Using Mplus To Do Dynamic Structural Equation Modeling

Bengt Muthén
Professor Emeritus, UCLA
Mplus: https://www.statmodel.com
bmuthen@statmodel.com

Tihomir Asparouhov
Mplus

Mplus Web Talks: No. 6
February 2023

We thank Ellen Hamaker and Loes Keijsers for helpful comments and guidance on data sets, and Thuy Nguyen and Noah Hastings for expert assistance.

Outline

- **Section 1**: Introducing the example
  - At-risk mood profiles related to depression in adolescents
- **Section 2**: How to handle varying (random) times of measurement
  - Understanding the TINTERVAL option
    - also useful for creating time manually
- **Section 3**: Descriptive analyses to understand the data
  - TYPE = TWOLEVEL BASIC
  - Histograms
- **Section 4**: TYPE = TWOLEVEL (RANDOM), regular twolevel analysis
  - Estimating the intraclass correlation, random variance, random covariance

Outline Continued

- **Section 5**: Two-Level Analysis Bringing Time Into the Model
  - Univariate DSEM
  - Model diagram, equations, Mplus input, plots
- **Section 6**: Two-level DSEM and RDSEM analysis
  - Bivariate DSEM/RDSEM
- **Section 7**: Categorical outcome
- **Section 8**: Cross-classified analysis
  - Looking for trends over time
- **Section 9**: How large do N and T have to be?
  - Checklist
  - Monte Carlo simulations
- **Section 10**: References

Section 1 Introducing the Example

Grumpy or Depressed Data

- Data from a study designed to detect at-risk mood profiles related to depression in adolescents
- ESM questionnaires measuring positive and negative affect in Dutch adolescents ages 12 to 16, 63% girls
- Positive and negative affect are measured as an average of six 7-category items
- Tiredness is measured on a 7-point scale
- N = 240, several measures per day for 7 days, Tuesday - Monday
- Covariates: Gender, SDQ (measure of childhood emotional problems)
- de Haan-Rietdijk, Voelkle, Keijsers, Hamaker (2017). Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. frontiers in Psychology

Schedule of Measurements

- Participants filled out ESM questionnaires throughout the day, including during school hours
- Questionnaires delivered on the adolescents' own smartphones
- Eight measurements taken randomly in 3 blocks of time between 8 am and 10pm
  - A morning measurement between 8 am and 10 am
  - Six measurements between 10 am and 8 pm
  - An evening measurement between 8 pm and 10 pm

Section 2 How to Handle Varying Times of Measurement

Varying Times of Measurement: Understanding the TINTERVAL Option

- Times of measurement are not the same across people - data are misaligned with respect to time due to random measurement occasions
- Varying times of observation are common in longitudinal data but needs special attention in DSEM due to its use of lagged effects:
  - The time distance between measurements at time $t-1$ and time t needs to be the same across people
  - This can be accomplished by inserting missing data to align time: The TINTERVAL option

Hypothetical Case 1: Frequent Measurements

	Original data				Aligned data			
	ID = 1		ID = 2		ID = 1		ID = 2	
Data row	t	y	t	y	t	y	t	y
1	9	$y_{1,9}$	9	$y_{2,9}$	9	$y_{1,9}$	9	$y_{2,9}$
2	10	$y_{1,10}$	11	$y_{2,11}$	10	$y_{1,10}$	*	*
3	11	$y_{1,11}$	12	$y_{2,12}$	11	$y_{1,11}$	11	$y_{2,11}$
4	13	$y_{1,13}$			*	*	12	$y_{2,12}$
5					13	$y_{1,13}$	*	*

- Insertion of missing data rows using a time interval of 1 is obtained by the VARIABLE command option TINTERVAL = t(1); where t is the name of the time variable (e.g. hour) in the data set
  - Data row = bin number = time used in the analysis

Hypothetical Case 2: Infrequent Measurements

- With infrequent measurements, time interval = 1 gives many missing data rows: Wider time interval needed

	Original data		Tinterval = 2		Aligned data	
Data row	t	y	Bin	Bin range		
1	9	y_9	1	9-10	9	y_9
2	13	y_{13}	2	11-12	*	*
3	16	y_{16}	3	13-14	13	y_{13}
4	19	y_{19}	4	15-16	16	y_{16}
5	21	y_{21}	5	17-18	*	*
			6	19-20	19	y_{19}
			7	21-22	21	y_{21}

Choosing Tinterval for the Grumpy or Depressed Data

- The total number of hours for the study is 24 x 7 = 168
- The smaller the time interval, the better the match to the data, but the more missing data inserted
  - Time interval = 1 gives 24/1 = 24 time points per day. Maximum number of time points = 24 x 7 = 168. Coverage = 0.171
  - Time interval = 2 gives 24/2 = 12 time points per day. Maximum number of time points = 12 x 7 = 84. Coverage = 0.341
  - Time interval = 3 gives 24/3 = 8 time points per day which was the aim of the study. Maximum number of time points = 8 x 7 = 56. Coverage = 0.507
- Coverage = proportion not missing

Tinterval Examples: ID = 41 with Interval 3 (3 Hours)

Bin # (bin time)	Bin range	Observed time	Observed time rearranged	New time mid point of bin range
1	0-2	*	*	1.5
2	3-5	*	**7**	4.5
3	6-8	**7, 8**	8	7.5
4	9-11	11	11	10.5
5	12-14	**12, 14**	12	13.5
6	15-17	16	**14**	16.5
7	18-20	18	**16**	19.5
8	21-23	*	**18**	22.5 (SMSE=4.5)
9	24-26	*	*	25.5
10	27-29	*	**31**	28.5
11	30-32	**31, 32** (7, 8 am)	32	31.5
...				

- The range is more precisely stated as [low-high), e.g. [21 - 24): ≥ 21 and < 24
- If there is more than 1 observed time per bin, rearrange observed time into neighboring bins: 1 move due to bin 3 and 3 moves due to bin 5
- 4 of 8 observed times are misaligned in the first day: 7, 14, 16, and 18
- Misalignment is quantified by the difference between the observed time and the new time midpoint: Time interval plot with SMSE (squared root of mean square error = distance)

Tinterval Examples: ID = 41 with Interval 2

Bin # (bin time)	Bin range	Observed time
1	0-1	*
2	2-3	*
3	4-5	*
4	6-7	7
5	8-9	8
6	10-11	11
7	12-13	12
8	14-15	14
9	16-17	16
10	18-19	18
11	20-21	*
12	22-23	*
13	24-25	*
14	26-27	*
15	28-29	*
16	30-31	31
17	32-33	32
...		

- Perfect alignment, no need to rearrange observed time

ID = 20, Tinterval = 2, Day 1

Bin # (bin time)	Bin range	Observed time	Observed time rearranged	New time mid point of bin range
1	0-1	*	*	*
2	2-3	*	*	*
3	4-5	5	5	5
4	6-7	7	7	7
5	8-9	*	**10**	9
6	10-11	**10,11**	11	11
7	12-13	13	13	13
8	14-15	*	*	*
9	16-17	16	16	17
10	18-19	18	18	19
11	20-21	*	*	*
12	22-23	*	*	*

- 1 misalignment: 10

ID=20, Tinterval = 2, Day 2

Bin # (bin time)	Bin range	Observed time	Observed time rearranged
13	24-25	*	*
14	26-27	*	*
15	28-29	*	*
16	30-31	30	30
17	32-33	33	33
18	34-35	*	**36**
19	36-37	36	**38**
20	38-39	**38, 39**	39
21	40-41	41	41
22	42-43	*	*
23	44-45	*	*
24	46-47	*	*

- 2 misalignments. 2 moves.

Section 3 Descriptive Analyses to Understand the Data TYPE = TWOLEVEL BASIC

Data Structure and Analysis Steps

- Data in long format: Time (hours) nested in individuals (ID)

ID	HOURS	PA
1	9	4
1	10	5
1	11	5
1	13	6
2	9	5
2	11	7
3	...	

- Step 1: TYPE = TWOLEVEL BASIC to check histograms, within- and between-level variation and between-level plots (MLR)
- Step 2: TYPE = TWOLEVEL (RANDOM) modeling of variation in mean, variance, and correlation (Bayes), regular twolevel analysis
- Step 3: TYPE = TWOLEVEL (RANDOM) DSEM analysis, bringing in time (Bayes): Univariate
- Step 4: TYPE = TWOLEVEL (RANDOM) DSEM analysis, bringing in time (Bayes): Bivariate
- Step 5: TYPE = CROSSCLASSIFIED (Bayes)

Step 1: TYPE = TWOLEVEL BASIC for PA, NA, Tired, and Hrs. CLUSTER = ID

```
            USEVARIABLES = pa na tired hrs;
            CLUSTER = id;
ANALYSIS:
            TYPE = TWOLEVEL BASIC;
PLOT:
            TYPE = PLOT3;
```

```
*** WARNING
  One or more individual-level variables have no variation within a
  cluster for the following clusters.

   Variable        Cluster IDs with no within-cluster variation
     PA              240 249 78 531
     NA              240 442 45 249 319 531 263 200 160 2 320 352 260 119
                     254 503 347 385 570 523 313 338 462 256
     TIRED           29 383 442 249 531 24 160 320 165 406 454
```

- These individuals may be deleted (discussed later)

Step 1: Summary of Data

- 240 individuals with an average cluster size (number of time points) of 24

Size (s)	Cluster ID with Size s
1	238 549
2	161 29 **240** 383 442 **45** 572
3	206 **249**
4	319 176 197 **78** 276
5	456 **531** 415 571 196
6	263 24 200 380 51 388 392
7	359 100 160 224
9	92
10	71 2 13 320
11	12 7 204 341 352 357
12	205 22 561 28 485
13	314 260 119 433 254 348
14	503 107 47 552 277 137 188
15	5 361
16	36 401 507 518 403 373 428 280 221 347
17	407
18	165 351 521 528 4 447 452 554 560 302 414 155
19	1 49 186 406 295 203 310 385 152
20	514 122 31 207 454 547 144 457 463 258 491 360 56
21	445 175
22	411 32 266 520 64
23	371 102 211 243 570 469 470
24	557 235 374 340 85 556

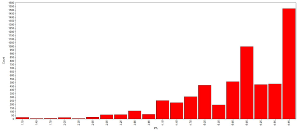
Size (s)	Cluster ID with Size s
25	129 43 393 159 111 309 555
26	465 386 419 475 523 558 83 148 542 313 327
27	95 213 178 455 296 446
28	6 372 338 94
29	279 417 52 365 499 429 87
30	232 480 460 40
31	66 30 305
32	278 462 75 177 286 234 389 130 300 398 546 23
33	473
34	543 126 288
35	317 439 298
36	225 227 163 394 69 65 118 **20**
37	287 468 50 171 89 438 101 573
38	209 307 297 541 162 70 16
39	109 246 440 256
40	510 537
41	181 220
43	508 46 431 358 79 328 346
44	478 382 252 27 228
45	81 501 322 443
46	331 **41** 525 466
47	343 448
48	370
49	106

Step 1: Intraclass Correlations

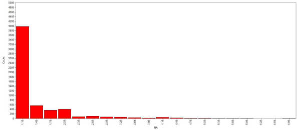
Estimated Intraclass Correlations for the Y Variables

Variable	Intraclass Correlation	Variable	Intraclass Correlation	Variable	Intraclass Correlation
PA	0.570	NA	0.467	TIRED	0.477

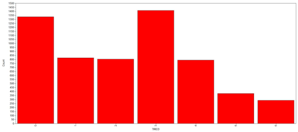
- ICC = V(Between)/(V(Between) + V(Within))
  = Variation across individuals / Total variation
- Variation across individuals is viewed as variation of a random mean, a variable corresponding to individuals' mean across time, that is, variation across individuals in the average level; a trait variable



PA: 22% ceiling effect.

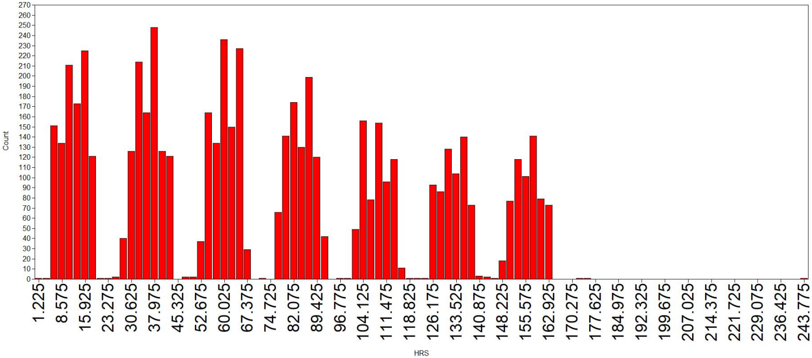


NA: 60% floor effect.



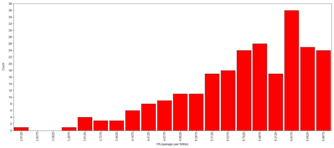
Tired: 23% floor effect.

