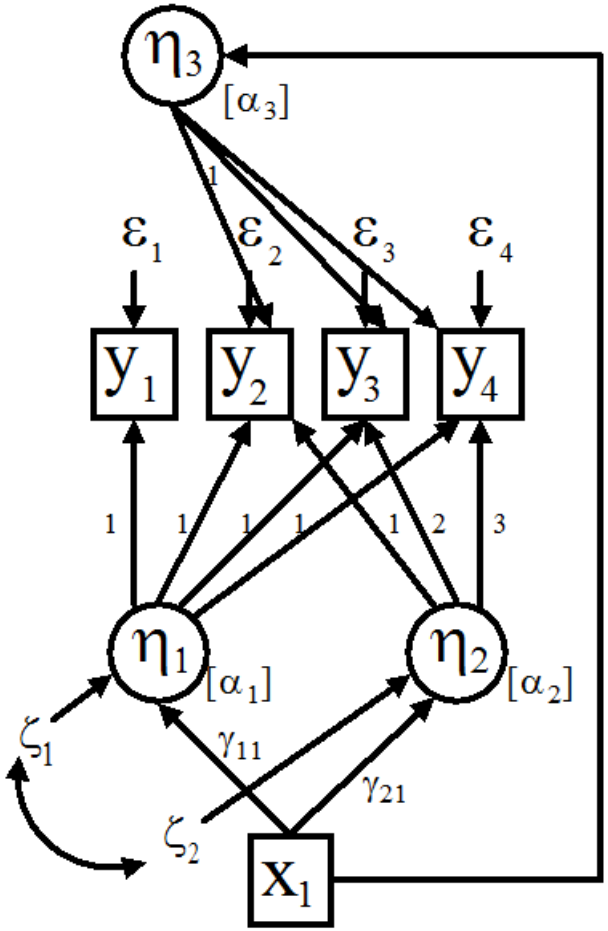
$$y_{ij} = \eta_{1i} + \eta_{2i} \times TIME_i + \eta_{3i} \times RETEST_i + \varepsilon_{ij}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i}$$

$$\eta_{2i} = \alpha_2 + \zeta_{2i}$$

$$\eta_{3i} = \alpha_3 + \zeta_{3i}$$

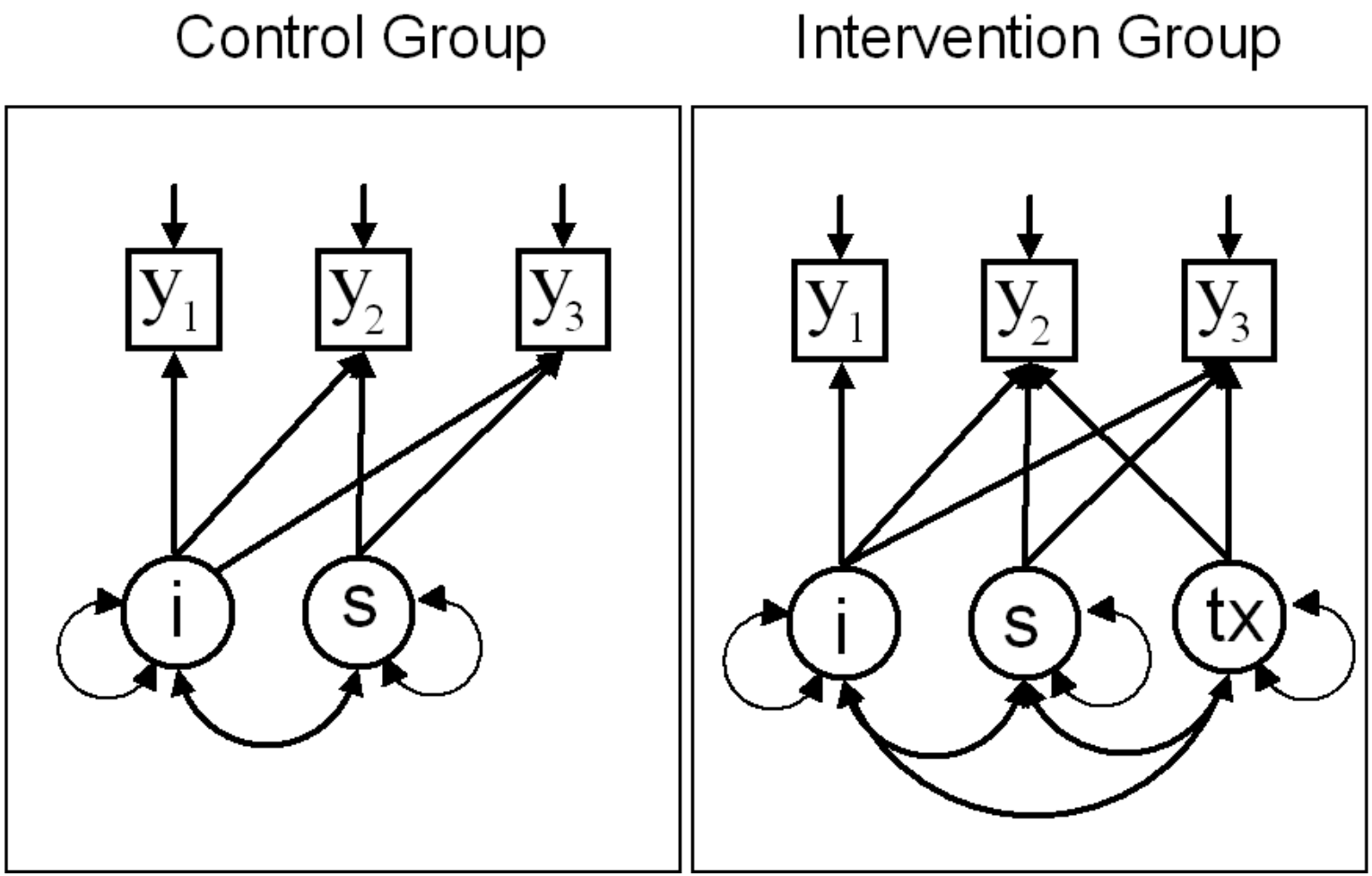





```
TITLE:    Latent Growth Curve
          with retest effect
DATA:     FILE = blah.dat ;
VARIABLE: NAMES= y1-y4 x1 ;
MODEL:
  i s | y1@0 y2@0 y3@0 y4@0 ;
  i s on x1 ;
  r by y2-y4@1 ;
  [r*] ;
  ! test what happens
  ! relaxing the constraints
  ! below...
  r@0 ;
  r with i @0 ;
  r with s @0 ;
```

