A Comparison of Cluster Analysis and Growth Mixture Modeling in the Recovery of Developmental Trajectory Classes

Madeline M. Carrig

and

Daniel J. Bauer

University of North Carolina at Chapel Hill

Introduction

- Theoretical and empirical work in a number of areas of developmental research suggests that populations of interest are often comprised of subpopulations characterized by qualitatively different patterns of change over time.
- One recent and significant innovation in quantitative methodology, Muthén's (in press) latent variable growth mixture modeling (LGM) procedure, promises to permit the study of growth trajectory mixtures in research samples.
- However, at present, this new technique is relatively unstudied, and its empirical properties are hence poorly understood.

Aims of the Present Study

Our goal was to utilize computer simulation technology to explore the large-sample properties of LGM. In so doing, we planned to:

- Investigate the ability of LGM to recover known growth function subgroups.
- Compare the performance of LGM to more traditional approaches to inferring population subgroup membership from observed data.
- Investigate the taxonomic utility of LGM and more traditional approaches across a variety of conditions commonly encountered in applied developmental research.

Sample

- Data sets simulating large random samples from populations comprised of distinct subpopulations were generated in SAS.
- The subpopulations giving rise to the large-sample data sets were defined as follows (see Figures 1 and 2):
  - **Group 1, *Low Stable***: μ (intercept) = 50, μ (slope) = 0
  - **Group 2, *Rapid Linear Growth***: μ (intercept) = 50, μ (slope) = 15
  - **Group 3, *Moderate Linear Growth***: μ (intercept) = 50, μ (slope) = 7.5
  - **Group 4, *High Stable***: μ (intercept) = 110, μ (slope) = 0
  - Within subpopulation, intercept and slope possess individual variability, and are normally distributed.

- Data observations were simulated as follows:
  - Five repeated measures were produced for each individual.
  - Time-specific disturbances in individual scores were generated as random normal deviates. The dispersion of disturbances was monotonically increasing over time, and was set such that modeled growth trajectories accounted for the desired level of variability in observed scores.
  - n = 2000 per subpopulation sampled.
- Nine simulated data sets were created, one for each of nine conditions defined by number of subpopulations present in population sampled (number of subgroups = 2, 3, 4) and amount of variability in observed scores explained by modeled growth trajectories (Multiple R^2 for repeated measures < .5, .7, .9).