Step 1: Histogram for Hours Variable

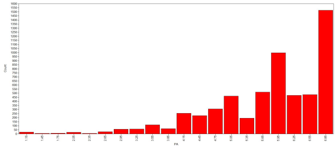


- Max time = 168 but some beyond (check data)
- Dropoff of people over time
- Few observations late at night and early morning
- Histogram display properties changed from default of 20 to 100 bins

- Between-level histogram of person average over time
- Less skewed than overall histogram - only 2% at max value vs 22%



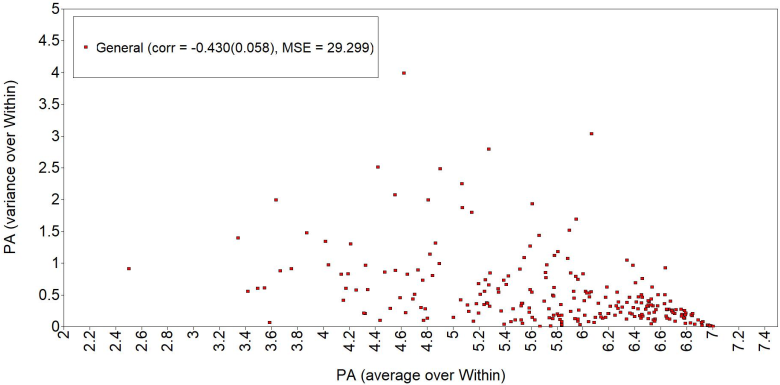
(a) Based on average over time



(b) Based on individual values

- Between-level histogram of averages over time corresponds to the distribution of random intercepts which is assumed to be normal

Step 1: Between-Level Scatter Plot of PA Mean and Variance



- The variance varies over people and is correlated -0.43 with the mean
- NA scatterplot shows a correlation of 0.65. Tired shows no correlation

Section 4 Regular twolevel analysis
TYPE = TWOLEVEL (RANDOM)

Step 2: TWOLEVEL Modeling of PA, NA, and Tired
Model 1: Random Mean

- What portion of the total variance is due to variation across individuals in their means across time ? What are the predictors of the variation?
  - Single outcome, random mean, fixed residual variance:

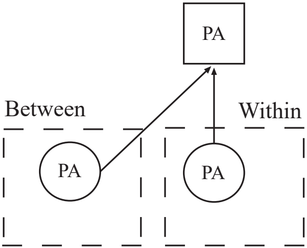
    $$y_{it} = \alpha_i + \varepsilon_{it}, \tag{1}$$

    where α_i is the random mean, that is, an average over time that varies across individuals, also seen as a latent trait. The residual $\varepsilon_{it} \sim N(0, \sigma^2)$ is assumed uncorrelated across individual and time
  - The random mean (latent trait) accounts for the correlation across time as expressed by the intraclass correlation

    $$ICC = V(\alpha)/V(y) = V(B)/(V(B) + V(W)) \tag{2}$$

  - The ICC is the R-square for the variance of the outcome explained by the random mean, that is, the R-square due to the latent trait
  - The random mean plays the role of the intercept growth factor ($\lambda = 1$) in a growth model analyzed in a twolevel, long format

- $y_{it} = \alpha_i + \varepsilon_{it}$
- y: observed PA variable (square)
- α: latent PA variable on Between (circle)
- ε: latent PA variable on Within (circle)
- Diagram drawn like RI-CLPM




```
                USEVARIABLES = pa;
                CLUSTER = id;
                ! 24*7 = 168 (same as time >84):
                USEOBSERVATIONS =
                hrs LE 168 AND id NE 240
                AND id NE 249 AND id NE 78
                AND id NE 531;
ANALYSIS:       TYPE = TWOLEVEL;
                ESTIMATOR = BAYES;
                BITERATIONS = (1000);
                PROCESSORS = 8;
MODEL:          %WITHIN%
                pa (w);
                %BETWEEN%
                pa(b);
MODEL
CONSTRAINT:     NEW(icc);
                icc = b/(b + w);
                ! PA icc estimate = 0.57 (.024)
                ! PA(W) = 0.56, PA(B) = 0.75
```

Step 2: TWOLEVEL Modeling of PA, NA, and Tired
Model 2: Random Residual Variance

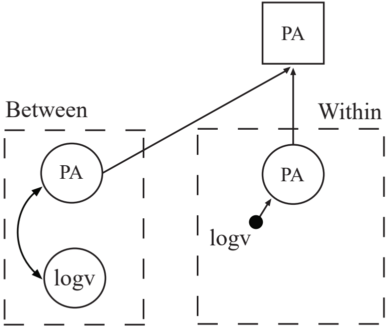
- How much does the residual variance vary across individuals? What are its predictors? What is the correlation between the mean and residual variance?
  - Single outcome, random mean, random residual variance:

$$y_{it} = \alpha_i + \varepsilon_{it}, \tag{3}$$

where α_i is the random mean and the residual $\varepsilon_{it} \sim N(0, \sigma_i^2)$, that is, the variance also varies across individuals ($log \sigma_i^2 \sim N(\mu, \sigma^2)$)

Step 2 Model 2 in Formula and Diagram Form

- $y_{it} = \alpha_i + \varepsilon_{it}$, $\varepsilon_{it} \sim N(0, \sigma_i^2)$
- y: observed PA variable
- α: latent PA variable on Between
- ε: latent PA variable on Within
- The random residual variance σ_i^2 is represented by logv



Step 2 Model 2 Input

```
                  USEVARIABLES = pa;
                  CLUSTER = id;
                  ! 24*7 = 168 (same as time >84):
                  USEOBSERVATIONS =
                  hrs LE 168
                  AND id NE 240 AND id NE 249
                  AND id NE 78 AND id NE 531;
ANALYSIS:
                  TYPE = TWOLEVEL RANDOM;
                  ESTIMATOR = BAYES;
                  BITERATIONS = (1000);
                  PROCESSORS = 8;
MODEL:
                  %WITHIN%
                  logv | pa; ! Random residual var
                  %BETWEEN%
                  pa WITH logv;
                  [logv] (m);
                  logv (s);
MODEL
CONSTRAINT:
                  NEW(meanv);
                  meanv = EXP(m + s/2);
OUTPUT:
                  STANDARDIZED TECH1
                  TECH4 TECH8;
PLOT:
                  TYPE = PLOT3;
```

Understanding logv

- Random effects are assumed to be normally distributed which is not suitable for a variance since has a chi-square like distribution and is never negative. Because of this, the log of the variance is instead modeled which is emphasized by calling it logv
- The mean m of the random variance logv can be negative because it is on the log scale (negative values for residual variances less than 1)
- To get the mean on the regular scale, the logv mean should be exponentiated. The correct exponentiation also involves s, the variance of logv: The mean of the variance = exp(m + s/2). The theory behind this expression draws on the mean of the log normal distribution.
- The median is exp(m) and is more useful for the skewed distribution of the variance
- The mean and median of the variance can be expressed in MODEL CONSTRAINT which also gives the confidence interval (Bayes allows a potentially non-symmetric CI)
- A fuller discussion of logv is presented later in connection with Step 3 Model 2

Step 2 Model 2 Output Excerpts

	Estimate	Posterior S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Within Level					
Between Level					
PA WITH					
LOGV	-0.464	0.074	-0.613	-0.330	*
Means					
PA	5.759	0.056	5.642	5.858	*
LOGV	-1.025	0.074	-1.168	-0.880	*
Variances					
PA	0.735	0.074	0.603	0.895	*
LOGV	1.108	0.115	0.902	1.339	*
New/Additional Parameters					
MEANV	0.626	0.058	0.527	0.760	*

- The S.D. column corresponds to SE for ML. Estimate/SD is approximately a z-score but the CI is a better statistic to report
- The correlation between the random mean PA and the random logv residual variance is -0.52 (the scatterplot correlation was -0.43)

Step 2: TWOLEVEL Modeling of PA, NA, and Tired Model 3: Random Residual Covariance New in Version 8.9

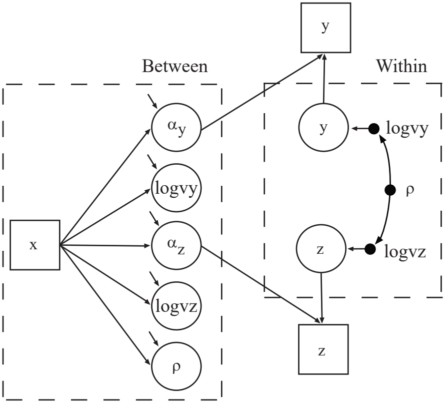
- How much do the residual covariances between different variables vary across individuals? What are their predictors?
  - Bivariate (or multivariate) outcome, random means, random residual variances and covariance:

$$y_{it} = \alpha_{yi} + \varepsilon_{yit}, \quad (4)$$

$$z_{it} = \alpha_{zi} + \varepsilon_{zit}, \quad (5)$$

where the residuals have individually-varying variances, $\varepsilon_{yit} \sim N(0, \sigma^2_{yi})$ and $\varepsilon_{zit} \sim N(0, \sigma^2_{zi})$, as well as individually-varying covariance $\rho_i \sqrt{\sigma^2_{yi}} \sqrt{\sigma^2_{zi}}$ where ρ_i is the correlation (Fisher z-transform of $\rho_i \sim N(\mu_r, \sigma^2_r)$; see Asparouhov & Muthén (2010). Bayesian analysis using Mplus: Technical implementation. `http://www.statmodel.com/download/Bayes3.pdf`)

Step 2 Model 3: Random Residual Variances and Residual Covariance for Two Outcomes with a Covariate



Input for Model 3 with Random Residual Covariances

```
            USEVARIABLES = pa na tired;
            CLUSTER = id;
            ! 24*7 = 168 (same as time >84):
            USEOBSERVATIONS = hrs le 168
            AND id ne 240 AND id ne 249
            AND id ne 78 AND id ne 531;
ANALYSIS:
            TYPE = TWOLEVEL RANDOM;
            ESTIMATOR = BAYES;
            BITERATIONS = (2000);
MODEL:
            %WITHIN%
            logvpa | pa;
            logvna | na;
            logvti | tired;
            cpn | pa with na;
            cpt | pa with tired;
            cnt | na with tired;
            %BETWEEN%
            pa na tired logvpa-cnt;
            [logvpa] (mpa);
            [logvna] (mna);
            [logvti] (mti);
            logvpa (spa);
            logvna (sna);
            logvti (sti);
            [cpn] (mc1);
            [cpt] (mc2);
            [cnt] (mc3);
MODEL
CONSTRAINT:
            NEW (meanvpa meanvna meanvti
            medrpn medrpt medrnt);
            meanvpa = exp(mpa+spa/2);
            meanvna = exp(mna+sna/2);
            meanvti = exp(mti+sti/2);
            ! Inverse of Fisher's z to get the median corr.:
            medrpn = (exp(2*mc1)-1)/(exp(2*mc1)+1);
            medrpt = (exp(2*mc2)-1)/(exp(2*mc2)+1);
            medrnt = (exp(2*mc3)-1)/(exp(2*mc3)+1);
OUTPUT:
            TECH1 TECH4 TECH8 STANDARDIZED;
PLOT:
            TYPE = PLOT3;
```

Output for Model 3 with Random Residual Covariances

	Estimate	Posterior S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Between Level					
Means					
PA	5.792	0.054	5.686	5.897	*
NA	1.336	0.029	1.281	1.393	*
TIRED	3.330	0.084	3.157	3.490	*
LOGVPA	-1.020	0.074	-1.164	-0.870	*
LOGVNA	-2.638	0.138	-2.908	-2.374	*
LOGVTI	0.164	0.086	-0.013	0.328	
CPN	-0.462	0.028	-0.516	-0.409	*
CPT	-0.250	0.019	-0.287	-0.212	*
CNT	0.094	0.020	0.054	0.132	*
Variances					
PA	0.669	0.066	0.553	0.819	*
NA	0.176	0.020	0.141	0.220	*
TIRED	1.567	0.158	1.296	1.903	*
LOGVPA	1.117	0.115	0.919	1.361	*
LOGVNA	4.279	0.413	3.560	5.181	*
LOGVTI	1.553	0.164	1.262	1.909	*
CPN	0.113	0.016	0.087	0.147	*
CPT	0.043	0.009	0.027	0.061	*
CNT	0.032	0.008	0.020	0.051	*
New/Additional Parameters					
MEANVPA	0.632	0.060	0.528	0.762	*
MEANVNA	0.607	0.170	0.389	1.029	*
MEANVTI	2.553	0.315	2.069	3.320	*
MEDRPN	-0.432	0.023	-0.475	-0.388	*
MEDRPT	-0.245	0.018	-0.279	-0.209	*
MEDRNT	0.094	0.020	0.054	0.132	*

Output for Model 3 with Covariates (Standardized)

	Estimate	Posterior S.D.	P-Value	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Between Level						
PA ON						
SDQEMOTAA	-0.331	0.046	0.000	-0.416	-0.239	*
GIRL	0.003	0.045	0.474	-0.083	0.088	
NA ON						
SDQEMOTAA	0.314	0.047	0.000	0.213	0.401	*
GIRL	-0.018	0.046	0.336	-0.108	0.072	
TIRED ON						
SDQEMOTAA	0.222	0.048	0.000	0.126	0.315	*
GIRL	0.104	0.046	0.014	0.012	0.193	*
LOGVPA ON						
SDQEMOTAA	0.194	0.050	0.000	0.093	0.286	*
GIRL	0.069	0.048	0.066	-0.023	0.163	
LOGVNA ON						
SDQEMOTAA	0.264	0.046	0.000	0.173	0.354	*
GIRL	0.000	0.046	0.498	-0.089	0.088	
LOGVTI ON						
SDQEMOTAA	0.005	0.052	0.460	-0.097	0.101	
GIRL	0.047	0.049	0.172	-0.051	0.144	
CPN ON						
SDQEMOTAA	-0.066	0.059	0.137	-0.186	0.048	
GIRL	-0.077	0.057	0.093	-0.190	0.029	
CPT ON						
SDQEMOTAA	-0.050	0.070	0.231	-0.190	0.089	
GIRL	-0.027	0.070	0.340	-0.162	0.111	
CNT ON						
SDQEMOTAA	0.213	0.081	0.004	0.052	0.368	*
GIRL	0.042	0.076	0.293	-0.110	0.187	

Why Are These Twolevel Models Not Sufficient for Intensive Longitudinal Data?