Analysis of Randomized Studies

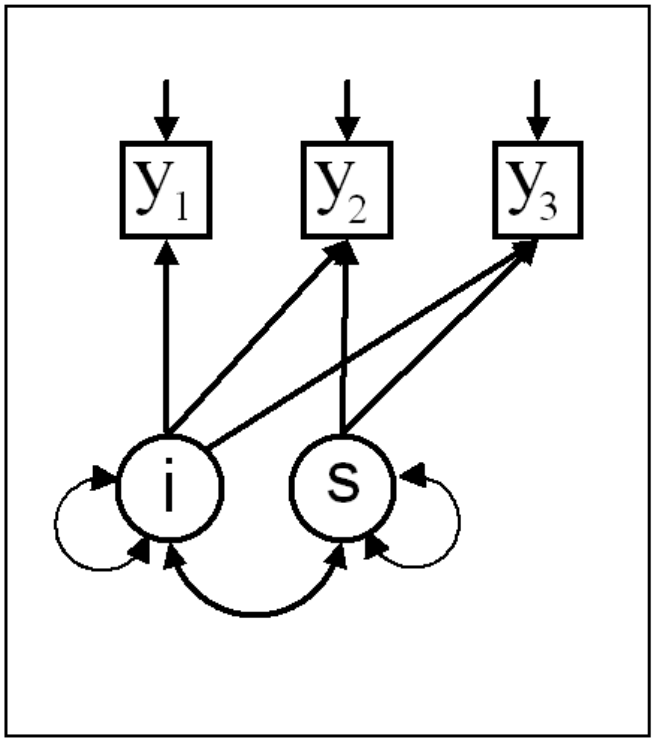
Or natural experiments: Treatment effect as a latent variable (with a variance



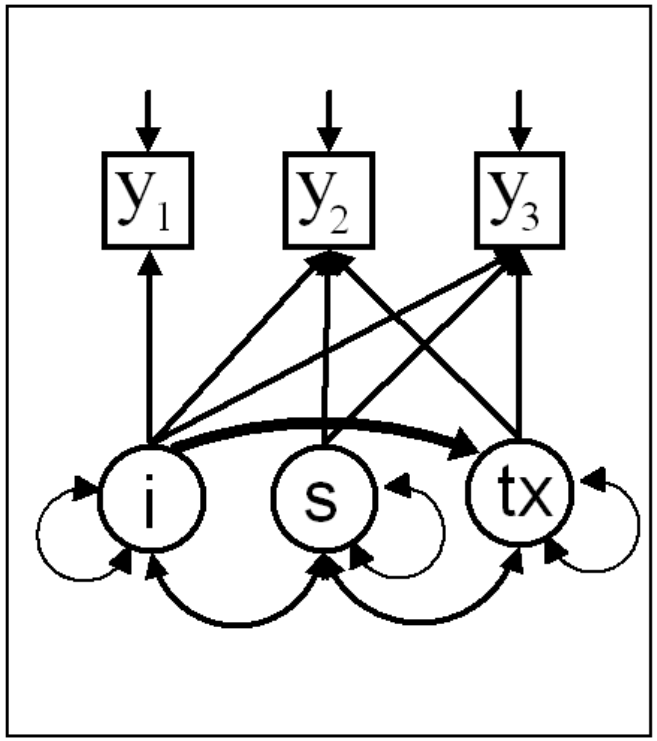
Analysis of Randomized Studies

Treatment effect dependent on baseline

Control Group

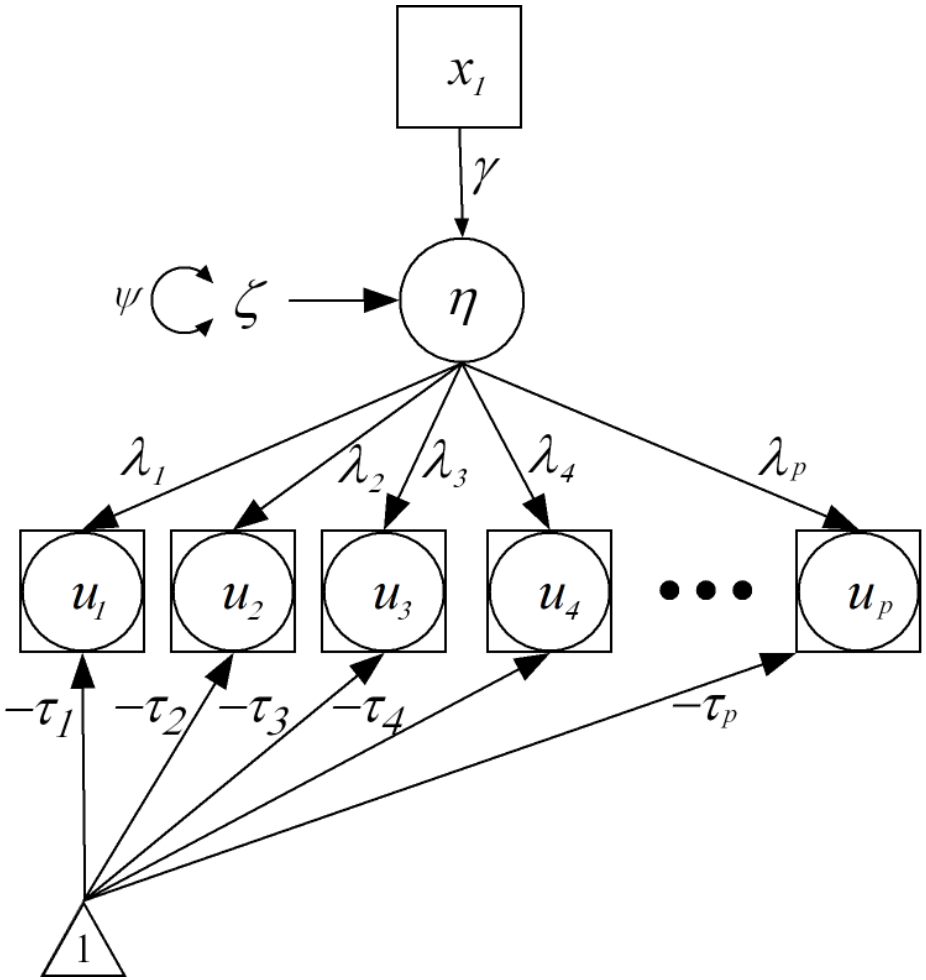


Intervention Group



Baseline-dependent treatment effect

Discrete time survival analysis



$$u_j = \begin{cases} 0 & \text{if no event at time } j \ \& \ u_{j-1} = 0 \\ 1 & \text{if event at time } j \\ . & \text{if } u_{j-1} = 1 | u_{j-1} = . | \text{ censored at time } j \end{cases}$$

