Further Readings On Multilevel Regression Analysis

Ludtke Marsh, Robitzsch, Trautwein, Asparouhov, Muthen (2007). Analysis of group level effects using multilevel modeling: Probing a latent covariate approach. Submitted for publication.

Raudenbush, S.W. & Bryk, A.S. (2002). Hierarchical linear models: Applications and data analysis methods. Second edition. Newbury Park, CA: Sage Publications.

Snijders, T. & Bosker, R. (1999). Multilevel analysis. An introduction to basic and advanced multilevel modeling. Thousand Oakes, CA: Sage Publications.

37

Logistic And Probit Regression

38

Categorical Outcomes: Logit And Probit Regression

Probability varies as a function of x variables (here x_1, x_2)

$$P(u = 1 \mid x_1, x_2) = F[\beta_0 + \beta_1 x_1 + \beta_2 x_2], \qquad (22)$$

$P(u = 0 \mid x_1 , x_2) = 1 - P[u = 1 \mid x_1 , x_2]$, where F[z] is either the standard normal ($\Phi[z]$) or logistic ($1/[1 + e^{-z}]$) distribution function.

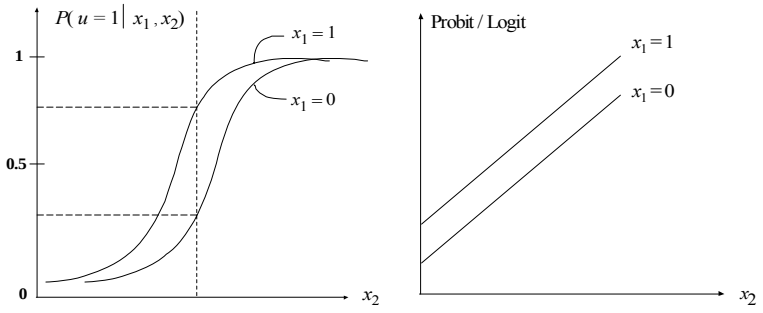
Example: Lung cancer and smoking among coal miners

u	lung cancer ($u = 1$) or not ($u = 0$)
x_1	smoker ($x_1 = 1$), non-smoker ($x_1 = 0$)
x_2	years spent in coal mine

39

Categorical Outcomes: Logit And Probit Regression

$$P(u = 1 \mid x_1, x_2) = F [\beta_0 + \beta_1 x_1 + \beta_2 x_2], \qquad (22)$$



40

Interpreting Logit And Probit Coefficients

- Sign and significance
- Odds and odds ratios
- Probabilities

41

Logistic Regression And Log Odds

$$Odds\ (u = 1 \mid x) = P(u = 1 \mid x) / P(u = 0 \mid x)$$
$$= P(u = 1 \mid x) / (1 - P(u = 1 \mid x)).$$

The logistic function

$$P(u = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

gives a log odds linear in x,

$$logit = log\ [odds\ (u = 1 \mid x)] = log\ [P(u = 1 \mid x) / (1 - P(u = 1 \mid x))]$$

$$= \log \left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \ / \ (1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}) \right]$$

$$= \log \left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} * \frac{1 + e^{-(\beta_0 + \beta_1 x)}}{e^{-(\beta_0 + \beta_1 x)}} \right]$$

$$= \log \left[e^{(\beta_0 + \beta_1 x)} \right] = \beta_0 + \beta_1 x$$

42

Logistic Regression And Log Odds (Continued)

- $logit = log\ odds = \beta_0 + \beta_1 x$
- When x changes one unit, the *logit* (*log odds*) changes β_1 units
- When x changes one unit, the *odds* changes e^{β_1} units

43

Two-Level Logistic Regression

With j denoting cluster,

$\text{logit}_{ij} = \log (P(u_{ij} = 1)/P(u_{ij} = 0)) = \alpha_j + \beta_j * x_{ij}$

where

$\alpha_j = \alpha + u_{0j}$
$\beta_j = \beta + u_{1j}$

High/low α_j value means high/low logit (high log odds)