- What is lacking in the two-level analyses just presented? Model 1 has

$$y_{it} = \alpha_i + \varepsilon_{it}. \tag{6}$$

- This model accounts for correlation of observations across time using the random mean α_i (the latent trait)
- But there is further correlation to take into account: The residuals ε_{it} are likely correlated across time due to measurements close in time. For two time points t1 and t2, $Corr(\varepsilon_{it1}, \varepsilon_{it2}) = \rho^{|t1-t2|}$, that is, there is a non-zero correlation with size depending on the distance in time
- A common way to model this correlation is the lag-1 auto-correlation model $\varepsilon_{it} = \beta\varepsilon_{it-1} + \zeta_{it}$ (see, e.g., Chi & Reinsel, 1989, JASA)
  - Not only does this take into account residual correlation but it also offers substantively interesting modeling - DSEM

Relation to DSEM

- Consider the model with auto-correlated residuals,

$$y_{it} = \alpha_i + \varepsilon_{it}, \tag{7}$$
$$\varepsilon_{it} = \beta\,\varepsilon_{it-1} + \zeta_{it}. \tag{8}$$

Inserting (8) into (7) says that

$$y_{it} = \alpha_i + \beta\,\varepsilon_{it-1} + \zeta_{it}. \tag{9}$$

Equation (7) for t-1 implies that $\varepsilon_{it-1} = y_{it-1} - \alpha_i$. Inserting this into (9), shows that the model of (7) and (8) implies a within-level part of DSEM with latent variable centering of y_{it-1},

$$y_{it} = \alpha_i + \beta\,(y_{it-1} - \alpha_i) + \zeta_{it}, \tag{10}$$
$$y_{it} - \alpha_i = \beta\,(y_{it-1} - \alpha_i) + \zeta_{it}, \tag{11}$$
$$y^W_{it} = \beta\,y^W_{it-1} + \zeta_{it}, \tag{12}$$

where $y^W_{it} = \varepsilon_{it}$.

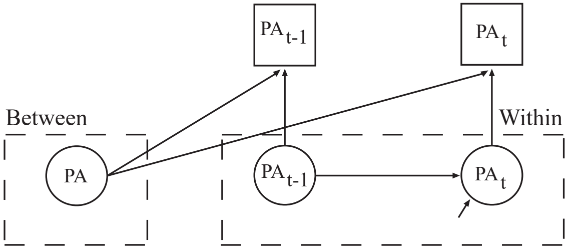
Section 5 Two-Level Analysis Bringing Time Into the Model: DSEM for a Single Outcome

Step 3: Two-Level Modeling Using DSEM
Univariate Analysis of PA

- 3 ways to describe the model:
  - Model diagram
  - Formulas
  - Mplus input
- Alternative ways to draw the model diagrams:
  - Balance between statistical accuracy and visual simplicity

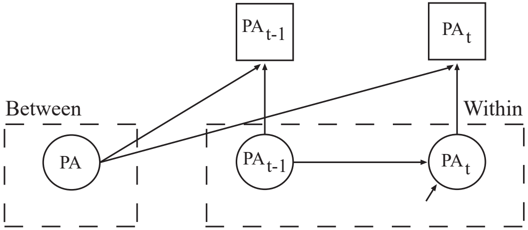
Two-Level Modeling Using DSEM: Model Diagram

- Within level: Variation over time (lag-1 shown in figure)
- Between level: Variation over individuals

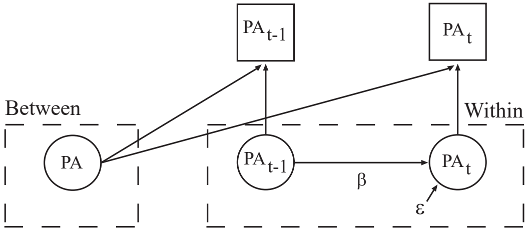


- The observed variable is decomposed into a Within and a Between part. The Between part is a random intercept varying over persons (also referred to as a latent trait)
- The model diagram is in line with the drawing of the random intercept in Mplus Web Talk 3 and RI-CLPM (Mulder & Hamaker, 2021)

Two-Level Modeling Using DSEM: Model Diagram Drawing Conventions



- Only two adjacent time points are shown
- The observed PA variables are shown outside the within and between boxes because they are neither within-only or between-only variables
- The regression coefficients for the observed PA variable on the latent between PA variable and latent within PA variables are all fixed at 1 and the residual is zero (the observed PA is the sum of the two latent PAs reflecting a within-between decomposition of the observed variable)
- The latent within-level PA variables have residuals
- The full within-part of the model is shown only for time point *t*



- The observed variable PA_t is expressed in terms of a random intercept α (labeled PA in the diagram) and the latent-variable centered PA_{it-1},

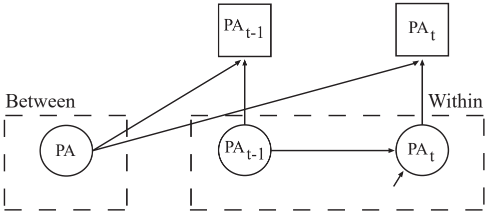
$$PA_{it} = \alpha_i + \beta \, (PA_{it-1} - \alpha_i) + \varepsilon_{it}, \qquad (13)$$

  where the latent variable centering avoids Nickell's bias (Asparouhov et al., 2018). Hamaker & Grasman (2015) also discuss centering

- Defining the latent within variable $PA_{it}^{(W)} = PA_{it} - \alpha_i$, (13) is the same as

$$PA_{it}^{(W)} = \beta \, PA_{it-1}^{(W)} + \varepsilon_{it}. \qquad (14)$$

Step 3 Model 1: Mplus Input



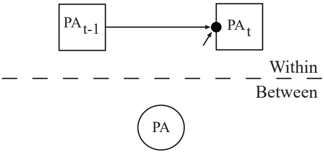
- Mplus MODEL command:

```
MODEL:    %WITHIN%
          pa ON pa&1;
          %BETWEEN%
          pa;
```

- The within-level slope (AR) and residual variance can be random (individual-specific), adding latent variables to the between level
- Mplus allows a full SEM to be specified on both levels

Two Alternative Ways to Draw DSEM Model Diagrams

- Mplus User's Guide with random effects marked as filled circles on within and open circles (latent variables) on between



- Hamaker papers/talks emphasizing decomposition into within- and between-level variation

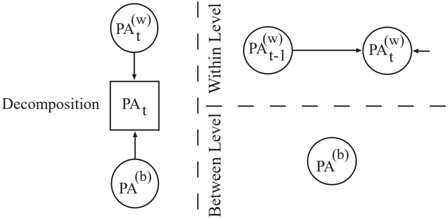
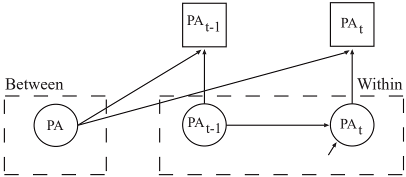


Diagram Style Chosen for this Web Talk



MODEL: %WITHIN%
pa ON pa&1;
%BETWEEN%
pa;

- Compared to the User's Guide style, there is no filled circle for the random intercept and it is made clear that the regression is for the within variables, implying latent variable centering
- Compared to the Hamaker style, there is no decomposition drawn, the role of the Between PA as a random intercept is emphasized, and there are no (w), (b) superscripts for the variables but instead the variable names match those of the input

What Defines a DSEM (Time Series) Run in Mplus as Opposed to a Regular Two-Level Analysis?

- The LAGGED option triggers DSEM (time series analysis) by allowing &
- The TINTERVAL option can be used only with LAGGED
- TINTERVAL is not necessary for LAGGED
- Not using the TINTERVAL option assumes that time is 1, 2, 3,...

Stationarity

- Stationarity of the time series implies that the mean and variance of the outcome are time invariant
  - This is not a modeling limitation but is necessary to match the type of data we are considering
- Without variance stationarity, data would have values that explode over time due to increasingly large variances, that is, data with exponential growth that we typically don't see in our applications
- Consider again the AR(1) model with the within-level part (dropping the W superscript as in the input)

$$PA_{it} = \beta PA_{it-1} + \varepsilon_{it}, \tag{15}$$

  with the residual variance $V(\varepsilon_{it}) = \theta$
- Stationarity requires that $|\beta| < 1$ (auto-regression slope less than 1)

Stationarity vs Non-Stationarity

- What is the variance of $PA_{it} = \beta PA_{it-1} + \varepsilon_{it}$? Stationarity implies

$$V(PA_{it}) = \beta^2 V(PA_{it}) + \theta, \quad (16)$$

$$V(PA_{it}) = \theta/(1-\beta^2); \ \ |\beta| < 1 \quad (17)$$

- The variance stationarity also implies that the unstandardized β estimate is the same as the standardized
- Non-stationarity example: Consider 10 consequtive timepoints 1, 2, 3, ..., 10 where V(PA_{i1}) = 1:

$$V(PA_{i2}) = V(\beta PA_{i1} + \varepsilon_{i2}) = \beta^2 + ..., \quad (18)$$

$$V(PA_{i3}) = V(\beta(\beta PA_{i1} + \varepsilon_{i2}) + \varepsilon_{i3}) = \beta^4 + ... \quad (19)$$

etc

  - The variance of PA_{i10} is dominated by the first term $\beta^{2\times 9}$
    - For example, $\beta = 1.5$ results in V(PA_{i10}) $= 1.5^{18} = 1,478!$
- Non-stationarity of the mean may occur with a trend over time
  - A function of time may be used as a covariate, so that RDSEM can be applied with stationarity for the residuals

Step 3 Model 1: Full Input Including Intraclass Correlation in DSEM

```
            USEVARIABLES = pa;
            ! 24*7 = 168 (same as time >84):
            USEOBSERVATIONS = hrs le
            168 AND id ne 240
            AND id ne 249 AND id ne 78
            AND id ne 531;
            CLUSTER = ID;
            ! The LAGGED option
            ! triggers time series analysis:
            LAGGED = pa(1);
            ! New in version 8.9:
            TINTERVAL = hrs (2 time);
ANALYSIS:
            TYPE = TWOLEVEL;
            ESTIMATOR = BAYES;
            BITERATIONS = (1000);
            PROCESSORS = 8;
```

```
MODEL:        %WITHIN%
              pa ON pa&1 (r);
              pa (resvar);
              %BETWEEN%
              pa (bvar);
MODEL
CONSTRAINT:   NEW(wvar icc);
              ! wvar: total within var
              ! which is time invariant
              ! so that
              ! wvar = r*r*wvar + resvar
              ! which means that:
              wvar = resvar/(1-r*r);
              icc = bvar/(bvar + wvar);
              ! icc = 0.56 (.025)
OUTPUT:       STANDARDIZED
              TECH1 TECH4 TECH8;
PLOT:         TYPE = PLOT3;
```

TINTERVAL Change in Version 8.9

- Version 8.9 has changed TINTERVAL to work the same way for TWOLEVEL and CROSSCLASSIFIED DSEM
- Initial missing values for an individual are no longer dropped for TWOLEVEL
  - The change causes minimal differences in estimates
  - The old approach can be obtained by adding DROP as a third argument in parenthesis for TINTERVAL
- The orginal time variable (hrs in the previous example) is no longer changed by TINTERVAL but a new time variable that the user names is used in the analysis (time in the example) and can be modified in DEFINE
  - If no new time variable is named by the user, _BINT is used

Step 3 Model 1 Output Excerpts

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%	Significance
Within Level					
PA ON					
PA&1	0.414	0.014	0.387	0.441	*
Residual Variances					
PA	0.483	0.009	0.466	0.502	*
Between Level					
Means					
PA	5.756	0.058	5.639	5.871	*
Variances					
PA	0.735	0.074	0.611	0.908	*
New/Additional Parameters					
WVAR	0.583	0.011	0.561	0.605	*
ICC	0.557	0.025	0.508	0.610	*