$$\boldsymbol{\tau} = \begin{pmatrix} * & * & * & * & * \end{pmatrix}$$
$$\boldsymbol{\Lambda}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$
$$\boldsymbol{\Psi} = 0$$
$$\boldsymbol{\Gamma} = \begin{pmatrix} * \end{pmatrix}$$

$$\hat{h}(j) = \frac{1}{1+e^{-\tau_j+\gamma}}$$

Joint continuous time survival and growth curve model

Mcardle et al., (2005) J. Geriatr. Psychiatry Neurol. 18(4):234

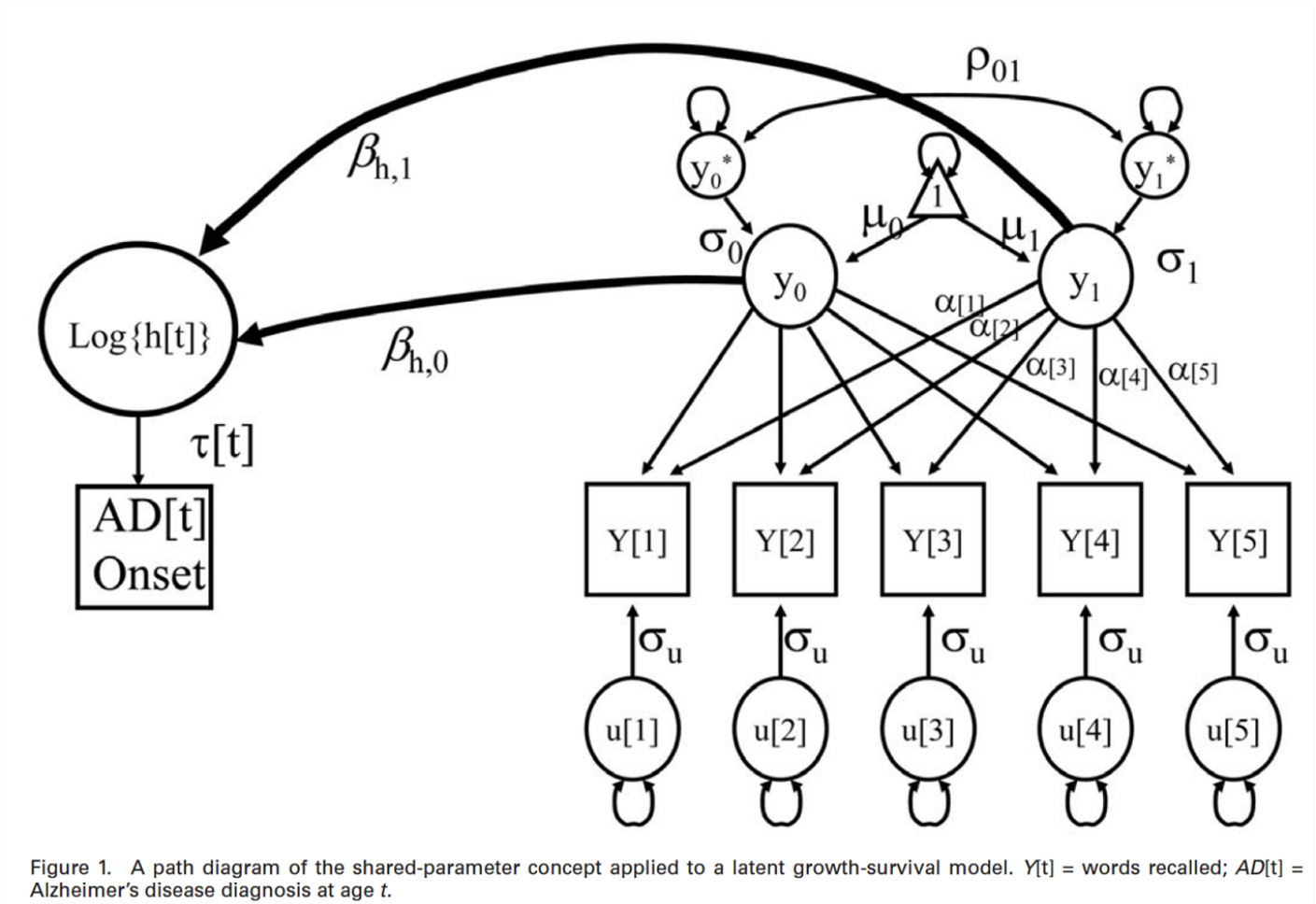
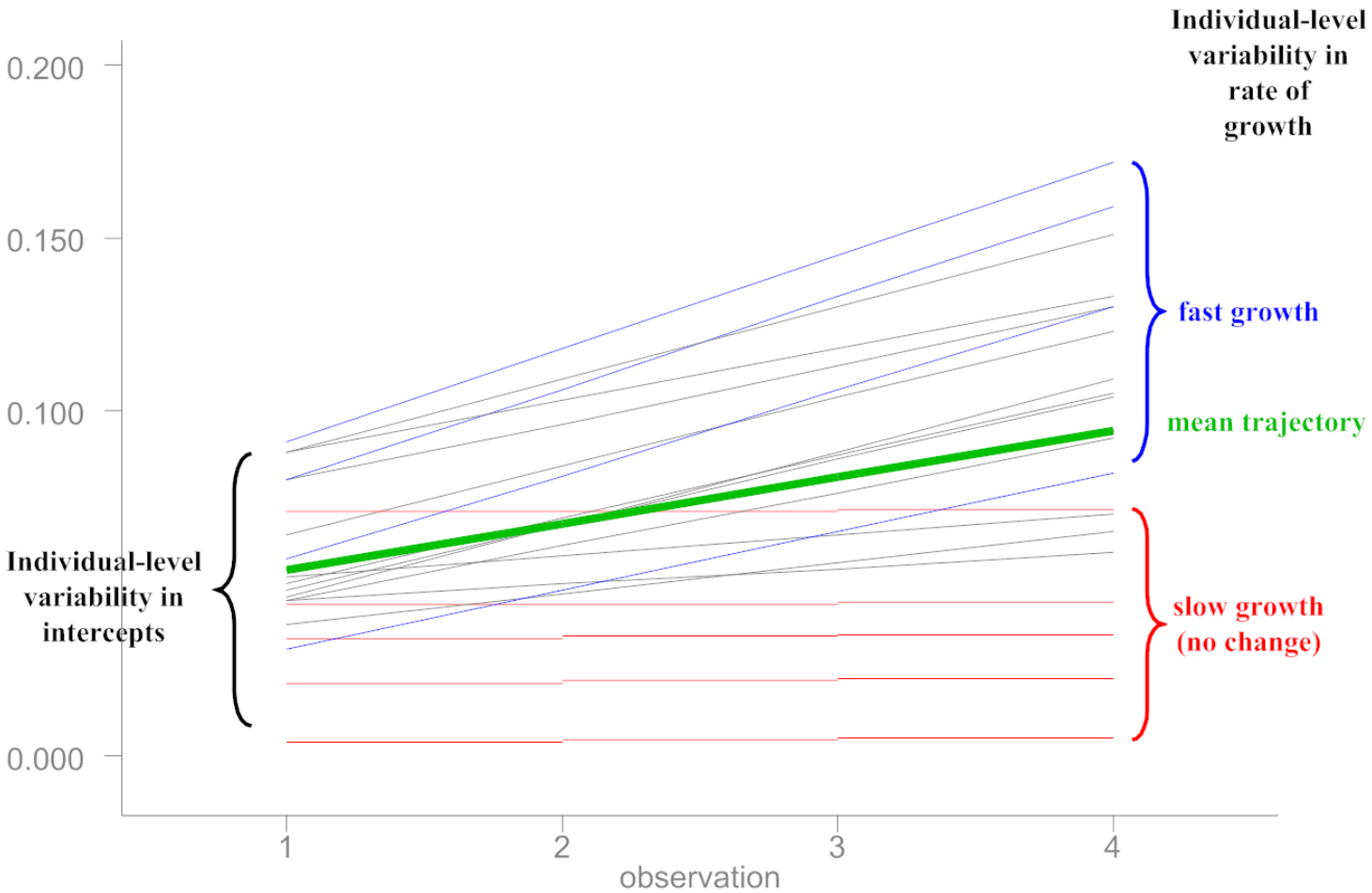
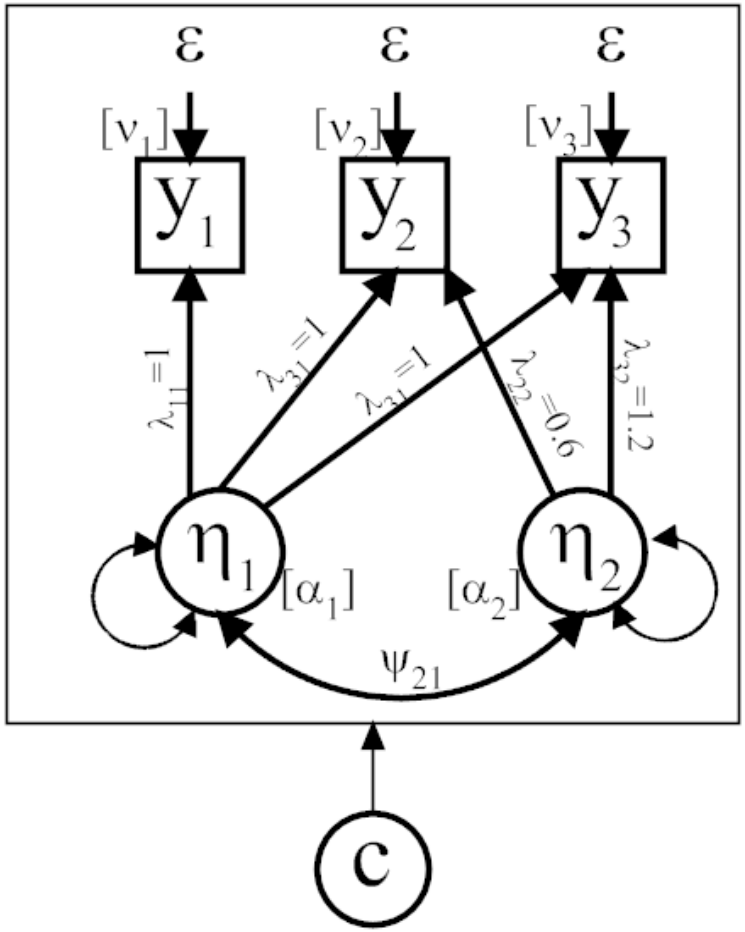
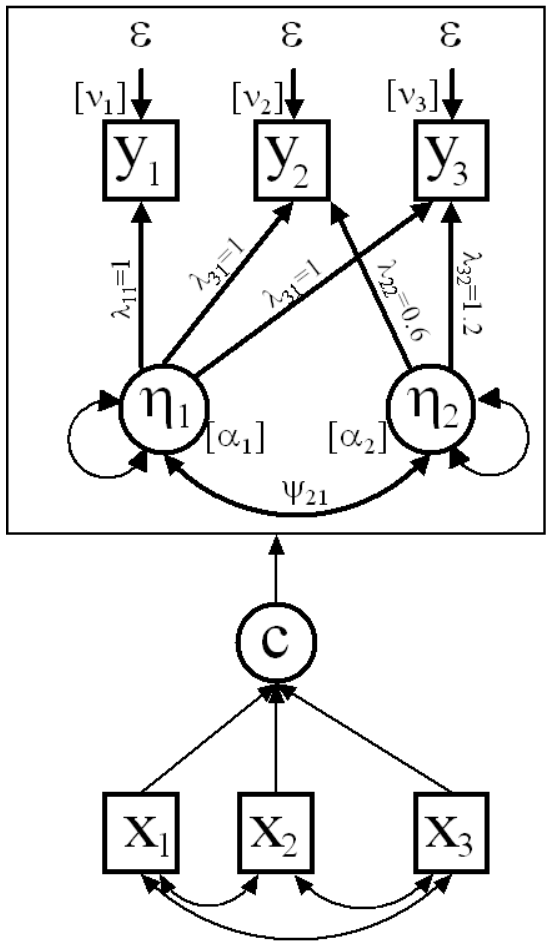


