A Framework for Investigating the Performance of Latent Growth Mixture Models.

Paul Dudgeon

Melbourne School of Psychological Sciences
The University of Melbourne. Vic. 3010
AUSTRALIA
dudgeon@unimelb.edu.au

Modern Modeling Methods (M^3) Conference
University of Conneticut
21-22 May 2013

Monte Carlo Simulation Studies in SEM

Multivariate Behavioral Research, *35* (2), 137-167

Design and Analysis of Monte Carlo Experiments: Attacking the Conventional Wisdom

Anders Skrondal
Department of Epidemiology
National Institute of Public Health, Oslo

Many published simulation studies rely on...

- A small number of conditions being investigated.
- Reporting descriptive "eyeballing" of findings as tables.

Skrondal instead advocated researcher to...

- Use large factorial designs, especially fractional types
- Apply statistical models to outcome measure (meta-model)
- Quantify main and interaction effects using a meta-model
- Report results of meta-model rather than tables

Monte Carlo Simulation Studies in SEM (cont.)

Skrondal's recommendations have found limited appeal (so far!).

- Cited in 11 peer-reviewed articles.
- 4 published applications using his approach (De Roover et al. 2013)

Possible reasons:

- Long-held conventions about fully-crossed factorial approaches
- Highly cited Monte Carlo studies (e.g., two by Hu & Bentler in 1990s) have become templates
- Skrondal's extensive description of FFDs may be off-putting.
- Fractional factorial designs (FFDs) rarely encountered (perhaps?)

Current presentation partly aims to renew interest in his approach—and extend it into a general framework.

Some Initial Definitions

Statistical Model (or just *model*)

It is the (complex) model we wish to investigate in the Monte Carlo simulation. It could be, e.g., a latent growth mixture model (LGMM) or a multilevel SEM, and it is some aspect of its performance or behaviour (e.g., GOF measures) that is the dependent variable of the Monte Carlo study.

Meta-Model

A formal statistical model used to analyse the outcome measure of the Monte Carlo study. It is typically a general(ized) linear model appropriate for the outcome measure, and the independent variables of the meta-model reflect the factors of the experimental design used in the Monte Carlo simulation.

Always refer to *meta-model* explicitly to ensure the two are differentiated when ambiguity may arise.

Why Undertake Monte Carlo Simulations?

Monte Carlo simulations can be motivated by different rationales:

1. Evaluate model fidelity

Why Undertake Monte Carlo Simulations?

Monte Carlo simulations can be motivated by different rationales:

1. Evaluate model fidelity
2. Investigate robustness properties

Why Undertake Monte Carlo Simulations?

Monte Carlo simulations can be motivated by different rationales:

1. Evaluate model fidelity
2. Investigate robustness properties
3. Diagnostic assessments

Why Undertake Monte Carlo Simulations?

Monte Carlo simulations can be motivated by different rationales:

1. Evaluate model fidelity
2. Investigate robustness properties
3. Diagnostic assessments
4. Avoid applying for IRB approval

Why Undertake Monte Carlo Simulations?

Monte Carlo simulations can be motivated by different rationales:

1. Evaluate model fidelity
2. Investigate robustness properties
3. Diagnostic assessments
4. Avoid applying for IRB approval
5. “Poking-and-fiddling” to understand limitations and strengths

Why Undertake Monte Carlo Simulations?

Monte Carlo simulations can be motivated by different rationales:

1. Evaluate model fidelity
2. Investigate robustness properties
3. Diagnostic assessments
4. Avoid applying for IRB approval
5. “Poking-and-fiddling” to understand limitations and strengths

Models such as LGMMs and MSEMs are very complex, and have many hidden potential problems.

- Often unaware of problems, even when models (seem to) work
- Any model almost certainly wrong when applied to sample data
- Better understanding of possible consequences from being wrong
- Use and interpret complex models more prudently

Proposed framework enables systematic investigation of potential threats from different sources, and potential interactions among such sources.