- The S.D. column corresponds to SE for ML. Estimate/SD is approximately a z-score but the CI is a better statistic to report

Step 3 Model 1 Output Continued

STANDARDIZED MODEL RESULTS
STDYX Standardization

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Within Level					
PA ON					
PA&1	0.414	0.014	0.387	0.441	
Residual Variances					
PA	0.828	0.012	0.805	0.850	*
Between Level					
Means					
PA	6.710	0.337	6.024	7.372	*
Variances					
PA	1.000	0.000	1.000	1.000	

- Within-level R-square is 1 - 0.828 = 0.172

Summary of Results for PA Analysis using TINTERVAL = 1, 2, 3 (No Logv)

	Tinterval = 1	Tinterval = 2	Tinterval = 3
Time interval plot information (metric is hours)			
Avg. SMSE	0.500	0.832	1.872
Max. SMSE	0.500	1.732	4.517
Missing data coverage	0.171	0.341	0.507
Within Est's			
AR (1)	0.547 (.014)	0.414 (.014)	0.378 (.014)
Resvar	0.410 (.010)	0.483 (.009)	0.501 (.010)
Between Est's			
Mean	5.754 (.058)	5.756 (.058)	5.756 (.057)
Var	0.734 (.073)	0.735 (.074)	0.725 (.072)

- Tinterval = 1 matches the data best but the low coverage (many missing data rows) makes convergence difficult for complex models
- AR(1) values decline with increasing Tinterval but 0.414 is not 0.547*0.547 as expected, suggesting that lag-1 is not sufficient

Taking a Step Back: N = 1 DSEM (Classic Time Series Analysis)

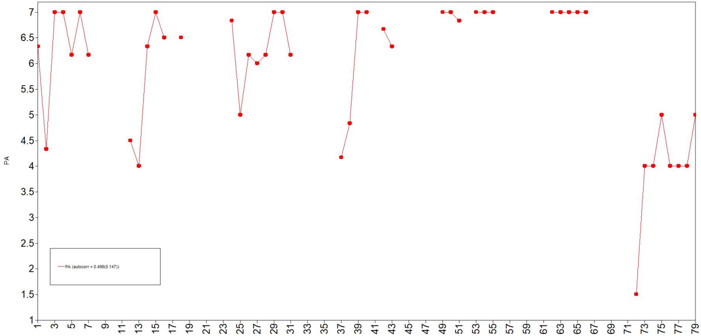
- This model has 3 parameters
- Common multivariate model: Vector auto-regressive (VAR) - like CLPM
- Mplus allows a full SEM to be specified

```
             USEVARIABLES = pa;
             USEOBSERVATIONS = id eq 41;
             LAGGED = pa(1);
             TINTERVAL = hrs (2 time);
ANALYSIS:    ESTIMATOR = BAYES;
             BITERATIONS = (1000);
             PROCESSORS = 8;
MODEL:       pa ON pa&1;
OUTPUT:      STANDARDIZED TECH1
             TECH4 TECH8;
PLOT:        TYPE = PLOT3;
```

- A between model is not specified

N = 1 DSEM for PA using ID = 41 (7 days: Tue - Mon)

- Time Series plot:



- X-axis: 79 time points (12 occasions per day for 7 days = max 84)
  - 46 of them non-missing: Sample size = 46

Output for N = 1 DSEM for PA using ID = 41 (N = 46)

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
PA ON					
PA&1	0.636	0.120	0.373	0.848	*
Intercepts					
PA	2.175	0.725	0.842	3.735	*
Residual Variances					
PA	1.071	0.274	0.705	1.757	*

- The N = 1 model has $PA_t = \nu + \beta PA_{t-1} + \varepsilon_t$
- Stationarity implies $E(PA_t) = E(PA_{t-1}) = \mu$ so that

$$\mu = \nu + \beta\mu, \tag{20}$$
$$\mu = \nu/(1-\beta). \tag{21}$$

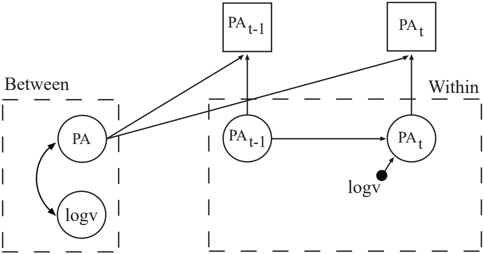
- Here, $\mu = 5.97$ as shown in TECH4 and RESIDUAL output

Joint N = 1 Analysis of All Individuals

- N = 1 analysis of all individuals together can be carried out and estimates averaged using the TYPE = MONTECARLO option of the DATA command in line with User's Guide ex 12.6
- Mplus DSEM input shown at `https://ellenhamaker.github.io/DSEM-book-chapter/`
- See section Empirical Illustration: Part 1 on page 580 of Hamaker, Asparouhov & Muthén (2023). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling. Chapter 31 in Handbook of Structural Equation Modeling (2nd edition); Rick H. Hoyle (Ed.); Publisher: Guilford Press.
`http://www.statmodel.com/download/HamakerAsparouhovMuthen21.pdf`

Step 3 Model 2: Two-Level Modeling Using DSEM with Random Residual Variance

- The Step 1 TYPE = TWOLEVEL BASIC between-level scatterplot of PA variance and mean showed strong heteroscedastic variance which was correlated with the mean



- The residual variance is interpreted as variation in innovation or random shocks (different reactions to events)
- Random AR(1) can also be added (interpreted as inertia)

Step 3 Model 2 Twolevel DSEM Input

```
           USEVARIABLES = pa;
           ! 24*7 = 168 (same as time >84):
           USEOBSERVATIONS = hrs LE 168 AND
           id NE 240 AND id NE 249 AND id NE 78 AND id NE 531;
           CLUSTER = id;
           LAGGED = pa(1);
           TINTERVAL = hrs (2 time); ! New in version 8.9
ANALYSIS:
           TYPE = TWOLEVEL RANDOM
           ESTIMATOR = BAYES;
           BITERATIONS = (1000);
           PROCESSORS = 8;
MODEL:
           %WITHIN%
           pa ON pa&1;
           logv | pa;
           %BETWEEN%
           pa WITH logv;
OUTPUT:    STANDARDIZED TECH1 TECH8;
PLOT:      TYPE = PLOT3;
           FACTORS = ALL(50);
SAVEDATA:  FILE = patint2.dat;
```

Number of Time Points for Different Individuals (Clusters)

SUMMARY OF DATA

Number of clusters 240

Size (s)	Cluster ID with Size s
6	**45**
9	238
10	197 196 319
17	572
18	161
19	456
21	415
22	357
31	276
33	176 107
34	29 47
35	352
40	254 457 549 204
41	469 224 205
42	165 36 200
43	433 406 371
44	314 13 320 347 7
46	221 51 383 385 277 175
53	351
55	122 313
56	2 160 554 49
58	503 556 71
64	560 130
65	445 428
66	92 442
67	28
68	317 463 361 155 537

Size (s)	Cluster ID with Size s
69	129
70	452 454 491 278 **20**
75	310
76	1 234 40 552 102 260 12
	571 119
77	309 263 507 394 542 32
	338 280 373 438 473 485
78	558 118 83 66
79	211 528 43 94 152 111
	322 480 557 380 159 499 206
80	411 307 419 101 85 258 188
	89 547 374 220 24 6 388 465
	203 561 401 296
81	341 31 455 137 207 209 279
	466 144 213 297 300 302 305
	178 392 514 523 181 148 227
	543 56 4 246 69 429 22 327
	23 443 570 340 448
82	16 393 75 398 286 403 287
	407 288 414 126 417 79 298
	81 431 30 50 439 440 87 **41** 52
	446 447 27 95 100 225 162 228
	460 462 232 328 331 468 163
	470 235 475 478 64 343 346 243
	501 171 252 508 510 65 518 520
	521 358 525 359 360 541 256 365
	546 370 106 372 177 555 5 266
	382 109 46 386 70 389 573
83	295 348 186

Number of Time Points Continued

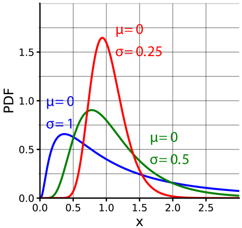
- Note that the number of time points per individual in a run using the TINTERVAL option includes time points with missing data that was inserted by the TINTERVAL option to synchronize the timing of the measurements
  - Consider for example the first individual on the list, ID = 45 which has 6 time points - 4 of those timepoints have missing data
- To see the number of time points without inserted missing data, do a run without TINTERVAL
  - See the data summary for the Step 1 TYPE = TWOLEVEL BASIC run - this shows that ID = 45 is observed at 2 time points
- It is important to have a sufficient number of time points without missing data when estimating random effects such as regression coefficients because the estimation is based on computing the random effect value for each individual for which "N" is the number of observed timepoints (Asparouhov & Muthén, 2022, Practical Aspects of Dynamic Structural Equation Models. Technical Report)

Model Results Without and With Logv

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%	Sig.
Within Level						
PA ON PA&1	**0.414**	0.014	0.000	0.387	0.441	*
Residual Var	0.483	0.009	0.000	0.466	0.502	*
Between Level						
Mean	5.756	0.058	0.000	5.639	5.871	*
Variance	0.735	0.074	0.000	0.611	0.908	*
Within Level						
PA ON PA&1	**0.426**	0.015	0.000	0.396	0.453	*
Residual Var	NA					
Between Level						
PA WITH LOGV	**-0.472**	0.073	0.000	-0.628	-0.344	*
Mean PA	5.753	0.056	0.000	5.644	5.867	*
Mean LOGV	-1.161	0.072	0.000	-1.314	-1.014	*
Var. PA	0.735	0.074	0.000	0.611	0.908	*
Var. LOGV	1.101	0.120	0.000	0.896	1.375	*

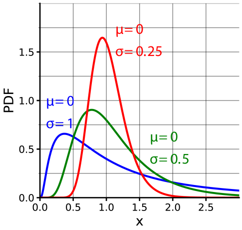
- Jongerling, Laurenceau, Hamaker (2015) in MBR: Bias when ignoring random logv

- V(ε) = v, where v has a lognormal distribution ($>$ 0 and skewed):



- Because v has a lognormal distribution, logv $\sim N(\mu, \sigma^2)$
  - This fits with the assumption of normally distributed random effects in Mplus
- The lognormal distribution implies that v has
  - Median = e^{μ}
  - Mean = $e^{\mu+\sigma^2/2}$
  - Variance = $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$
- Equivalent expression: V(ε) = e^s (variance $>$ 0) so that log V(ε) = s has a normal distribution

Understanding the Random Residual Variance, Continued



- For a skewed distribution like the lognormal for the random residual variance v, quartiles are better summaries than means and variances
- The quartiles can be obtained from the normal logv distribution
- 50% of the normal distribution lies between the 25th and 75th quartiles (0.675 SD from the mean)
- Because the lognormal is a monotonic function of the normal, the quartiles for the residual variance are obtained by exponentiating the normal distribution quartiles

Understanding the Random Residual Variance Continued

- The quartile estimates and their CIs are obtained by adding MODEL CONSTRAINT with parameter labels from MODEL:

```
MODEL:      %WITHIN%
            pa ON pa&1;
            logv | pa;
            %BETWEEN%
            pa WITH logv;
            [logv] (m);
            logv (s);
```

```
MODEL
CONSTRAINT: NEW(medianv 25qv 75qv);
            medianv = EXP(m);
            ! 25th and 75th quartiles
            ! based on normal dist of logv:
            25qv = exp(m - sqrt(s)*0.675);
            75qv = exp(m + sqrt(s)*0.675);
```

- medianv = 0.313 (0.023), CI = [0.269 0.363]
  25qv = 0.154 (0.013), CI = [0.129 0.178]
  75qv = 0.635 (0.052), CI = [0.543 0.751
- The non-random residual variance in Model 1 is 0.483
- The square roots of the quartiles correspond to SDs which can be related to the scale of the variable

TECH4 Output

- With TYPE = RANDOM, the within-level covariance matrix changes over individuals. TECH4 gives the average over individuals for within (TECH4(CLUSTER) gives cluster-specific results)

AVERAGE ESTIMATES DERIVED
FROM THE MODEL FOR WITHIN
ESTIMATED MEANS
FOR THE LATENT VARIABLES

LOGV	PA	PA&1
-1.093	5.741	5.741

ESTIMATED COVARIANCE MATRIX
FOR THE LATENT VARIABLES

	LOGV	PA	PA&1
LOGV	0.000		
PA	0.000	0.642	
PA&1	0.000	0.272	0.642

- The between-level part of TECH4 is as usual

SAVEDATA Information (New in Version 8.9)

Save file
  patint2.dat

Order and format of variables

PA	F10.3
PA&1	F10.3
HRS	F10.3
_NEWTIME	F10.3
TIME	F10.3
ID	I4

- HRS: Original hours variable
- _NEWTIME: Midpoints of time ranges implied by TINTERVAL (only used in the Time interval plot)
- TIME: Name given by user to the time variable created by TINTERVAL = hrs(2 time) and used in the analysis (same as bin number). If name not given by user, the default name is _BINT