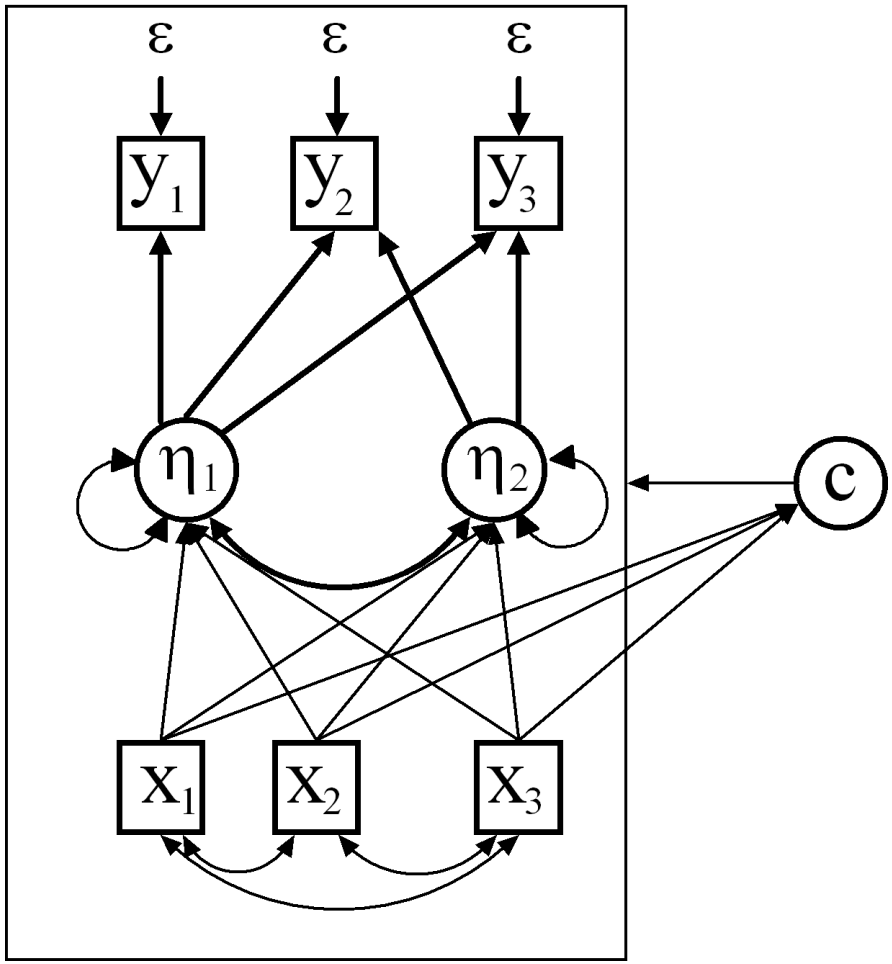
Figure 1. A path diagram of the shared-parameter concept applied to a latent growth-survival model. *Y*[t] = words recalled; *AD*[t] = Alzheimer's disease diagnosis at age *t*.