44

Predicting Juvenile Delinquency From First Grade Aggressive Behavior

- Cohort 1 data from the Johns Hopkins University Preventive Intervention Research Center
- n= 1,084 students in 40 classrooms, Fall first grade
- Covariates: gender and teacher-rated aggressive behavior

45

Input For Two-Level Logistic Regression

```
TITLE:
        Hopkins Cohort 1 2-level logistic regression
DATA:
        FILE = Cohort1_classroom_ALL.DAT;
VARIABLE:

        NAMES =  prcid juv99 gender stub1F bkRule1F harmO1F
                 bkThin1F yell1F takeP1F fight1F lies1F
                 tease1F;
        CLUSTER = classrm;
        USEVAR = juv99 male aggress;
        CATEGORICAL =  juv99;
        MISSING = ALL (999);
        WITHIN = male aggress;
DEFINE:
        male = 2 - gender;
        aggress = stub1F + bkRule1F + harmO1F + bkThin1F +
                  yell1F + takeP1F + fight1F + lies1F + tease1F;
```

46

Input For Two-Level Logistic Regression (Continued)

```
ANALYSIS:
        TYPE = TWOLEVEL MISSING;
        PROCESS = 2;
MODEL:
        %WITHIN%
        juv99 ON male aggress;
        %BETWEEN%
OUTPUT:
        TECH1 TECH8;
```

47

Output Excerpts Two-Level Logistic Regression

```
MODEL RESULTS
                    Estimates    S.E    Est./S.E.
Within Level
  JUV99         ON
    MALE             1.071    0.149     7.193
    AGGRESS          0.060    0.010     6.191

Between Level
Thresholds
  JUV99$1            2.981    0.205    14.562

Variances
  JUV99              0.807    0.250     3.228
```

48

Understanding The Between-Level Intercept Variance

- Intra-class correlation
  - ICC = $0.807/(\pi^2/3 + 0.807)$
- Odds ratios
  - Larsen & Merlo (2005). Appropriate assessment of neighborhood effects on individual health: Integrating random and fixed effects in multilevel logistic regression. American Journal of Epidemiology, 161, 81-88.
  - Larsen proposes MOR:
    "Consider two persons with the same covariates, chosen randomly from two different clusters. The MOR is the median odds ratio between the person of higher propensity and the person of lower propensity."

    $MOR = \exp(\sqrt{(2* \sigma^2)} * \Phi^{-1} (0.75))$

    In the current example, ICC = 0.20, MOR = 2.36
- Probabilities
  - Compare α_j=1 SD and α_k=-1 SD from the mean