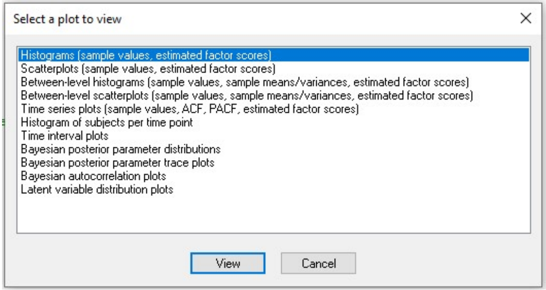
Saved Data for First Day of ID = 20 (Tinterval = 2)

PA	PA&1	HRS	_NEWTIME	TIME	ID
*	*	*	*	1.000	20
*	*	*	*	2.000	20
6.167	*	**5.000**	5.000	3.000	20
5.833	6.167	7.000	7.000	4.000	20
5.667	5.833	**10.000**	**9.000**	5.000	20
5.833	5.667	11.000	11.000	6.000	20
6.000	5.833	13.000	13.000	7.000	20
*	6.000	*	*	8.000	20
5.833	*	16.000	17.000	9.000	20
5.667	5.833	18.000	19.000	10.000	20
*	5.667	*	*	11.000	20
*	*	*	*	12.000	20

- Compare with slide 14 showing what Tinterval does
- First response at 5am (HRS column)
- Misalignment at 10am (10 and 11 in the same 2-hour bin so 10 is moved to earlier)
- Handling night time is a research topic (night AR $\neq$ day AR)

Overview of Plot Menu Choices

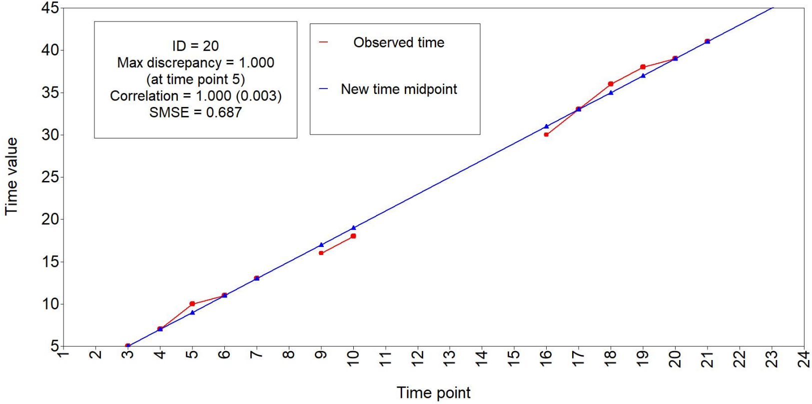
- Plot - View plots:



- Histograms
- Scatterplots
- Between level scatterplots
- Time series plots
- Histogram of subjects per time point
- Time interval plots
- Bayesian posterior parameter distributions
- Bayesian posterior parameter trace plots
- Bayesian autocorrelation plots

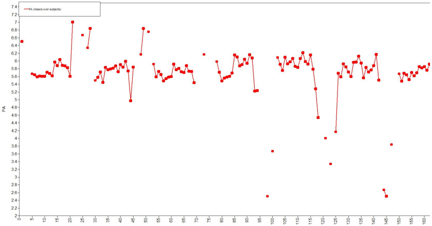
Time Interval Plot for ID = 20 (see Slides 14, 72)

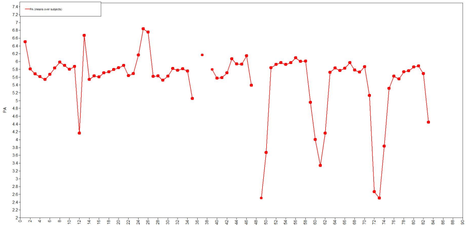
- 12 occasions per day for 2 days = 24 time points on the x-axis
- The time points on the x-axis represent the values of the TIME variable



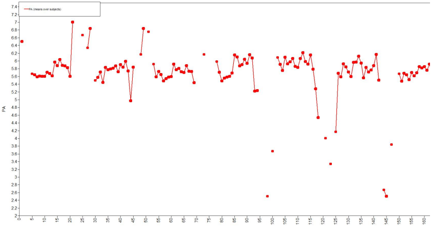
- SMSE = square root of mean square error over all time points (average distance between observed time and new time midpoint)

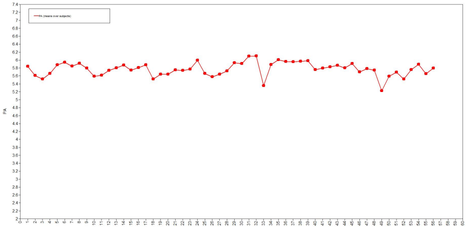
- Tinterval = 1 vs 2



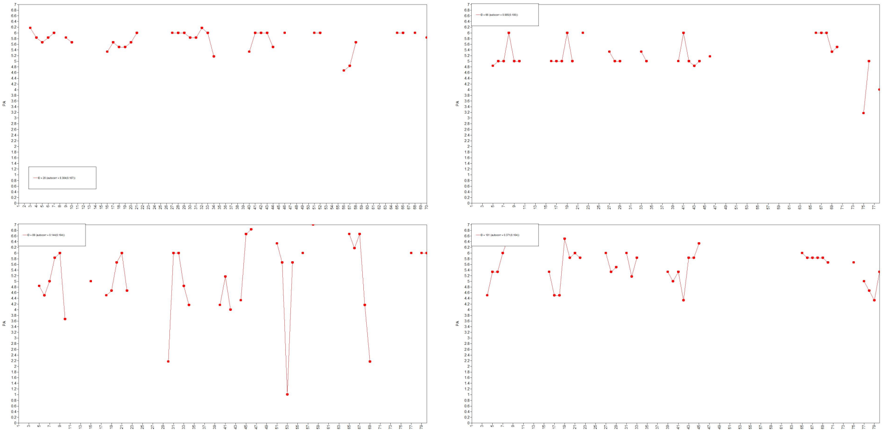


- Tinterval = 1 vs 3





Time Series Plot of PA for 4 Persons (Tinterval=2)



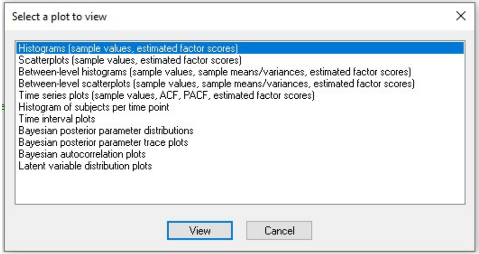
- Hard to see a pattern from any one individual - DSEM combines information across individuals to get the pattern for the population

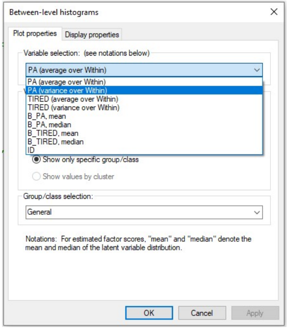
Three Kinds of Between-Level Histograms of Means

- (1) The observed mean for each person
    - Average over time for each person
- (2) The estimated between-level mean for each person
    - Requires FACTORS = ALL (50) in the PLOT command: The histogram is based on the Bayes estimates of each person's random effect value (factor score) averaged over the 50 draws from the posterior distribution of all parameter estimates (which includes random effects)
    - Referred to as B_PA, mean in the Between-level Histograms choice of the Plot menu
- (3) Cluster mean
    - For each person, this is the average of the random mean over all iterations during the estimation. Used in the OUTPUT option RESIDUAL(CLUSTER)
    - Referred to as PA (estimated cluster mean) in the Between-level Histograms choice of the Plot menu

Three Kinds of Between-Level Histograms, Cont'd

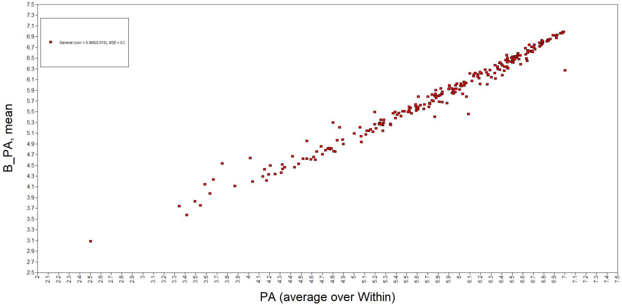
- Summary of Between-level histogram choices (3rd line):
  - (1) PA (average over Within)
  - (2) B_PA, mean
  - (3) PA (estimated cluster mean)



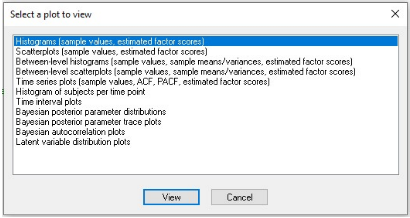


Between-Level Scatterplots

- y-axis: Estimated cluster mean factor scores (2)
- x-axis: Observed average over within (1)

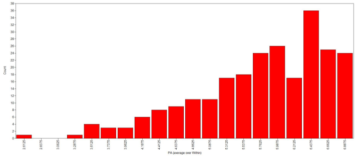


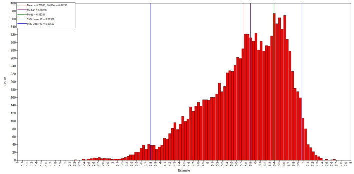
- Some clusters show a discrepancy - most likely due to missing data handled better by the estimated cluster mean (y-axis), drawing on correlated variables in the model
- For more on factor scores, see multiple imputation on slide 95



- Latent variable distribution plots (last line)
  - Requires FACTORS = ALL (50) in the PLOT command: The estimated betwen-level mean for each person using several imputations (draws) from the posterior distribution of all parameter estimates (which includes random effects)
    - Each person's random effect (or factor) distribution is created by 50 draws after the last Bayes iteration, and the whole distribution is based on the number of people (clusters) times 50. This gives a smoother representation of the histogram (2) of estimated between-level means
    - Imputations - also called plausible values - allow uncertainty in the estimates to be accounted for (not just a point estimate for each person)

Histogram of Estimated Between-Level Mean vs Latent Variable Distribution





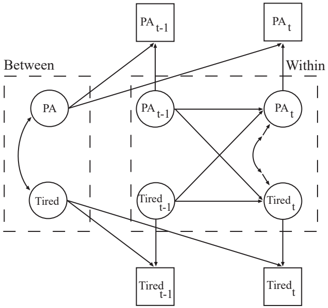
- Similar shape (normality is assumed in the model but the posterior updates this prior using the data)

Mplus Runs

- Mplus outputs and plots
  - Step 1: TYPE = TWOLEVEL BASIC
  - Step 2: TYPE = TWOLEVEL RANDOM (regular twolevel)
  - Step 3, Model 2: Twolevel DSEM with random residual variance
- Plotting for Mac and Linux users: The GH5 file that the PLOT command produces can be used for plots in R: `http://www.statmodel.com/mplus-R/`
  - New feature: Simple user interface added for R plots with the use of R Shiny

Section 6 Two-level DSEM and RDSEM Analysis Two Outcomes

Step 4 Model 1: Bivariate Modeling in DSEM Cross-Lagged Analysis of PA and Tired



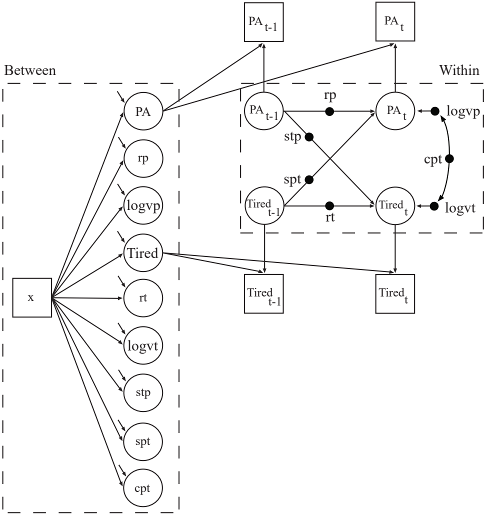

```
MODEL:          %WITHIN%
                pa ON pa&1 tired&1;
                tired ON tired&1 pa&1;
                %BETWEEN%
                PA WITH tired;
```

Step 4 Model 1 Stand'd Within Estimates for Cross-Lagged DSEM Analysis of PA and Tired (TINTERVAL = 2)

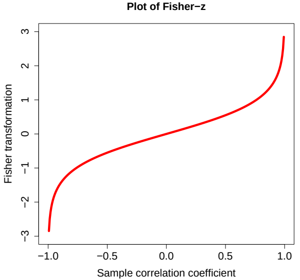
		Posterior	95% C.I.		
	Estimate	S.D.	Lower 2.5%	Upper 2.5%	Significance
Within Level					
PA ON					
PA&1	0.398	0.015	0.369	0.426	*
Tired&1	**-0.084**	0.016	-0.114	-0.052	*
Tired ON					
Tired&1	0.460	0.015	0.431	0.489	*
PA&1	**-0.007**	0.015	-0.035	0.022	
Tired with PA	-0.189	0.013	-0.215	-0.163	*
Residual var's					
PA	0.819	0.013	0.795	0.843	*
Tired	0.786	0.013	0.760	0.811	*

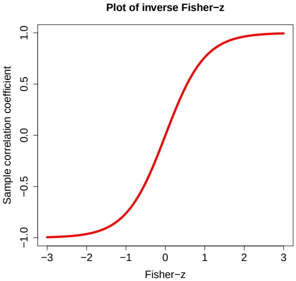
- Small effect of $\text{Tired}_{t-1} \rightarrow \text{PA}_t$. Larger lag0 effect?