Growth Mixture Modeling




ε
ε
ε
[ν1]
[ν2]
[ν3]
y1
y2
y3
λ11=1
λ31=1
λ31=1
λ22=0.6
λ32=1.2
η1
η2
[α1]
[α2]
ψ21
c


ε
ε
ε
[ν1]
[ν2]
[ν3]
y1
y2
y3
λ11=1
λ31=l
λ31=l
λ32=0.6
λ32=1.2
η1
η2
[α1]
[α2]
ψ21
c
x1
x2
x3


ε
ε
ε
y1
y2
y3
η1
η2
c
x1
x2
x3

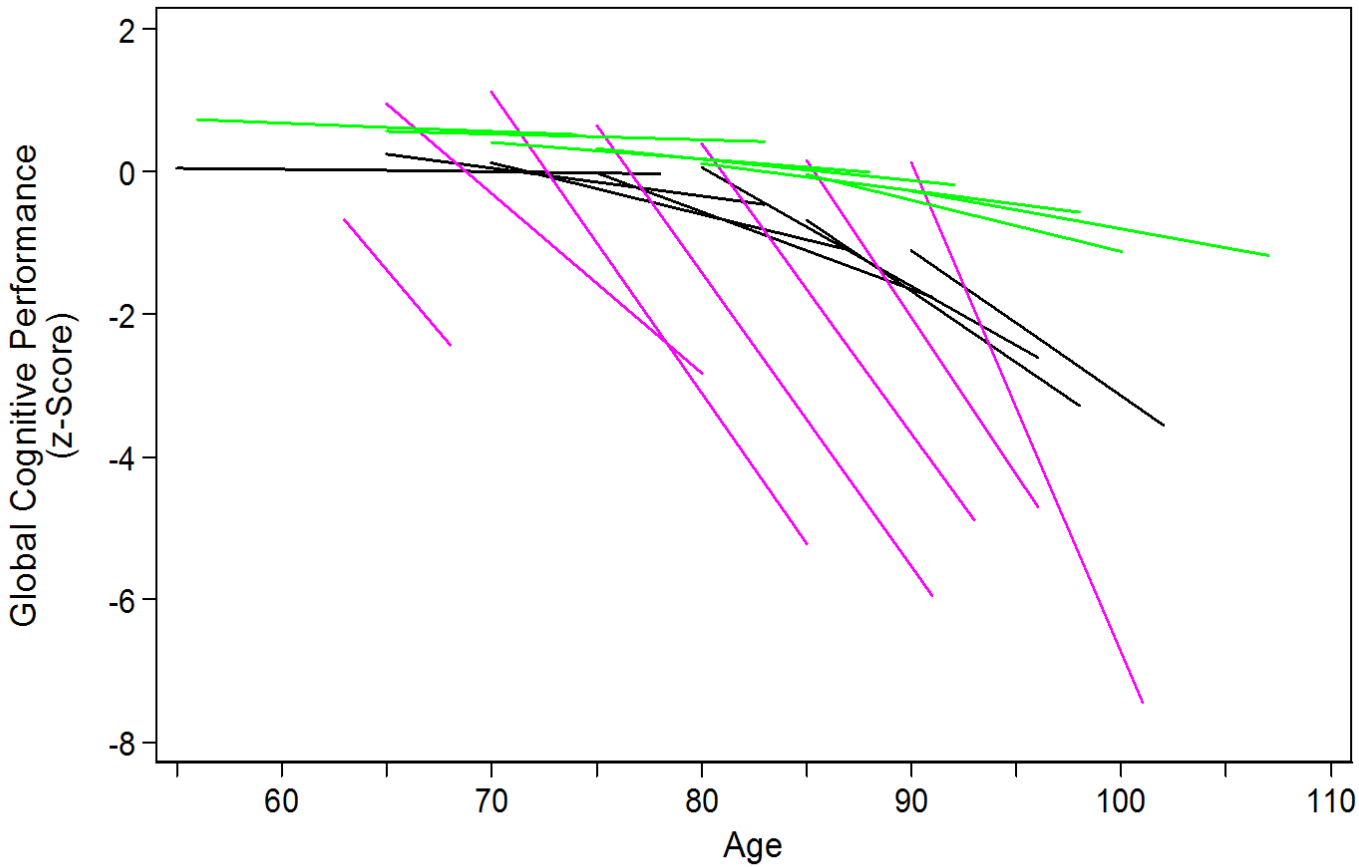
Growth Mixture Modeling

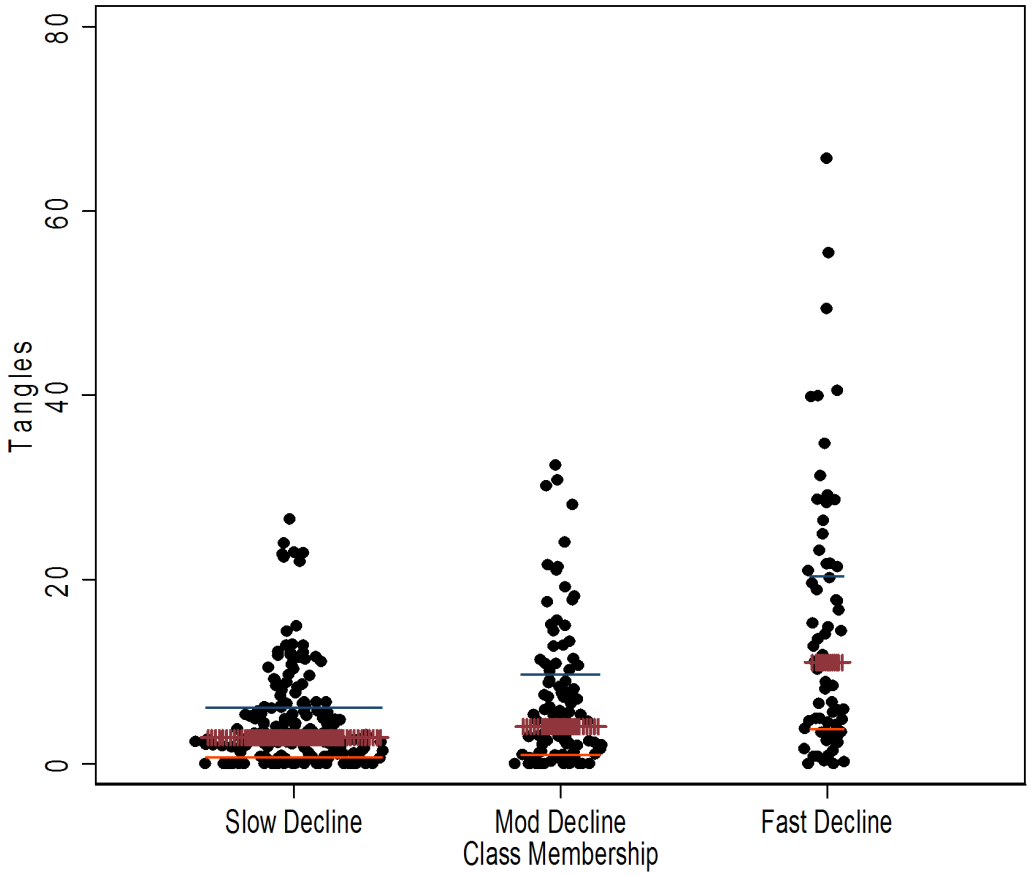
Identify population sub-samples with different growth trajectories

Age and Ageing 2011; **40:** 684–689
doi: 10.1093/ageing/afr101
Published electronically 2 September 2011

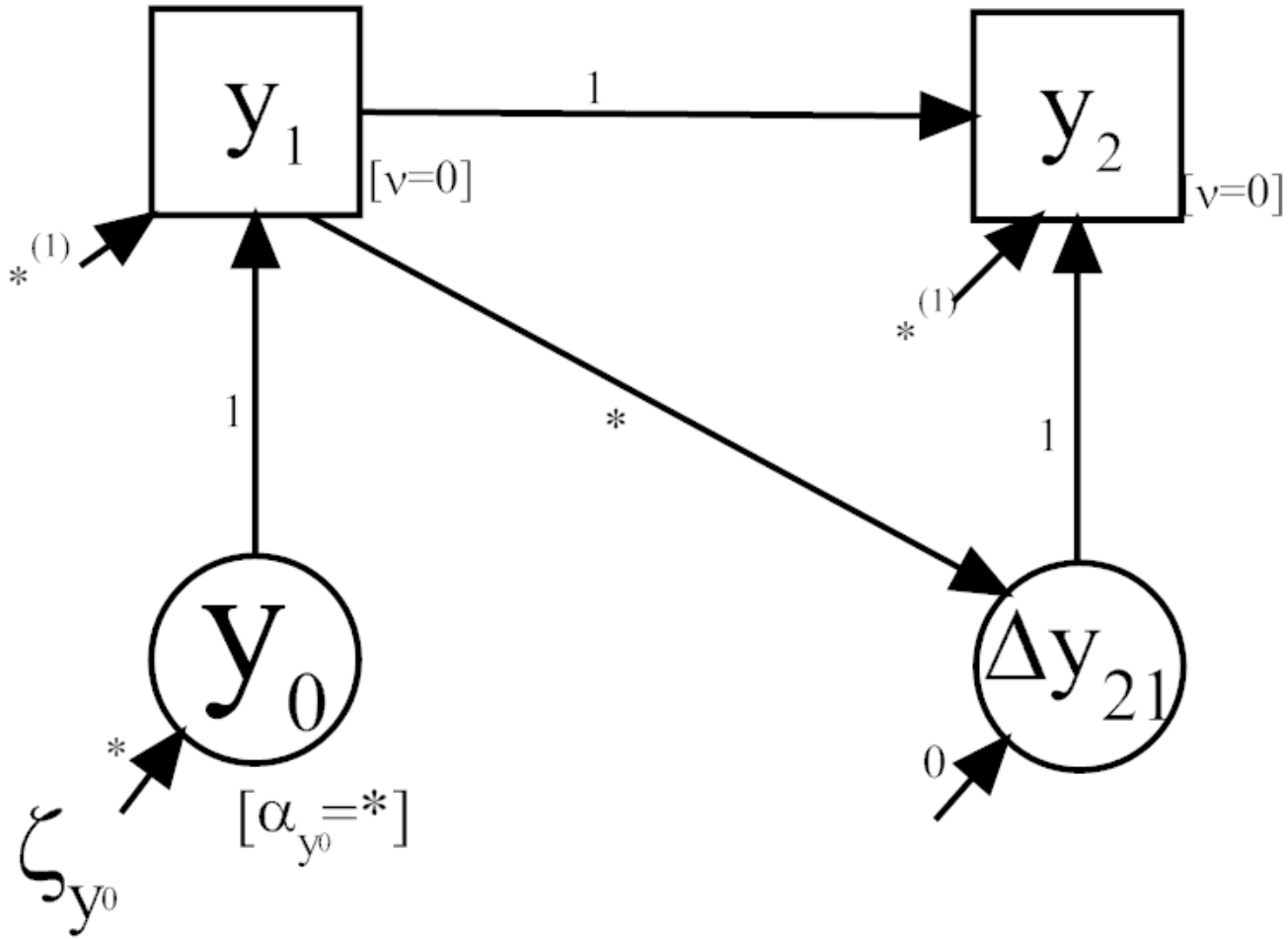
Cognitive decline in the elderly: an analysis of population heterogeneity