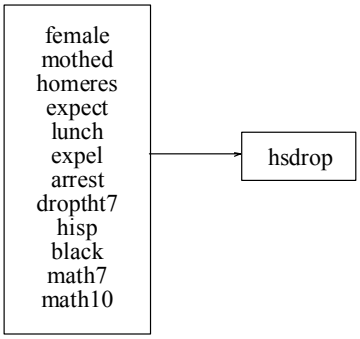
49

Two-Level Path Analysis

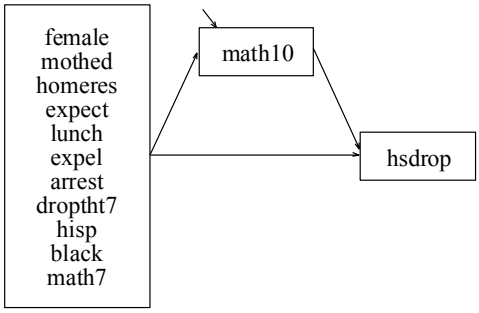
50

A Path Model With A Binary Outcome And A Mediator With Missing Data

Logistic Regression



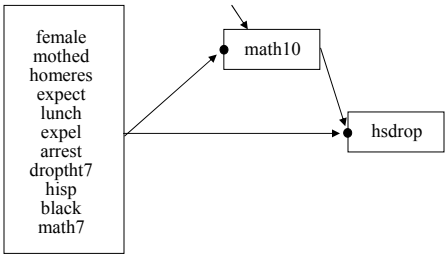
Path Model



51

Two-Level Path Analysis

Within



Between

52

Input For A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable

```
TITLE:      a twolevel path analysis with a categorical outcome
            and missing data on the mediating variable
DATA:       FILE = lsayfull_dropout.dat;
VARIABLE:   NAMES = female mothed homeres math7 math10 expel
            arrest hisp black hsdrop expect lunch droptht7
            schcode;
            MISSING = ALL (9999);
            CATEGORICAL = hsdrop;
            CLUSTER = schcode;
            WITHIN = female mothed homeres expect math7 lunch
            expel arrest droptht7 hisp black;
ANALYSIS:   TYPE = TWOLEVEL MISSING;
            ESTIMATOR = ML;
            ALGORITHM = INTEGRATION;
            INTEGRATION = MONTECARLO (500);
```


Input For A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable (Continued)

```
MODEL:
            %WITHIN%
            hsdrop ON female mothed homeres expect math7 math10
            lunch expel arrest droptht7 hisp black;
            math10 ON female mothed homeres expect math7 lunch
            expel arrest droptht7 hisp black;

            %BETWEEN%
            hsdrop*1; math10*1;


OUTPUT:     PATTERNS SAMPSTAT STANDARDIZED TECH1 TECH8;
```

Output Excerpts A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable

Summary Of Data

```
     Number of patterns             2
     Number of clusters            44

Size (s)       Cluster ID with Size s
   12            304
   13            305
   36            307    122
   38            106    112
   39            138    109
   40            103
   41            308
   42            146    120
   43            102    101
   44            303    143
   45            141
```

55

Output Excerpts A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable (Continued)

```
Size (s)       Cluster ID with Size s
   46            144
   47            140
   49            108
   50            126    111    110
   51            127    124
   52            137    117    147    118    301    136
   53            142    131
   55            145    123
   57            135    105
   58            121
   59            119
   73            104
   89            302
   93            309
  118            115
```

56

Output Excerpts A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable (Continued)

Model Results

	Estimates	S.E.	Est./S.E.	Std	StdYX
Within Level					
HSDROP ON					
FEMALE	0.323	0.171	1.887	0.323	0.077
MOTHED	-0.253	0.103	-2.457	-0.253	-0.121
HOMERES	-0.077	0.055	-1.401	-0.077	-0.061
EXPECT	-0.244	0.065	-3.756	-0.244	-0.159
MATH7	-0.011	0.015	-0.754	-0.011	-0.055
MATH10	-0.031	0.011	-2.706	-0.031	-0.197
LUNCH	0.008	0.006	1.324	0.008	0.074
EXPEL	0.947	0.225	4.201	0.947	0.121
ARREST	0.068	0.321	0.212	0.068	0.007
DROPTHT7	0.757	0.284	2.665	0.757	0.074
HISP	-0.118	0.274	-0.431	-0.118	-0.016
BLACK	-0.086	0.253	-0.340	-0.086	-0.013

57

Output Excerpts A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable (Continued)

	Estimates	S.E.	Est./S.E.	Std	StdYX
MATH10 ON					
FEMALE	-0.841	0.398	-2.110	-0.841	-0.031
MOTHED	0.263	0.215	1.222	0.263	0.020
HOMERES	0.568	0.136	4.169	0.568	0.070
EXPECT	0.985	0.162	6.091	0.985	0.100
MATH7	0.940	0.023	40.123	0.940	0.697
LUNCH	-0.039	0.017	-2.308	-0.039	-0.059
EXPEL	-1.293	0.825	-1.567	-1.293	-0.026
ARREST	-3.426	1.022	-3.353	-3.426	-0.054
DROPTHT7	-1.424	1.049	-1.358	-1.424	-0.022
HISP	-0.501	0.728	-0.689	-0.501	-0.010
BLACK	-0.369	0.733	-0.503	-0.369	-0.009