Step 4 Model 1 with All Random Effects and Covariates



- The model allows individually-varying covariance $\rho_i \sqrt{\sigma^2_{PA_i}} \sqrt{\sigma^2_{Tired_i}}$ where ρ_i is the correlation
- But ρ_i is not normally distributed so we transform it
- Fisher z-transform: $z = \frac{1}{2} ln[(1+\rho_i)/(1-\rho_i)], z \sim N(\mu, \sigma^2)$
- The reverse formula is $\rho = (e^{2z}-1)/(e^{2z}+1)$
- The median of the original ρ is obtained as $(e^{2\mu}-1)/(e^{2\mu}+1)$
- z and ρ are almost identical for ρ values between -0.5 and +0.5





- Asparouhov & Muthén (2010). Bayesian analysis using Mplus: Technical implementation. `http://www.statmodel.com/download/Bayes3.pdf`

Step 4 Model 1 with All Random Effects: Input

```
            USEVARIABLES = pa tired
            SDQemotAA girl;
            LAGGED = pa(1) tired(1);
            TINTERVAL = hrs(2 time);
            CLUSTER = id;
            BETWEEN = SDQemotAA girl;
            ! 24*7 = 168 (same as time >84):
            USEOBSERVATIONS = hrs le 168
            AND id ne 240 AND id ne 249
            AND id ne 78 AND id ne 531;
DEFINE:     girl = sexAA - 1;
            CENTER
            SDQemotAA(GRANDMEAN);
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
            ESTIMATOR = BAYES;
            BITERATIONS = (2000);
            THIN = 10;
            PROCESSORS = 8;
MODEL:      %WITHIN%
            rp | pa ON pa&1;
            logvp | pa;
            rt | tired ON tired&1;
            logvt | tired;
            spt | pa ON tired&1;
            stp | tired ON pa&1;
            cpt | pa WITH tired;
            %BETWEEN%
            pa tired rp-cpt ON SDQemotAA girl;
            pa tired rp-cpt WITH pa tired rp-cpt;
            spt (vspt);
            stp (vstp);
            [logvp] (mpa);
            [logvt] (mti);
            logvp (spa);
            logvt (sti);
            [cpt] (mc);
MODEL
CONSTRAINT: NEW(sdspt sdstp);
            sdspt = sqrt(vspt);
            sdstp = sqrt(vstp);
            NEW (meanvpa meanvti medrpt);
            meanvpa = exp(mpa+spa/2);
            meanvti = exp(mti+sti/2);
            medrpt = (exp(2*mc)-1)/(exp(2*mc)+1);
```

Step 4 Model 1 with All Random Effects: Analysis Strategies

- It is important to have a sufficient number of time points without missing data when estimating random effects such as regression coefficients because the estimation is based on computing the random effect value for each individual for which "N" is the number of observed timepoints - many individuals with few timepoints may complicate the convergence
- Asparouhov & Muthén (2022). Practical Aspects of Dynamic Structural Equation Models. Technical Report
  - Convergence made harder by including individuals that do not vary across time
  - Number of random effects should ideally be smaller than the number of time points
  - Problems can be avoided by building up the model in steps: Random intercepts only, adding random effects looking for z-values > 3 for their variances, don't correlate the random effects right away

Number of Time Points for Each Individual (Cluster)

- Is the number of time points sufficient for 9 random effects?
- Tinterval = 2 with inserted missing data versus the original data

Size (s)	Cluster ID with Size s
6	45
9	238
10	**197 196 319**
17	**572**
18	**161**
19	**456**
21	**415**
22	**357**
31	276
33	176 107
34	29 47
35	352
40	254 457 549 204
41	469 224 205
42	165 36 200
43	433 406 371
44	314 13 320 347 7

Size (s)	Cluster ID with Size s
1	238 549
2	**161** 29 240 383 442 45 **572**
3	206 249
4	**319** 176 **197** 78 276
5	**456** 531 **415** 571 **196**
6	263 24 200 380 51 388 392
7	359 100 160 224
9	92
10	71 2 13 320
11	12 7 204 341 352 **357**
12	205 22 561 28 485
13	314 260 119 433 254 348
14	503 107 47 552 277 137 188
15	5 361
16	36 401 507 518 403 373 428 280 221 347
17	407
18	165 351 521 528 4 447 452 554 560 302 414 155
19	1 49 186 406 295 203 310 385 152
20	514 122 31 207 454 547 144 457 463 258 491 360 56

Standardization with TYPE = RANDOM

- Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker (2016). How to compare cross-lagged associations in a multilevel autoregressive model. Psychological Methods, 21, 206-221
- Standardization using individual-specific variances for models with random slopes or variances
- Analogous to N = 1 analysis
- Mplus computes the standardized values for each individual and presents the average random slope over individuals

Step 4 Model 1 with All Random Effects: Non-Stationarity Warning

WARNING: PROBLEMS OCCURRED IN SEVERAL ITERATIONS IN THE COMPUTATION OF THE STANDARDIZED ESTIMATES FOR SEVERAL CLUSTERS. THIS IS MOST LIKELY DUE TO AR COEFFICIENTS GREATER THAN 1 OR PARAMETERS GIVING NON-STATIONARY MODELS. SUCH POSTERIOR DRAWS ARE REMOVED. THE FOLLOWING CLUSTERS HAD SUCH PROBLEMS:

45 238 197 196 319 572 161 456 415 357 276 176 107 29 254 549 204 224
205 165 200 433 406 371 314 13 320 347 7 221 51 383 385 277 175 351 122
2 160 554 49 503 556 71 560 445 428 92 442 317 463 361 454 20 234 552
260 309 263 394 542 32 338 373 438 485 66 211 152 480 557 380 499 419
101 85 258 188 374 220 388 465 203 561 401 341 31 207 279 144 302 392
514 148 327 443 570 448 393 398 286 403 407 414 30 439 440 41 27 100
225 228 462 328 468 163 243 508 65 518 521

- Far fewer clusters are mentioned when the auto-correlations are not random

Non-Stationarity Warning Continued

- FAQ: Standardized coefficients in DSEM/RDSEM `http://www.statmodel.com/download/FAQ-DSEMStand.pdf`
- Message can typically be ignored (Mplus handles it optimally)
- Typical reasons:
  - Major cause is random auto-correlations where in some iterations for some individuals, some part of the posterior distribution is > 1 so that variances are negative under stationarity assumption which means that standardization cannot be done
  - Small sample size or small number of timepoints for many individuals
- Further information about causes can be obtained by
  - N = 1 analyses
  - Examining factor scores to find clusters with large auto-correlations (see later slide)

- Unstandardized between-level estimates

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Intercepts					
PA	5.755	0.088	5.582	5.927	*
TIRED	3.139	0.144	2.859	3.433	*
RP	0.346	0.044	0.257	0.427	*
SPT	**-0.063**	0.019	-0.103	-0.028	*
RT	0.352	0.041	0.268	0.428	*
STP	-0.027	0.060	-0.140	0.094	
LOGVPA	-1.426	0.131	-1.685	-1.167	*
LOGVTI	-0.153	0.149	-0.449	0.141	
CPT	-0.193	0.035	-0.262	-0.122	*
Residual Variances					
PA	0.617	0.068	0.501	0.766	*
TIRED	1.483	0.153	1.228	1.820	*
RP	0.064	0.011	0.046	0.088	*
SPT	0.005	0.002	0.002	0.009	*
RT	0.048	0.009	0.033	0.070	*
STP	0.029	0.013	0.011	0.061	*
LOGVPA	1.176	0.132	0.948	1.463	*
LOGVTI	1.661	0.192	1.325	2.057	*
CPT	0.033	0.009	0.018	0.053	*
New/Additional Parameters					
SDSPT	**0.069**	0.012	0.045	0.094	*
SDSTP	0.171	0.037	0.104	0.247	*
MEANVPA	0.433	0.067	0.327	0.594	*
MEANVTI	1.972	0.362	1.419	2.845	*
MEDRPT	-0.191	0.034	-0.256	-0.121	*

- The unstandardized, non-random estimate for Model 1 without covariates = -0.048

Within-Level STD Estimates Averaged Over Clusters

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
RP | PA ON PA&1	0.352	0.019	0.315	0.389	*
SPT | PA ON TIRED&1	**-0.056**	0.012	-0.081	-0.035	*
RT | TIRED ON TIRED&1	0.417	0.019	0.380	0.453	*
STP | TIRED ON PA&1	-0.032	0.033	-0.094	0.037	
CPT | PA WITH TIRED	-0.166	0.015	-0.199	-0.138	*
LOGVPA | PA	0.458	0.019	0.425	0.499	*
LOGVTI | TIRED	1.437	0.058	1.340	1.566	*

- STD standardization has no effect here due to no latent variables, making the results comparable to the unstandardized between-level estimates
- Note that LOGV entries are on the variance scale (so pos. values; no exp needed)
- SE's of these cluster averages decrease as a function of cluster size
- For significance testing of a random slope mean, the unstandardized mean should be used, not the cluster-averaged standardized slope

Step 4 Model 1 Regression on Covariates

- Unstandardized between-level estimates

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
RP ON					
SDQEMOTAA	0.001	0.012	-0.022	0.023	
GIRL	0.011	0.054	-0.094	0.115	
SPT ON					
SDQEMOTAA	-0.004	0.005	-0.015	0.005	
GIRL	0.012	0.020	-0.029	0.051	
RT ON					
SDQEMOTAA	0.007	0.010	-0.013	0.029	
GIRL	0.106	0.049	0.009	0.203	*
STP ON					
SDQEMOTAA	0.005	0.013	-0.019	0.030	
GIRL	-0.006	0.065	-0.142	0.118	
LOGVPA ON					
SDQEMOTAA	0.125	0.038	0.048	0.196	*
GIRL	0.230	0.165	-0.107	0.566	
LOGVTI ON					
SDQEMOTAA	-0.011	0.044	-0.099	0.077	
GIRL	0.099	0.190	-0.290	0.459	
CPT ON					
SDQEMOTAA	-0.007	0.009	-0.026	0.011	
GIRL	-0.007	0.042	-0.089	0.076	
PA ON					
SDQEMOTAA	-0.178	0.027	-0.233	-0.126	*
GIRL	0.004	0.115	-0.225	0.225	
TIRED ON					
SDQEMOTAA	0.187	0.043	0.105	0.272	*
GIRL	0.346	0.184	-0.021	0.705	

3 Ways to Examine the Individuals' Values of the Random Effects (Factor Score Values)

- (1) To view the factor score distribution:
  - PLOT command using FACTORS = ALL (50), or FACTORS = list, and using the Latent variable distribution plot
- (2) To get factor score mean, median, variance and percentile summaries saved in a file together with the rest of the data for a follow-up analysis:
  - SAVEDATA command using SAVE = FSCORES (50 10), where 10 refers to thinning
- (3) To get multiple imputations of factors scores and save all the between-level information per imputation to be subsequently analyzed in a single-level model using TYPE = IMPUTATION in the DATA command:
  - SAVEDATA command using SAVE = FSCORES (200), FACTORS = list, and FILE = name imp*.dat
  - Asparouhov & Muthén (2010). Plausible values for latent variables using Mplus. Technical Report `http://www.statmodel.com/download/Plausible.pdf`

Example of (2): Saving Factor Score Summaries

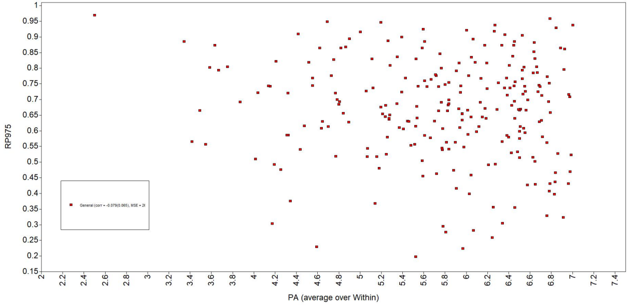
- SAVEDATA: SAVE = FSCORES(50 10); FILE = fscoresM1.dat;

PA	F10.3
TIRED	F10.3
PA&1	F10.3
TIRED&1	F10.3
HRS	F10.3
_NEWTIME	F10.3
TIME	F10.3
RP Mean	F10.3
RP Median	F10.3
RP Standard Deviation	F10.3
RP 2.5% Value	F10.3
RP 97.5% Value	F10.3
SPT Mean	F10.3
SPT Median	F10.3
SPT Standard Deviation	F10.3
SPT 2.5% Value	F10.3
SPT 97.5% Value	F10.3
RT Mean	F10.3
RT Median	F10.3
RT Standard Deviation	F10.3
RT 2.5% Value	F10.3
RT 97.5% Value	F10.3
STP Mean	F10.3
STP Median	F10.3