Kathleen M. Hayden[1], Bruce R. Reed[2,3], Jennifer J. Manly[4], Douglas Tommet[5], Robert H. Pietrzak[6], Gordon J. Chelune[7], Frances M. Yang[5], Andrew J. Revell[8], David A. Bennett[9], Richard N. Jones[5]



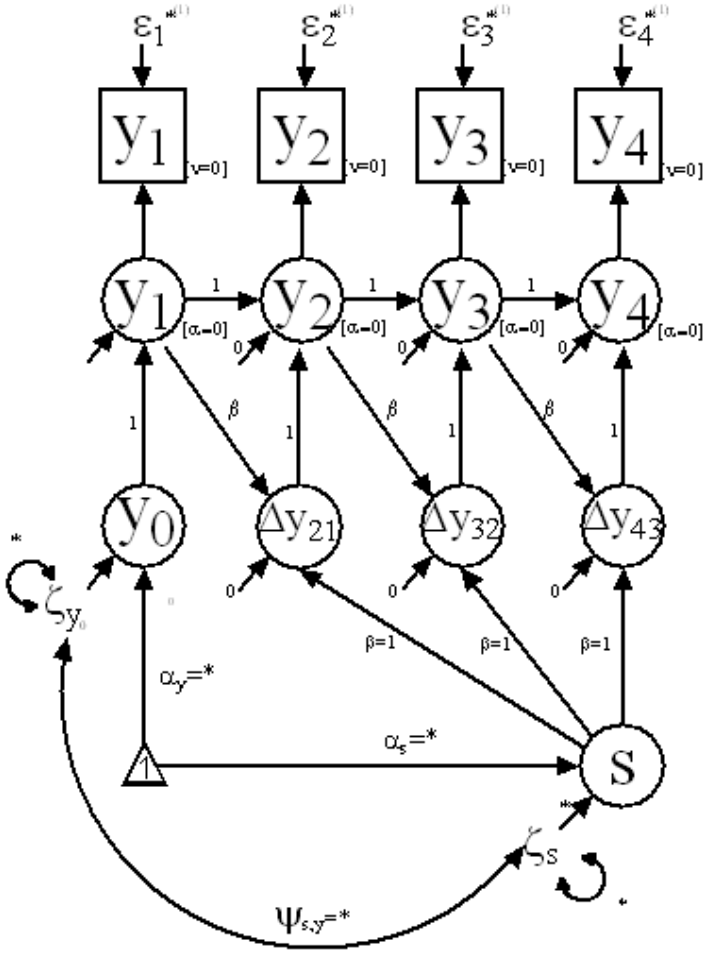

Tangles
0
20
40
60
80
Slow Decline
Mod Decline
Fast Decline
Class Membership

Latent Difference Score Models A different approach to LDA with SEM: the change scre as a latent variable



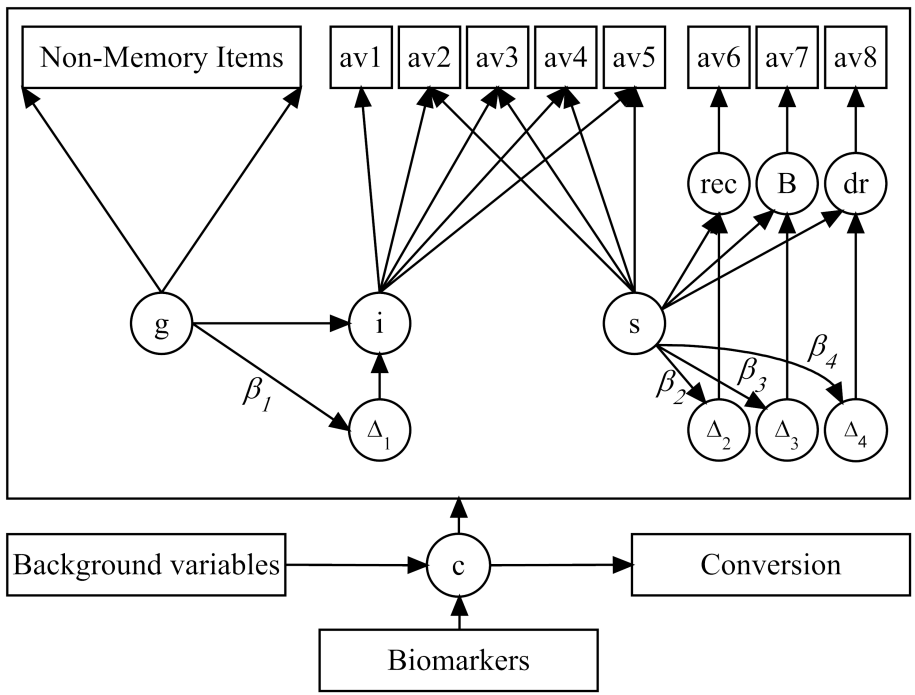
Dual Change Score Model

Change expressed with two parameters: autoregressive change score (Δ_y), and a systematic part (s). Flexible curve shapes can be estimated.



A Complex Model with LGCM, LDS, Latent Classes

This is an attempt to build a better way to use neuropsychological performance data to identify persons at risk for cognitive decline (conversion to clinical MCI or dementia). Gives priority to modeling word list learning (AVLT, `av1-av5`), discrepancies in word list learning and word list recognition and delayed recall trials (`av6`, `av8` via Δ_2, Δ_4) and incorporating background risk factors and biomarker information (Yang et al., 2015, *in preparation*, and/or Gross et al., 2015, *in preparation*).

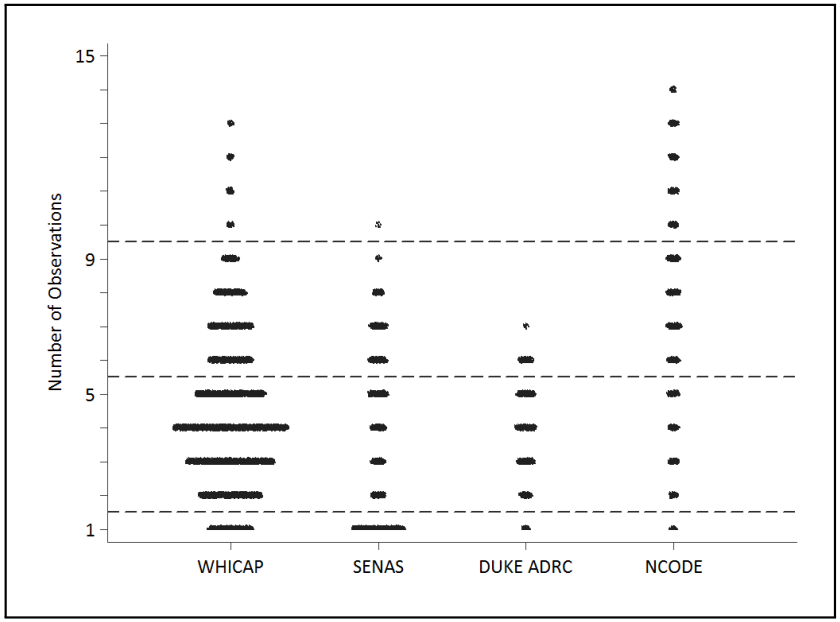


The Question

Start with a important question: You can do so much with modern SEM software that you can essentially just relax about the *analysis* and think about what the important *question* is

When to bring in the methodologist: Early, but after the question has been articulated.