Figure 1

Mean Trajectories of 4 Subpopulations

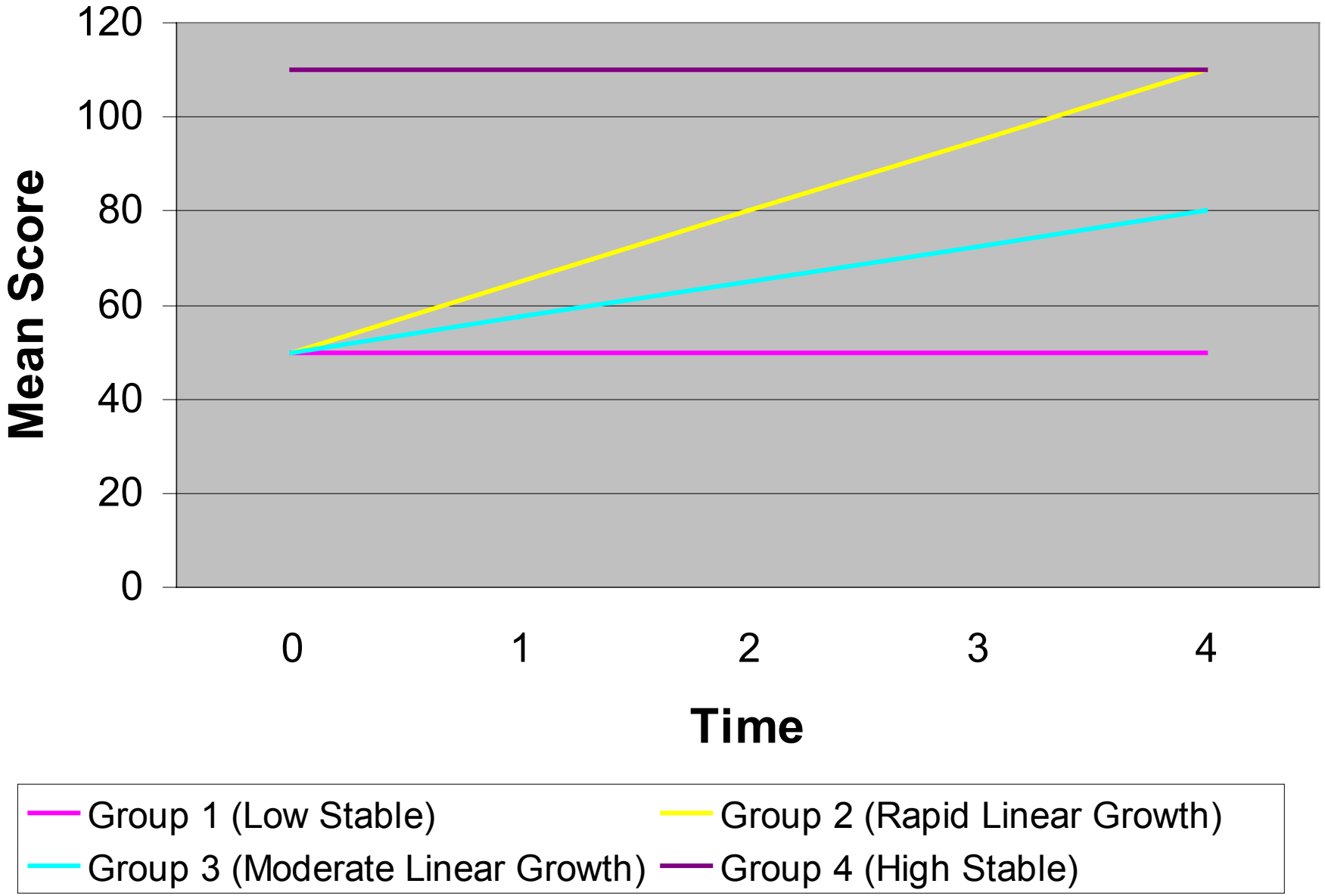
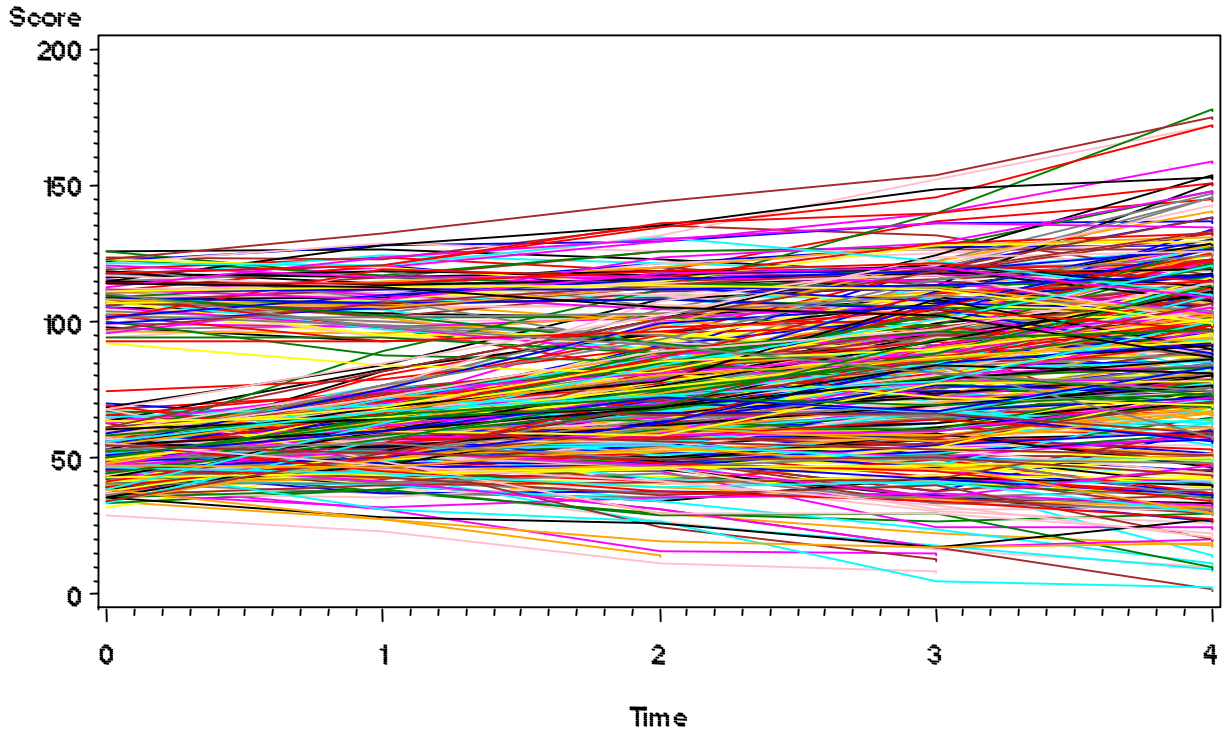