STP Standard Deviation	F10.3
STP 2.5% Value	F10.3
STP 97.5% Value	F10.3
LOGVPA Mean	F10.3
LOGVPA Median	F10.3
LOGVPA Standard Deviation	F10.3
LOGVPA 2.5% Value	F10.3
LOGVPA 97.5% Value	F10.3
LOGVTI Mean	F10.3
LOGVTI Median	F10.3
LOGVTI Standard Deviation	F10.3
LOGVTI 2.5% Value	F10.3
LOGVTI 97.5% Value	F10.3
B_PA Mean	F10.3
B_PA Median	F10.3
B_PA Standard Deviation	F10.3
B_PA 2.5% Value	F10.3
B_PA 97.5% Value	F10.3
B_TIRED Mean	F10.3
B_TIRED Median	F10.3
B_TIRED Standard Deviation	F10.3
B_TIRED 2.5% Value	F10.3
B_TIRED 97.5% Value	F10.3
ID	I4

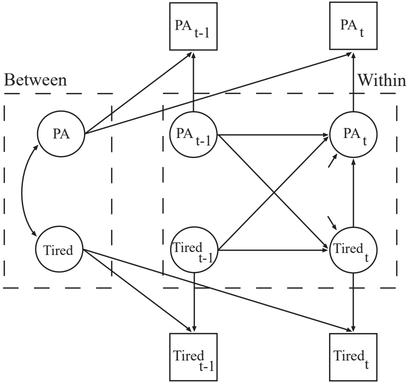
Follow-Up Two-Level Basic Analysis of the Saved Factor Scored Summaries

- Between-level scatterplot of PA Auto-Correlation's 97.5 Percentile vs PA Average



- Pointing at top left individual (cluster) with the highest percentile value (y-axis) of 0.97 shows that it is ID = 414
- Unusual individual with a very low PA average - outlier to be deleted?

Step 4 Model 2: Adding a Lag0 Effect of $\text{Tired}_t \rightarrow \text{PA}_t$ to the Cross-Lagged Analysis of PA and Tired



MODEL: %WITHIN%

pa ON pa&1 **tired** tired&1;

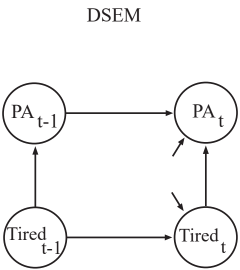
tired ON tired&1 pa&1;

Model 2 Standard'd Estimates for Cross-Lagged Analysis of PA and Tired Without and With Lag0 Effect of $Tired_t \rightarrow PA_t$

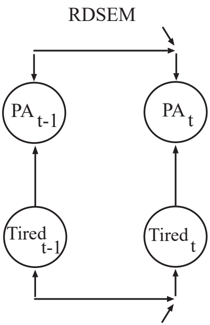
	Estimate	Posterior S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
		Only Lag1 effect of Tired			
Within Level					
PA ON					
PA&1	0.398	0.015	0.369	0.426	*
Tired&1	**-0.084**	0.016	-0.114	-0.052	*
		Lag0 and Lag1 effect of Tired			
Within Level					
PA ON					
PA&1	0.396	0.015	0.366	0.423	*
Tired	**-0.193**	0.014	-0.220	-0.165	*
Tired&1	-0.005	0.018	-0.029	0.039	

Step 4 Model 3: RDSEM

- Within-level part of the model:



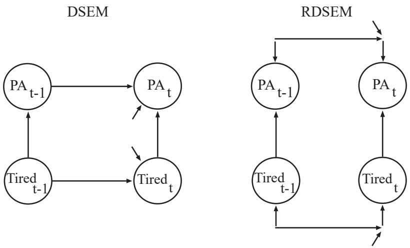

```
%WITHIN%
PA ON TIRED;
PA ON PA&1;
TIRED ON TIRED&1;
```




```
%WITHIN%
PA ON TIRED;
PA^ ON PA^1;
TIRED^ ON TIRED^1;
```

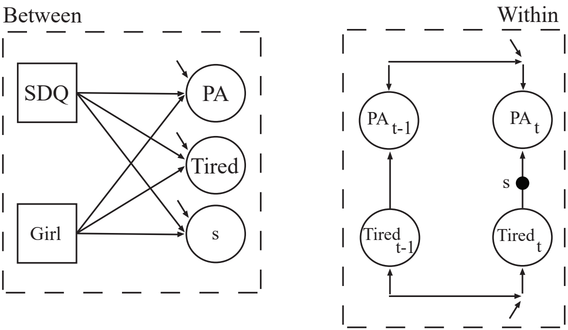
- RDSEM is like regular twolevel analysis regressing PA on Tired: No lagged effects but instead auto-correlated residuals
- RDSEM = DSEM if no covariates (Tired in this case)

DSEM vs RDSEM Continued



- Apart from $Tired_t \rightarrow PA_t$, the two models have different implications:
  - DSEM: Indirect influence $Tired_{t-1} \rightarrow PA_{t-1} \rightarrow PA_t$
  - RDSEM: Only the residual of PA_{t-1} influences PA_t
- The 2 models give different estimates of the regression of PA_t on $Tired_t$ - small difference in this example but this is not always the case
- Asparouhov & Muthén (2019). Comparison of models for the analysis of intensive longitudinal data. SEM journal. Tables 6 and 7
- RDSEM can be used to study the relationship between two outcomes controlling for covariates, e.g., a trend such as daily cycles

RDSEM Analysis with a Random Slope Regressed on Between-Level Covariates



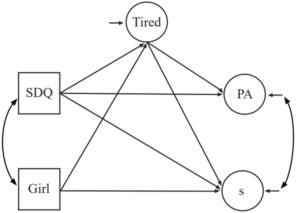
RDSEM Input

```
USEVARIABLES = pa tired SDQemotAA girl;
LAGGED = pa(1) tired(1);
TINTERVAL = hrs(2 time);
CLUSTER = id;
BETWEEN = SDQemotAA girl;
! 24*7 = 168 (same as time >84):
USEOBSERVATIONS = hrs LE 168
AND id NE 240 AND id NE 249
AND id NE 78 AND id NE 531;
DEFINE:
girl = sexAA - 1;
CENTER SDQemotAA(GRANDMEAN);
ANALYSIS:
TYPE = TWOLEVEL RANDOM;
ESTIMATOR = BAYES;
BITERATIONS = (1000);
THIN = 10;
PROCESSORS = 8;
MODEL:
%WITHIN%
s | pa ON tired;
pa^ ON pa^1;
tired^ ON tired^1;
%BETWEEN%
pa tired s ON SDQemotAA girl;
pa tired s WITH pa tired s;
OUTPUT:
STANDARDIZED TECH1
TECH4 TECH8;
PLOT:
TYPE = PLOT3;
```

RDSEM Output, Standardized Estimates

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Within-Level Standardized Estimates Averaged Over Clusters					
S | PA ON					
TIRED	-0.199	0.015	-0.228	-0.169	*
PA^ ON					
PA^1	0.390	0.016	0.359	0.419	*
TIRED^ ON					
TIRED^1	0.460	0.015	0.431	0.488	*
Residual Variances					
PA	0.781	0.012	0.758	0.805	*
TIRED	0.788	0.014	0.762	0.814	*
Between Level					
S ON					
SDQEMOTAA	-0.168	0.068	-0.297	-0.029	*
GIRL	0.005	0.070	-0.133	0.145	
PA ON					
SDQEMOTAA	-0.330	0.046	-0.418	-0.241	*
GIRL	0.007	0.047	-0.090	0.091	
TIRED ON					
SDQEMOTAA	0.227	0.049	0.132	0.323	*
GIRL	0.100	0.049	0.000	0.190	*

Alternative RDSEM Between Model: Mediation



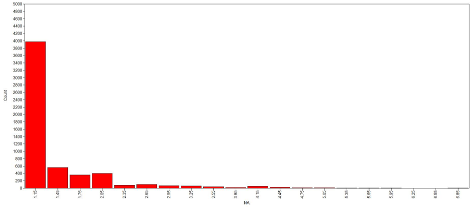

```
%BETWEEN%
pa ON tired SDQemotAA girl;
tired ON SDQemotAA girl;
s ON tired SDQemotAA girl;
pa WITH s;
```

- Standardized between-level estimates

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
PA ON					
TIRED	-0.435	0.055	-0.536	-0.319	*
SDQEMOTAA	-0.230	0.047	-0.320	-0.137	*
GIRL	0.055	0.043	-0.035	0.139	
TIRED ON					
SDQEMOTAA	0.228	0.047	0.142	0.326	*
GIRL	0.104	0.049	0.015	0.202	*
S ON					
TIRED	-0.332	0.118	-0.525	-0.049	*
SDQEMOTAA	-0.092	0.073	-0.233	0.037	
GIRL	0.039	0.075	-0.098	0.192	

Section 7 Categorical Outcome

Categorical Outcome: Negative Affect



- 60% at the lowest value of NA: Treating the variable as continuous leads to model misspecification of linear models such as DSEM typically causing underestimated regression slopes
- Dichotomize the variable (trichotomize?). Censored? Two-part?
- Asparouhov et al. (2018). Dynamic structural equation models. SEM
    - Normally distributed latent response variable Y* underlying the categorical observed variable Y is assumed with the regular linear DSEM applied to Y* (see also Mplus Web Talk 4, Part 2)
    - Nominal and count not available for DSEM

Binary Outcome: Input for Dichotomized NA Individuals not Changing over Time Deleted

```
USEOBSERVATIONS = hrs le 168
AND id ne 240 AND id ne 249
AND id ne 78 AND id ne 531
AND id ne 45 AND id ne 319
AND id ne 352 AND id ne 254
AND id ne 200 AND id ne 320
AND id ne 347 AND id ne 385
AND id ne 313 AND id ne 2
AND id ne 160 AND id ne 503
AND id ne 442 AND id ne 260
AND id ne 119 AND id ne 263
AND id ne 338 AND id ne 523
AND id ne 570 AND id ne 462
AND id ne 256;

USEVARIABLES = pa nabin;
CATEGORICAL = nabin;
CLUSTER = id;
TINTERVAL = hrs (2 time);
LAGGED = pa(1) nabin(1);
```

```
DEFINE:
        IF(na GT 1)THEN nabin = 1
        ELSE nabin = 0;
        ! Alternative: Use CUT na(1);
        ! and keep the variable name na
ANALYSIS:
        TYPE = TWOLEVEL;
        ESTIMATOR = BAYES;
        BITERATIONS = (2000);
        THIN = 10;
        PROCESSORS = 8;
MODEL:
        %WITHIN%
        pa ON pa&1 nabin&1;
        nabin ON nabin&1 pa&1;
        %BETWEEN%
        pa nabin WITH pa nabin ;
```

Binary Outcome: Original NA vs Binary NA Standardized Output

- Original NA (treated as continuous)

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
Within Level					
PA ON					
PA&1	0.414	0.017	0.379	0.445	*
NA&1	**-0.002**	0.017	-0.034	0.032	
NA ON					
NA&1	0.251	0.019	0.210	0.285	*
PA&1	**-0.039**	0.019	-0.075	-0.001	*

- Binary NA (dichotomized)

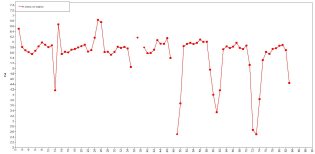
Within Level					
PA ON					
PA&1	0.368	0.017	0.336	0.403	*
NABIN&1	**-0.192**	0.022	-0.236	-0.149	*
NABIN ON					
NABIN&1	0.501	0.033	0.437	0.566	*
PA&1	**-0.089**	0.022	-0.133	-0.045	*

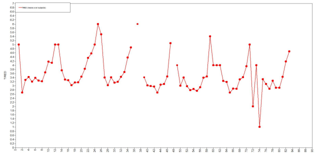
Random Residual Correlation with Binary NA

```
MODEL:
        %WITHIN%
        pa ON pa&1 nabin&1;
        nabin ON nabin&1 pa&1;
        logvp | pa;
        ! binary na does not have a free
        ! residual variance parameter
        c | pa WITH nabin;
        ! random biserial correlation
        %BETWEEN%
        pa nabin logvp c WITH pa nabin logvp c;
        [logvp] (mp);
        logvp (sp);
        [c] (mc);
MODEL
CONSTRAINT:
        NEW (meanvp medr);
        meanvp = exp(mp+sp/2);
        medr = (exp(2*mc)-1)/(exp(2*mc)+1);
```

Section 8 Cross-Classified Analysis

- Time series plots of PA and Tired averages in the sample:





- The sample averages have varying precision over time
  - can we get a time series plot for model-estimated values?
- Yes, by cross-classified analysis

Cross-Classified Time Series Analysis ($N > 1$)