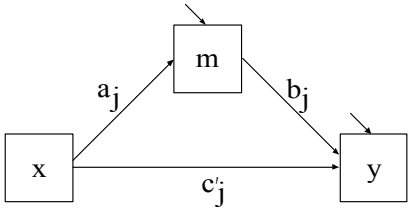
58

Output Excerpts A Two-Level Path Analysis Model With A Categorical Outcome And Missing Data On The Mediating Variable (Continued)

	Estimates	S.E.	Est./S.E.	Std	StdYX
Residual Variances					
MATH10	62.010	2.162	28.683	62.010	0.341
Between Level					
Means					
MATH10	10.226	1.340	7.632	10.226	5.276
Thresholds					
HSDROP$1	-1.076	0.560	-1.920		
Variances					
HSDROP	0.286	0.133	2.150	0.286	1.000
MATH10	3.757	1.248	3.011	3.757	1.000

59

Two-Level Mediation



Indirect effect:

$$\alpha + \beta + Cov\,(a_j, b_j)$$

Bauer, Preacher & Gil (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. Psychological Methods, 11, 142-163.

60

Input For Two-Level Mediation

```
MONTECARLO:
        NAMES ARE y m x;
        WITHIN = x;
        NOBSERVATIONS = 1000;
        NCSIZES = 1;
        CSIZES = 100 (10);
        NREP = 100;

MODEL POPULATION:
        %WITHIN%
        c | y ON x;
        b | y ON m;
        a | m ON x;
        x*1; m*1; y*1;
        %BETWEEN%
        y WITH m*0.1 b*0.1 a*0.1 c*0.1;
        m WITH b*0.1 a*0.1 c*0.1;
        a WITH b*0.1 c*0.1;
        b WITH c*0.1;
        y*1 m*1 a*1 b*1 c*1;
        [a*0.4 b*0.5 c*0.6];
```

61

Input For Two-Level Mediation (Continued)

```
ANALYSIS:
       TYPE = TWOLEVEL RANDOM;
MODEL:
        %WITHIN%
        c | y ON x;
        b | y ON m;
        a | m ON x;
        m*1; y*1;
        %BETWEEN%
        y WITH M*0.1 b*0.1 a*0.1 c*0.1;
        m WITH b*0.1 a*0.1 c*0.1;
        a WITH b*0.1 (cab);
        a WITH c*0.1;
        b WITH c*0.1;
        y*1 m*1 a*1 b*1 c*1;
        [a*0.4] (ma);
        [b*0.5] (mb);
        [c*0.6];

MODEL CONSTRAINT:
        NEW(m*0.3);
        m=ma*mb+cab;
```

62

Output Excerpts Two Level Mediation

	Estimates			S.E.	M. S. E.	95%	% Sig
	Population	Average	Std.Dev.	Average		Cover	Coeff
Within Level							
Residual variances							
Y	1.000	1.0020	0.0530	0.0530	0.0028	0.960	1.000
M	1.000	1.0011	0.0538	0.0496	0.0029	0.910	1.000
Between Level							
Y WITH							
B	0.100	0.1212	0.1246	0.114	0.0158	0.910	0.210
A	0.100	0.1086	0.1318	0.1162	0.0173	0.910	0.190
C	0.100	0.0868	0.1121	0.1237	0.0126	0.940	0.090
M WITH							
B	0.100	0.1033	0.1029	0.1085	0.0105	0.940	0.120
A	0.100	0.0815	0.1081	0.1116	0.0119	0.950	0.070
C	0.100	0.1138	0.1147	0.1165	0.0132	0.970	0.160
A WITH							
B	0.100	0.0964	0.1174	0.1101	0.0137	0.920	0.150
C	0.100	0.0756	0.1376	0.1312	0.0193	0.910	0.110