The data



There is a lot of data

I would reccomend random effects modeling, either

- Random effects model with `TSCORES`
- Multilevel models

Multilevel Approach in Mplus

Requires long vs wide daya layout

LONG

Vertical

Multiple record

id	time	y1	x1
1	1	y11	x11
1	2	y12	x12
1	3	y13	x13
2	1	y21	x21
2	2	y22	x22
2	3	y23	x23
3	1	y31	x31
...			
n	p	ynp	xnp

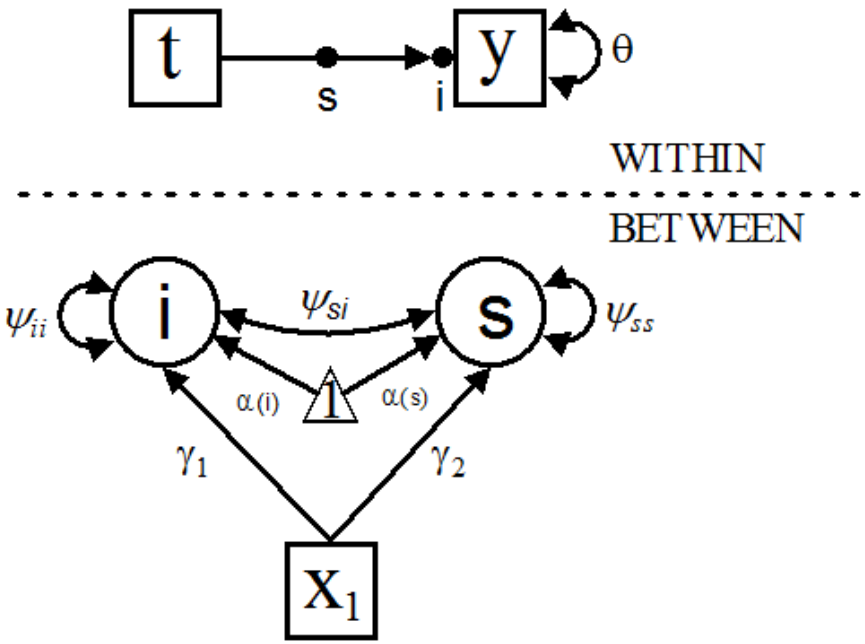
WIDE

Horizontal

Single record

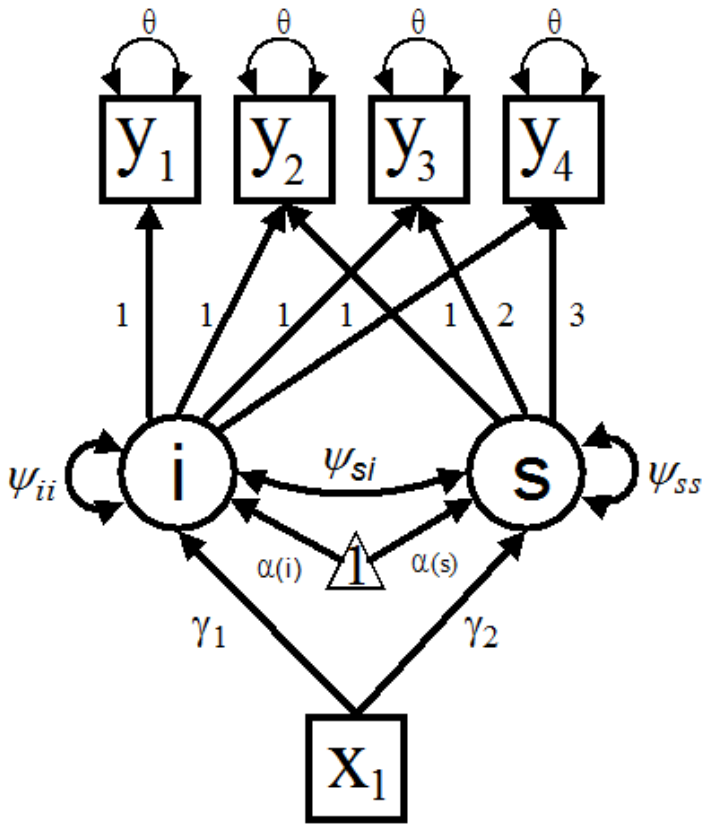
id	y1	y2	y3	x1	x2	x3
1	y11	y12	y13	x11	x12	x13
2	y21	y22	y23	x21	x22	x23
...						
n	yn1	yn2	yn3	xn1	xn2	xn3

Multilevel Model



$$y_{it} = \mathbf{i}_i + \mathbf{s}_i t_i + \epsilon_i$$
$$\mathbf{i}_i = \alpha(\mathbf{i}) + x_{1i}\gamma_1 + \zeta(\mathbf{i})_i$$
$$\mathbf{s}_i = \alpha(\mathbf{s}) + x_{1i}\gamma_2 + \zeta(\mathbf{s})_i$$

LGMC



$$y_{it} = \mathbf{i}_i + \mathbf{s}_i t + \epsilon_i$$
$$\mathbf{i}_i = \alpha(\mathbf{i}) + x_{1i}\gamma_1 + \zeta(\mathbf{i})_i$$
$$\mathbf{s}_i = \alpha(\mathbf{s}) + x_{1i}\gamma_2 + \zeta(\mathbf{s})_i$$

```
TITLE:    Multilevel Growth Model (ex02-03b.inp)
DATA:     FILE = ex02-03b.dat ;
VARIABLE: NAMES = id age msqtot male black ;
          MISSING ARE ALL (-9999) ;
          WITHIN = age ;
          BETWEEN = male black ;
          CLUSTER = id ;
ANALYSIS: TYPE = twolevel random ;
MODEL:    %within%
          s | msqtot on age ;
          %between%
          msqtot on male black ;
          s on male black ;
          msqtot with s ;
```