- Two between-level cluster variables: person crossed with time (one observation for a given person at a given time point)
- Generalization of the two-level model providing more flexibility: random effects can vary across not only persons but also time
- Consider the two-level model with a random intercept/mean:

$$y_{it} = \underbrace{\alpha + \alpha_i}_{\text{Between person}} + \underbrace{\beta\, y_{w,it-1} + \varepsilon_{it}}_{\text{Within person}}. \tag{22}$$

The corresponding cross-classified model is:

$$y_{it} = \underbrace{\alpha + \alpha_i}_{\text{Between person}} + \underbrace{\alpha_t}_{\text{Between time}} + \underbrace{\beta\, y_{w,it-1} + \varepsilon_{it}}_{\text{Within person}}. \tag{23}$$

  - α_i and α_t are normally distributed with zero means
  - The Bayes MCMC algorithm is more complex and slower

Step 5: Cross-Classified Analysis of PA and Tired Mplus Input

```
          USEVARIABLES = pa tired;
          LAGGED = pa(1) tired(1);
          TINTERVAL = hrs(2 time);
          CLUSTER = id time;
ANALYSIS:
          TYPE = CROSSCLASSIFIED;
          ESTIMATOR = BAYES;
          BITERATIONS = (2000);
          THIN = 10;
          PROCESSORS = 12;
```

```
MODEL:
          %WITHIN%
          pa ON pa&1;
          tired ON tired&1;
          pa ON tired tired&1;
          tired ON pa&1;
          %BETWEEN id%
          pa WITH tired;
          %BETWEEN time%
          pa WITH tired;
OUTPUT:
          STANDARDIZED TECH1 TECH8;
PLOT:
          TYPE = PLOT3;
          FACTORS = ALL(50);
```

Comparing Two-Level and Cross-Classified Estimates

- Standardized within-level estimates

	Estimate	Posterior S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
		Two-level			
Within Level					
PA ON					
PA&1	0.395	0.015	0.366	0.424	*
Tired	**-0.193**	0.014	-0.220	-0.166	*
Tired&1	-0.005	0.018	-0.030	0.042	
		Cross-classified			
Within Level					
PA ON					
PA&1	0.383	0.016	0.393	0.413	*
Tired	**-0.193**	0.013	-0.218	-0.166	*
Tired&1	-0.012	0.017	-0.020	0.045	

ID Between-Level Estimates: Two-Level vs Cross-Classified

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%	Significance
Two-level					
PA WITH					
TIRED	-0.549	0.085	-0.737	-0.401	*
Means					
PA	5.747	0.058	5.630	5.858	*
TIRED	3.386	0.084	3.222	3.558	*
Variances					
PA	0.738	0.074	0.610	0.898	*
TIRED	1.487	0.156	1.229	1.850	*
Cross-classified					
PA WITH					
TIRED	-0.555	0.084	-0.730	-0.408	*
Means					
PA	5.746	0.061	5.630	5.865	*
TIRED	3.394	0.112	3.169	3.612	*
Variances					
PA	0.738	0.076	0.604	0.896	*
TIRED	1.550	0.155	1.253	1.857	*

Between-Level Estimates of Cross-Classified Analysis

	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%	Significance
Between ID Level					
PA WITH					
TIRED	-0.555	0.084	-0.730	-0.408	*
Means					
PA	5.746	0.061	5.630	5.865	*
TIRED	3.394	0.112	3.169	3.612	*
Variances					
PA	0.738	0.076	0.604	0.896	*
TIRED	1.550	0.155	1.253	1.857	*
Between TIME Level					
PA WITH					
TIRED	-0.022	0.013	-0.052	-0.001	*
Variances					
PA	0.012	0.004	0.006	0.023	*
TIRED	0.265	0.062	0.173	0.412	*

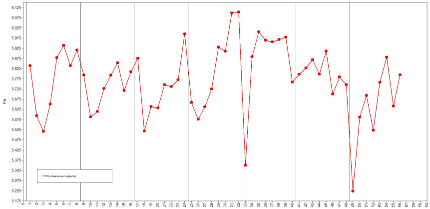
Comparing Stand'd Results for Tinterval = 1, 2, 3

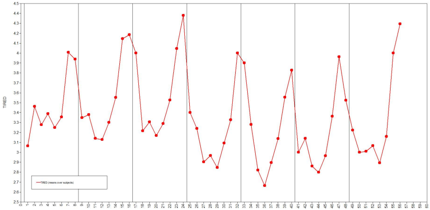
	Posterior Estimate	S.D.	95% C.I. Lower 2.5%	Upper 2.5%	Significance
			Tinterval = 1		
Within Level					
PA ON					
PA&1	0.517	0.016	0.486	0.547	*
TIRED	-0.163	0.016	-0.194	-0.129	*
TIRED&1	0.017	0.020	-0.023	0.055	
TIRED ON					
TIRED&1	0.565	0.014	0.534	0.592	*
PA&1	-0.031	0.015	-0.060	-0.002	*
			Tinterval = 2		
Within Level					
PA ON					
PA&1	0.383	0.016	0.353	0.413	*
TIRED	-0.193	0.013	-0.218	-0.166	*
TIRED&1	0.012	0.017	-0.020	0.045	
TIRED ON					
TIRED&1	0.416	0.015	0.387	0.445	*
PA&1	-0.026	0.016	-0.057	0.005	
			Tinterval = 3		
Within Level					
PA ON					
PA&1	0.349	0.015	0.320	0.378	*
TIRED	-0.204	0.013	-0.230	-0.177	*
TIRED&1	0.008	0.015	-0.022	0.039	
TIRED ON					
TIRED&1	0.373	0.014	0.344	0.400	*
PA&1	-0.024	0.014	-0.054	0.004	

Time Series Plot of Estimated Random Effects (Factor Scores) for PA and Tired

- Cross-classified analysis with CLUSTER = id time
  - The two between levels are referred to as level2a for time and level2b for id
  - The acronym B2a is used for the between-time level factor score to be plotted
- Plotting:
  - PLOT option: FACTORS=ALL(50)
  - Plot menu option: Time series plots (sample value, ACF, PACF, **estimated factor score**)
    - B2a_PA, mean
    - B2a_Tired, mean

Time Series Plot for PA and Tired Factor Scores, Tue-Mon





- How do you model these trends/cycles? Future DSEM web talk
  - See also Short Course Topic 13, Part 8 with a heart rate example

Mplus Run for CrossClassified

- Plots for step 5 cross-classified analysis of PA and Tired
  - Time series plot for estimated factor scores: B2a_PA, B2a_Tired
  - Right-click options: Mean line, time lines
- More plot options discussed in future web talk
  - Loop insert (cycles, weekday effects)

Section 9 How Large do N and T Have to Be? Monte Carlo Simulations

How Large do N and T Have to Be? Checklist

- Do you have enough timepoints? Recommendations:
  - At least 15 - 20 (if not, do single-level, wide analysis)
  - At least 25 for good performance with random slopes and var's
  - At least 50 for good N = 1 performance
- Do enough individuals have enough timepoints without missing?
  - If not, random slope and variance estimation can be problematic
- Does your outcome have variation across time?
  - Delete individuals with no variation
- Do you have enough individuals for random effects modeling on between?
  - A minimum of 50 recommended but many more may be needed for between-level variances (500 in the simulation below)
- Does your outcome have a strong trend that should be modeled?
  - If yes, use RDSEM with a function of time as covariate
- Do your own Monte Carlo simulation

How Large do N and T Have to Be? Monte Carlo

- Schultzberg & Muthén (2018). Number of subjects and time points needed for multilevel time series analysis: A simulation study of dynamic structural equation modeling. Structural Equation Modeling
  - 9 univariate DSEM models with varying complexity
  - Example: A univariate model with random intercept, random auto-regression and random residual variance needed T = 25 for N = 150 and T = 50 for N = 100
  - "What is worse, having a lower N or a lower T? Can large N compensate for small T better than large T can compensate for small N? The answer seems to be clear: Large N is better. That is, large N seems able to compensate for small T, better than large T can compensate for small N. — With that said, the random AR and residual variance do need fairly large T to be well estimated. If AR or residual variance is of substantive interest rather than just a heterogeneity feature to control for, many repeated measures will be needed." (p. 511)
- Current study: N = 240 and average T = 24, so sufficient for a univariate model - is it sufficient for a multivariate model?

Monte Carlo Simulations

- Parameter values obtained via SVALUES in the real-data run based on Step 4 Model 2 (lag-0 effect of Tired on PA)
- N = 250, T = 25

```
MONTECARLO:
        NAMES = pa tired;
        NOBSERVATIONS = 6250;
        NREPS = 500;
        CSIZES = 250(25);
        NCSIZES = 1;
        LAGGED = pa(1) tired(1);
        ! See UG ex12.6:
        ! REPSAVE = ALL;
        ! SAVE = step4Model2rep*.dat;
        ! Or for one replication:
        ! SAVE = step4Model2.dat;
ANALYSIS:
        TYPE = TWOLEVEL;
        ESTIMATOR = BAYES;
        BITERATIONS = (1000);
        PROCESSORS = 8;
```

```
MODEL
POPULATION:   %WITHIN%
        pa ON pa&1*0.39510;
        pa ON tired*-0.11099;
        pa ON tired&1*0.00309;
        tired ON tired&1*0.46071;
        tired ON pa&1*-0.01164;
        pa*0.46190;
        tired*1.39159;

        %BETWEEN%
        pa WITH tired*-0.54293;
        [ pa*5.74638 ];
        [ tired*3.39010 ];
        pa*0.73492;
        tired*1.48918;
MODEL:
        ! Copy MODEL POPULATION
```

Monte Carlo Results Summary

- N = 250, T = 25: Excellent results (estimates, SEs, coverage)
- N = 250, T = 15: Good results (a bit low coverage for Tired AR)
- N = 250, T = 10: Not acceptable results

- A more realistic analysis is obtained with varying cluster sizes (individuals with different number of time points)
- The real data has max 84 time points and average number of time points = 24
- An approximation to this is obtained by e.g. the 5 cluster sizes: CSIZES = 10(5) 65(15) 100(25) 65(50) 10(80);
  - 10 clusters of size 5, 65 clusters of size 15, etc for a total of 250 clusters (N)
  - Still excellent results

Monte Carlo Simulations with Random Slopes

- Parameter values obtained via SVALUES in the real-data run based on Step 4 Model 1 with added random slopes, random variances, and random covariance (total of 9 random effects)
- N = 250 with the 5 cluster sizes used on the previous slide
  - Good results, except between-level variances have somewhat biased estimates (overestimated) and SEs - N $\geq$ 500 is needed
- New in version 8.9 (message shown in TECH9):

  351 CLUSTERS WERE REMOVED BECAUSE THEIR GENERATED RANDOM EFFECTS PRODUCED NON-STATIONARY TIME SERIES OR NON-POSITIVE DEFINITE COVARIANCE MATRICES.
- With random slopes, variances, and covariances, non-stationarity is likely for some clusters in some replications
- In the past, such replications have been deleted leading to a low percentage of reported replications - now the cluster is thrown out but the replication kept
- 351 clusters is less than 1% of the 250*500 of the generated clusters

CrossClassified Monte Carlo

```
MONTECARLO:
        NAMES = pa tired;
        NOBSERVATIONS = 21000;
        NREPS = 100;
        CSIZES = 10[5(1)] 65[15(1)]
        100[25(1)] 65[50(1)] 10[80(1)];
        NCSIZES = 5[5];
        LAGGED = pa(1) tired(1);
ANALYSIS:
        TYPE = CROSSCLASSIFIED;
        ESTIMATOR = BAYES;
        BITERATIONS = (1000);
        PROCESSORS=8;
```

```
MODEL
POPULATION:  %WITHIN%
        pa ON pa&1*0.38327;
        pa ON tired*-0.11689;
        pa ON tired&1*0.00723;
        tired ON tired&1*0.41643;
        tired ON pa&1*-0.04240;
        pa*0.45668;
        tired*1.26884;
        %BETWEEN level2a%
        pa WITH tired*-0.02232;
        pa*0.01241;
        tired*0.26523;
        %BETWEEN level2b%
        pa WITH tired*-0.55530;
        [ pa*5.74610 ];
        [ tired*3.39425 ];
        pa*0.73831;
        tired*1.55047;
```

- In the SVALUES from the real-data run, %Between time% has to be replaced by %Between level2a% and %Between id% has to be replaced by %Between level2b%

Section 10 References

DSEM References and Workshop Videos

- Technical and applied papers:
  http://www.statmodel.com/TimeSeries.shtml
- Short Course YouTube videos and handouts:
  http://www.statmodel.com/topic12.shtml
  http://www.statmodel.com/topic13.shtml
- Hamaker YouTube video tutorials:
  https://www.youtube.com/watch?v=dA3HvJZDzeo&list=PLet3DgvxBn2S7N2hVW4COAwH3_VaRoujd
- Bayesian analysis in Mplus:
  - Short Course Topic 9:
    http://www.statmodel.com/topic9.shtml
  - Quick version; Short Course Topic 11:
    http://www.statmodel.com/topic11.shtml
  - Chapter 9 in the Muthén, Muthén & Asparouhov (2016) book Regression and Mediation Analysis Using Mplus