63

Output Excerpts Two-Level Mediation (Continued)

B WITH							
C	0.100	0.0892	0.1056	0.1156	0.0112	0.960	0.070
Y WITH							
M	0.100	0.1034	0.1342	0.1285	0.0178	0.940	0.140
Means							
Y	0.000	0.0070	0.1151	0.1113	0.0132	0.950	0.050
M	0.000	-0.0031	0.1102	0.1056	0.0120	0.950	0.050
C	0.600	0.5979	0.1229	0.1125	0.0150	0.930	1.000
B	0.500	0.5022	0.1279	0.1061	0.0162	0.890	1.000
A	0.400	0.3854	0.0972	0.1072	0.0096	0.970	0.970
Variances							
Y	1.000	1.0071	0.1681	0.1689	0.0280	0.910	1.000
M	1.000	1.0113	0.1782	0.1571	0.0316	0.930	1.000
C	1.000	0.9802	0.1413	0.1718	0.0201	0.980	1.000
B	1.000	0.9768	0.1443	0.1545	0.0212	0.950	1.000
A	1.000	1.0188	0.1541	0.1587	0.0239	0.950	1.000
New/Additional Parameters							
M	0.300	0.2904	0.1422	0.1316	0.0201	0.950	0.550

64

Two-Level Factor Analysis

65

Two-Level Factor Analysis

- Recall random effects ANOVA (individual i in cluster j):

$$y_{ij} = v + \eta_j + \varepsilon_{ij} = y_{B_j} + y_{W_{ij}}$$

- Two-level factor analysis ($r = 1, 2, \ldots, p$ items):

$$y_{rij} = v_r + \lambda_{B_r}\,\eta_{B_j} + \varepsilon_{B_{rj}} + \lambda_{W_r}\,\eta_{Wij} + \varepsilon_{Wrij}$$

(between-cluster variation) (within-cluster variation)

66

Two-Level Factor Analysis (Continued)

- Covariance structure:

$$V(\boldsymbol{y}) = V(\boldsymbol{y}_B) + V(\boldsymbol{y}_w) = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_w,$$

$$\boldsymbol{\Sigma}_B = \boldsymbol{\Lambda}_B \ \boldsymbol{\Psi}_B \ \boldsymbol{\Lambda}_B' + \boldsymbol{\Theta}_B,$$

$$\boldsymbol{\Sigma}_W = \boldsymbol{\Lambda}_W \ \boldsymbol{\Psi}_W \ \boldsymbol{\Lambda}_W' + \boldsymbol{\Theta}_W .$$

- Two interpretations:
  - variance decomposition, including decomposing the residual
  - random intercept model

67

Two-Level Factor Analysis And Design Effects

Muthén & Satorra (1995; Sociological Methodology): Monte Carlo study using two-level data (200 clusters of varying size and varying intraclass correlations), a latent variable model with 10 variables, 2 factors, conventional ML using the regular sample covariance matrix S_T, and 1,000 replications (d.f. = 34).

$$\Lambda_B = \Lambda_W = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Ψ_B, Θ_B reflecting different icc's

$$y_{ij} = \nu + \Lambda(\eta_{Bj} + \eta_{Wij}) + \varepsilon_{Bj} + \varepsilon_{Wij}$$

$$V(y) = \Sigma_B + \Sigma_W = \Lambda(\Psi_B + \Psi_W)\Lambda' + \Theta_B + \Theta_W$$

68

Two-Level Factor Analysis And Design Effects (Continued)

Inflation of χ^2 due to clustering

Intraclass Correlation		Cluster Size			
		7	15	30	60
0.05					
	Chi-square mean	35	36	38	41
	Chi-square var	68	72	80	96
	5%	5.6	7.6	10.6	20.4
	1%	1.4	1.6	2.8	7.7
0.10					
	Chi-square mean	36	40	46	58
	Chi-square var	75	89	117	189
	5%	8.5	16.0	37.6	73.6
	1%	1.0	5.2	17.6	52.1
0.20					
	Chi-square mean	42	52	73	114
	Chi-square var	100	152	302	734
	5%	23.5	57.7	93.1	99.9
	1%	8.6	35.0	83.1	99.4