Figure 2

Example Simulated Sample Data Set: 4 Subpopulations, Multiple $R^2 < 90\%$



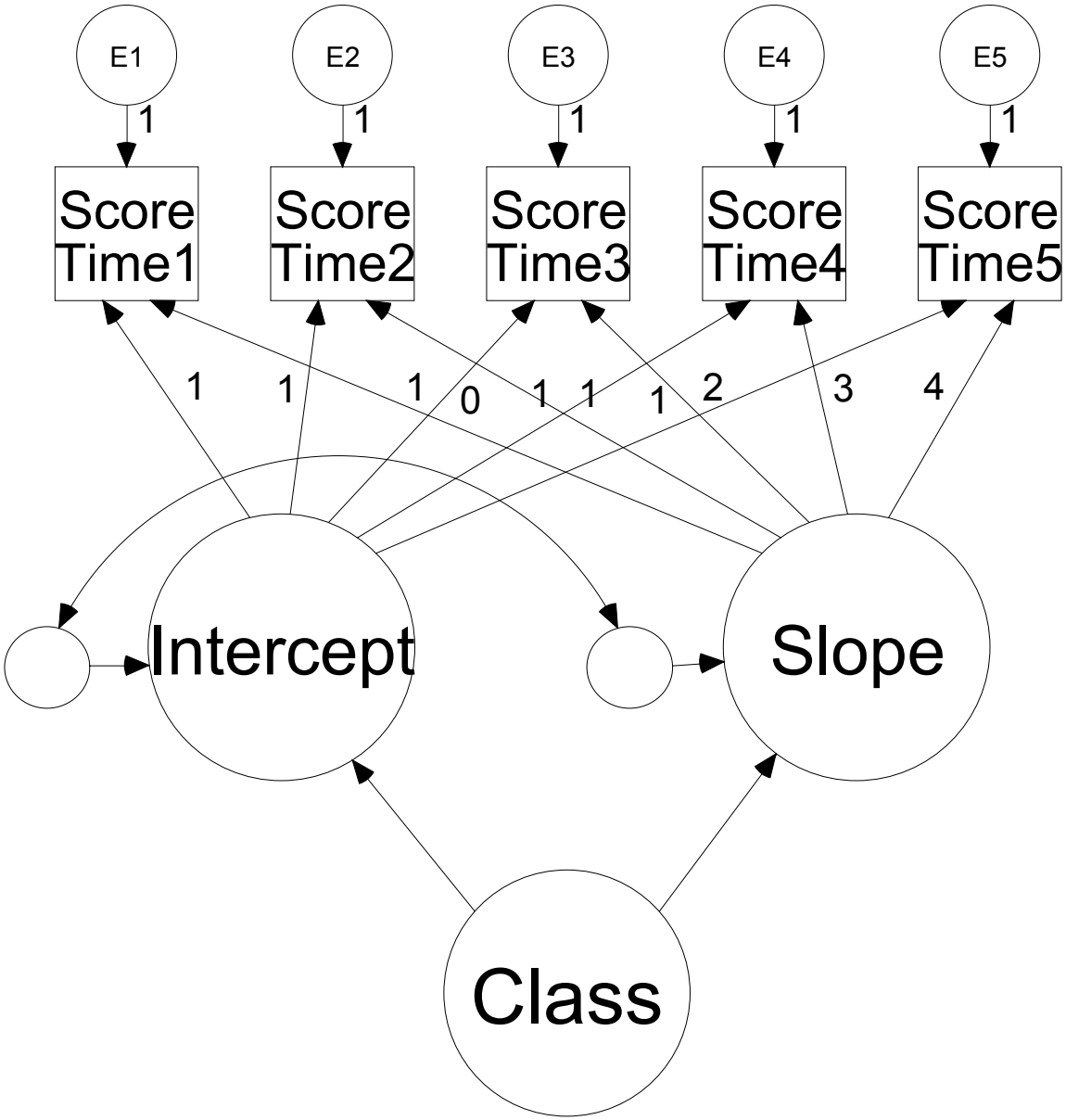
Procedure

- Resulting samples were analyzed using two approaches.
- **First,** latent growth curve mixture models (see Figure 3) were estimated using the statistical program Mplus (Muthén & Muthén, 1988).
  - Both fixed- (for which the variance of the intercept and slope factors were fixed to zero) and random- (for which individual variability in intercept and slope was modeled) effects models were estimated for each sample.
  - Mplus provides information about fuzzy class membership (i.e., for each individual, it provides a posterior probability for each class; probabilities range from 0 to 1 and sum to 1 across classes). For the present study, disjoint clusters were formed by assigning each individual to one, and only one, cluster on the basis of the highest posterior probability observed for that individual.

- **Second,** *k*-means, a traditional method for disjoint clustering of observations, was implemented using PROC FASTCLUS in SAS with options CONVERGE = 0, MAXITER = 100, and MAXCLUSTERS = number of subgroups in population sampled.
  - *k*-means clustering was performed on two sets of variables:
    - Repeated measures for t = 0,1,2,3,4; and
    - Individual OLS-regression estimates of intercept and slope.
  - Because variables with larger variances exert a larger influence in calculating a *k*-means cluster solution, regression estimates were standardized to unit variance prior to clustering.
- The clusters (classes) resulting from *k*-means and LGM analyses were associated with known population subgroups, and the extent to which individuals were correctly classified was assessed.

Figure 3

Latent Variable Growth Mixture Model