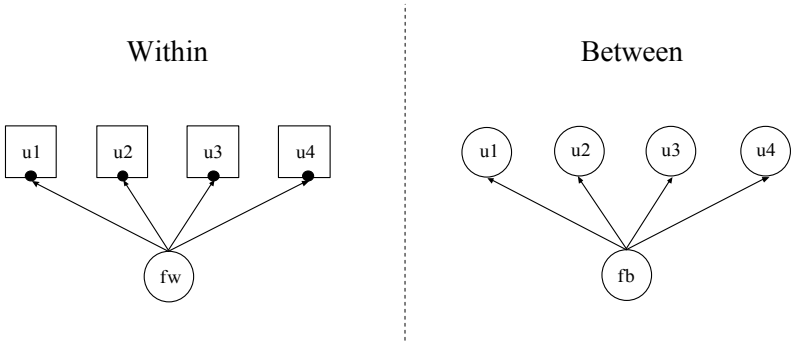
69

Two-Level Factor Analysis And Design Effects (Continued)

- Regular analysis, ignoring clustering
  - Inflated chi-square, underestimated SE's
- TYPE = COMPLEX
  - Correct chi-square and SE's but only if model aggregates, e.g. $\Lambda_B = \Lambda_W$
- TYPE = TWOLEVEL
  - Correct chi-square and SE's

70

Two-Level Factor Analysis (IRT)



$$u^*_{ij} = \lambda\,(f_{B_j} + f_{w_{ij}}) + \varepsilon_{ij}$$

71

Input For A Two-Level Factor Analysis (IRT) Model With Categorical Outcomes

```
TITLE:      this is an example of a two-level factor analysis
            model with categorical outcomes
DATA:       FILE = catrep1.dat;
VARIABLE:   NAMES ARE u1-u6 clus;
            CATEGORICAL = u1-u6;
            CLUSTER = clus;
ANALYSIS:   TYPE = TWOLEVEL;
            ESTIMATION = ML;
            ALGORITHM = INTEGRATION;
MODEL:
            %WITHIN%
            fw BY u1@1
            u2 (1)
            u3 (2)
            u4 (3)
            u5 (4)
            u6 (5);
```

72

Input For A Two-Level Factor Analysis (IRT) Model With Categorical Outcomes (Continued)

```
            %BETWEEN%
            fb BY u1@1
            u2 (1)
            u3 (2)
            u4 (3)
            u5 (4)
            u6 (5);
OUTPUT:     TECH1 TECH8;
```

73

Output Excerpts A Two-Level Factor Analysis (IRT) Model With Categorical Outcomes

Tests Of Model Fit

```
Loglikelihood

            H0 Value                          -3696.117
Information Criteria

            Number of Free Parameters                13
            Akaike (AIC)                       7418.235
            Bayesian (BIC)                     7481.505
            Sample-Size Adjusted BIC           7440.217
               (n* = (n + 2) / 24)
```

74

Output Excerpts A Two-Level Factor Analysis (IRT) Model With Categorical Outcomes (Continued)

Model Results

	Estimates	S.E.	Est./S.E.
Within Level			
FW BY			
U1	1.000	0.000	0.000
U2	0.915	0.146	6.264
U3	1.087	0.169	6.437
U4	1.058	0.164	6.441
U5	1.191	0.185	6.449
U6	1.143	0.178	6.439
Variances			
FW	0.834	0.191	4.360

75

Output Excerpts Two-Level Factor Analysis (IRT) Model With Categorical Outcomes (Continued)

	Estimates	S.E.	Est./S.E.
Between Level			
FB BY			
U1	1.000	0.000	0.000
U2	0.915	0.146	6.264
U3	1.087	0.169	6.437
U4	1.058	0.164	6.441
U5	1.191	0.185	6.449
U6	1.143	0.178	6.439
Thresholds			
U1$1	-0.206	0.096	-2.150
U2$1	0.001	0.091	0.007
U3$1	-0.016	0.100	-0.156
U4$1	-0.064	0.098	-0.652
U5$1	-0.033	0.105	-0.315
U6$1	-0.021	0.102	-0.209
Variances			
FB	0.496	0.139	3.562

76

SIMS Variance Decomposition

The Second International Mathematics Study (SIMS; Muthén, 1991, JEM).

- National probability sample of school districts selected proportional to size; a probability sample of schools selected proportional to size within school district, and two classes randomly drawn within each school
- 3,724 students observed in 197 classes from 113 schools with class sizes varying from 2 to 38; typical class size of around 20
- Eight variables corresponding to various areas of eighth-grade mathematics
- Same set of items administered as a pretest in the Fall of eighth grade and as a posttest in the Spring.

77

SIMS Variance Decomposition (Continued)

Muthén (1991). Multilevel factor analysis of class and student achievement components. Journal of Educational Measurement, 28, 338-354.

- Research questions: "The substantive questions of interest in this article are the variance decomposition of the subscores with respect to within-class student variation and between-class variation and the change of this decomposition from pretest to posttest. In the SIMS … such variance decomposition relates to the effects of tracking and differential curricula in eighth-grade math. On the one hand, one may hypothesize that effects of selection and instruction tend to increase between-class variation relative to within-class variation, assuming that the classes are homogeneous, have different performance levels to begin with, and show faster growth for higher initial performance level. On the other hand, one may hypothesize that eighth-grade exposure to new topics will increase individual differences among students within each class so that posttest within-class variation will be sizable relative to posttest between-class variation."

78

SIMS Variance Decomposition (Continued)

$$y_{rij} = v_r + \lambda_{Br}\,\eta_{Bj} + \varepsilon_{Brj} + \lambda_{wr}\,\eta_{wij} + \varepsilon_{wrij}$$

$$V(y_{rij}) = \quad \text{BF} \quad + \text{BE} \quad + \text{WF} \quad + \text{WE}$$

Between reliability: BF / (BF + BE)
– BE often small (can be fixed at 0)

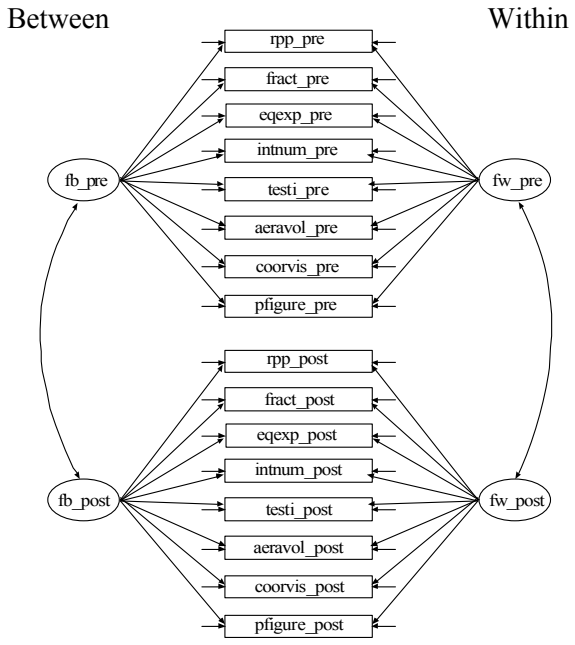
Within reliability: WF / (WF + WE)
– sum of a small number of items gives a large WE

Intraclass correlation:
ICC = (BF + BE) / (BF + BE + WF+ WE)

Large measurement error → large WE → small ICC

True ICC = BF / (BF + WF)

79



80