Percent of Individuals Correctly Classified
LGM

Number of Subpopulations/Group		Multiple R^2 Level					
		50%		70%		90%	
		Fixed	*Random*	*Fixed*	*Random*	*Fixed*	*Random*
2	G1: Low Stable	84.50%	91.35%	85.95%	92.95%	86.65%	92.90%
	G2: Rapid Linear	90.80%	86.30%	92.44%	88.88%	93.15%	91.00%
	Overall	87.65%	88.83%	89.19%	90.92%	89.90%	91.95%
3	G1: Low Stable	52.25%	93.45%	58.30%	92.15%	60.25%	93.50%
	G2: Rapid Linear	66.25%	78.10%	68.15%	89.28%	70.45%	90.05%
	G3: Mod. Linear	65.90%	1.85%	65.90%	0.40%	65.65%	0.05%
	Overall	61.47%	57.80%	64.12%	60.60%	65.45%	61.20%
4	G1: Low Stable	50.60%	94.00%	56.80%	90.30%	59.10%	87.75%
	G2: Rapid Linear	66.20%	77.95%	68.50%	91.04%	70.50%	2.70%
	G3: Mod. Linear	66.75%	1.30%	66.35%	0.20%	66.20%	60.45%
	G4: High Stable	99.10%	99.85%	99.30%	100.00%	99.25%	100.00%
	Overall	70.64%	68.25%	72.72%	70.36%	73.74%	62.69%

Percent of Individuals Correctly Classified
k-means

Number of Subpopulations/Group		Multiple R^2 Level					
		50%		70%		90%	
		Repeated Measures	*Regression Estimates*	*Repeated Measures*	*Regression Estimates*	*Repeated Measures*	*Regression Estimates*
2	G1: Low Stable	87.10%	72.50%	89.10%	78.40%	88.70%	89.20%
	G2: Rapid Linear	89.80%	67.45%	91.15%	71.10%	93.45%	76.60%
	Overall	88.45%	69.98%	90.13%	74.75%	91.08%	82.90%
3	G1: Low Stable	64.45%	55.45%	65.10%	58.50%	66.80%	61.50%
	G2: Rapid Linear	61.70%	63.75%	68.25%	66.50%	71.20%	66.45%
	G3: Mod. Linear	61.30%	36.25%	62.95%	38.65%	64.40%	40.95%
	Overall	62.48%	51.82%	65.43%	54.55%	67.47%	56.30%
4	G1: Low Stable	64.50%	59.05%	88.35%	63.65%	88.45%	93.20%
	G2: Rapid Linear	62.05%	57.15%	89.75%	62.70%	92.25%	88.10%
	G3: Mod. Linear	60.95%	57.55%	0.20%	60.35%	0.05%	0.15%
	G4: High Stable	95.20%	98.05%	50.05%	99.45%	51.00%	50.60%
	Overall	70.68%	67.95%	57.09%	71.54%	57.94%	58.01%

Patterns of Misclassification
4 Subpopulations, Multiple $R^2 < 50\%$

Clustering Approach	Group	%CC	%IC in G1	%IC in G2	%IC in G3	%IC in G4
Fixed-effects LGM	G1: Low Stable	50.60%	.	3.30%	46.05%	0.05%
	G2: Rapid Linear	66.20%	0.60%	.	32.95%	0.25%
	G3: Mod. Linear	66.75%	12.85%	20.35%	.	0.05%
	G4: High Stable	99.10%	0.00%	0.35%	0.55%	.
Random-effects LGM	G1: Low Stable	94.00%	.	4.80%	1.10%	0.10%
	G2: Rapid Linear	77.95%	20.25%	.	1.55%	0.25%
	G3: Mod. Linear	1.30%	65.50%	33.05%	.	0.15%
	G4: High Stable	99.85%	0.00%	0.10%	0.05%	.
***k*-means clustering, repeated measures**	G1: Low Stable	64.50%	.	2.45%	33.00%	0.05%
	G2: Rapid Linear	62.05%	2.05%	.	35.70%	0.20%
	G3: Mod. Linear	60.95%	20.10%	18.85%	.	0.10%
	G4: High Stable	95.20%	0.25%	0.10%	4.45%	.

Note: %CC = % Correctly Classified, %IC = % Incorrectly Classified.

Summary of Findings

Findings from LGM Analyses

- For all models estimated, convergence in Mplus was obtained only when known population parameters were provided as start values.
- LGM produced moderate-to-high correct classification rates (CCR), with overall (averaged across subgroups) CCR ranging from 57.80% to 91.95%, depending on sample and type of model estimated.
- As expected, in general, samples with higher Multiple R^2 levels were associated with higher CCR.
- The 3-subgroup condition was associated with substantially lower CCR than the 2-subgroup condition, likely due to the decrease in cluster separation resulting from the addition of the Moderate Linear Growth subgroup.

- CCR for subgroups 1, 2, and 3 in the 3-subgroup condition were approximately equal to CCR for these groups in the 4-subgroup condition, which reflected the addition of the High Stable subgroup. CCR for the High Stable subgroup, which was distinctly separated from all other subgroups in initial level, but not slope, approached 100%.
- Regardless of the number of total subpopulations sampled, in general, the random-effects model produced *higher* CCR for the Low Stable, Rapid Growth, and High Stable subgroups. In contrast, with one exception, use of a random-effects model resulted in dramatically *lower* CCR for the Moderate Linear Growth subgroup. This resulted in slightly lower CCR, overall, in the 3- and 4-subgroup conditions for the random-effects models.

Findings from *k*-means Clustering Analyses

- *k*-means clustering of repeated measures typically resulted in higher overall (averaged across subgroups) CCR than clustering of OLS regression estimates of individual intercept and slope.
- In general, LGM overall CCR equaled or exceeded *k*-means overall CCR for each sample.

Patterns of Misclassification Across Approaches

- In the present set of samples, fixed-effects LGM and *k*-means clustering of repeated measures were likely to misclassify *Low Stable* individuals in the *Moderate Linear Growth* subgroup.
- Whereas fixed-effects LGM and *k*-means clustering of repeated measures were likely to misclassify *Rapid Linear Growth* individuals in the *Moderate Linear Growth* subgroup, random-effects LGM was more likely to misclassify these individuals in the *Low Stable* subgroup.

Conclusions

- LGM appears to hold great promise as a new analytic tool for the testing of developmental theories of stability and change. In the present study, LGM produced moderate-to-high rates of correct classification of individuals into subgroups. Because disjoint (and not fuzzy) class assignment was evaluated, these rates may even underestimate somewhat LGM's true classification performance.
- Findings suggest that the ability of LGM to recover growth trajectory subpopulations, and patterns of misclassification, may vary depending on characteristics of the sampled population (e.g., number and separation of subpopulations). Surprisingly, random-effects models were not consistently superior to fixed-effects models in the recovery of developmental pathway subgroups.

- Findings should be replicated and extended in a study using multiple replications per experimental condition and incorporating additional conditions characterized by variations in sample size, number of time points, and functional form of trajectories.
- Mplus estimation of the LGM models required excellent start values (in this case, known subpopulation parameters) to achieve convergence. *k*-means clustering of repeated measures, which under many conditions produced disjoint cluster CCR rivaling those of LGM, could profitably be used for preliminary clustering of data and calculation of start values for later LGM in Mplus.