Three-Phase Hierarchical Framework

Systematically investigates three kinds of threats or problems in the different phases of fitting a statistical model to sample data

1. **Data generating process**

Three-Phase Hierarchical Framework

Systematically investigates three kinds of threats or problems in the different phases of fitting a statistical model to sample data

1. **Data generating process**

2. **Model specification** fitted to data.

Three-Phase Hierarchical Framework

Systematically investigates three kinds of threats or problems in the different phases of fitting a statistical model to sample data

1. **Data generating process**
2. **Model specification** fitted to data.
3. **Model estimation** options

Three-Phase Hierarchical Framework

Systematically investigates three kinds of threats or problems in the different phases of fitting a statistical model to sample data

1. **Data generating process**
   - Degree of class separation
   - Between-class variance structure
   - True number of latent groups
2. **Model specification** fitted to data.
3. **Model estimation** options

Three-Phase Hierarchical Framework

Systematically investigates three kinds of threats or problems in the different phases of fitting a statistical model to sample data

1. **Data generating process**
   - Degree of class separation
   - Between-class variance structure
   - True number of latent groups
2. **Model specification** fitted to data.
   - Imposed mean structure
   - Chosen number of latent groupings
   - Variance structure imposed between groups
3. **Model estimation** options

Three-Phase Hierarchical Framework

Systematically investigates three kinds of threats or problems in the different phases of fitting a statistical model to sample data

1. **Data generating process**
   - Degree of class separation
   - Between-class variance structure
   - True number of latent groups
2. **Model specification** fitted to data.
   - Imposed mean structure
   - Chosen number of latent groupings
   - Variance structure imposed between groups
3. **Model estimation** options
   - Number of sets of random starting values
   - Inequality constraints on variances
   - Method used to identify best number of latent groups

Three-Phase Hierarchical Framework (cont.)

Each successive phase represents an overarching design structure on all prior phases.

Phase 1		**Phase 2**		**Phase 3**
Data generation	$\Longrightarrow$	Model specification	$\Longrightarrow$	Estimation options

Within each phase, the meta-model enables investigation of:

- Multiple determinants of the statistical model's performance
- Interactions among determinants

Between different phases, the meta-model enables investigation of:

- Direct effects on performance by each phase
- Interactions among determinants.across phases.

Three-Phase Hierarchical Framework (cont.)

Differentiate each phase on at least five dimensions:

- Researchers' level of control
- Number of plausible factors for investigation
- Appropriate kind of experimental design
- Level of validity threats
- Software specificity

Differentiating Phases of the Framework

1. The **data generating process**
   - Outside the control of researchers
   - Large number of potentially relevant factors
   - Fractional factorial design most appropriate
   - Primarily external validity issues
   - Software independent

Differentiating Phases of the Framework

1. The **data generating process**
   - Outside the control of researchers
   - Large number of potentially relevant factors
   - Fractional factorial design most appropriate
   - Primarily external validity issues
   - Software independent

2. **Model specification** fitted to data
   - Partly under the control of researchers
   - Reduced number of relevant factors
   - Factorial design
   - Internal validity issues
   - Software independent

Differentiating Phases of the Framework

1. The **data generating process**
   - Outside the control of researchers
   - Large number of potentially relevant factors
   - Fractional factorial design most appropriate
   - Primarily external validity issues
   - Software independent

2. **Model specification** fitted to data
   - Partly under the control of researchers
   - Reduced number of relevant factors
   - Factorial design
   - Internal validity issues
   - Software independent

3. **Model estimation** options
   - Under the control of researchers
   - Moderate number of relevant factors.
   - Factorial design
   - Internal validity issues
   - Software dependent

The Latent Growth Mixture Model

Latent growth mixture models (LGMMs) extend the standard latent grown model (LGM) by assuming latent heterogeneity, such that observed scores $\boldsymbol{y}_i$ arise from K discrete unobserved groupings (i.e., latent classes).

LGMMs hypothesise that the multivariate normal density $f(\boldsymbol{y}_i)$ comprises a finite mixture of K component normal distributions

$$f(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k \left[\boldsymbol{y}_i; \boldsymbol{\alpha}_k(\boldsymbol{\theta}_k), \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k)\right]$$

Restrictive structures imposed on $\boldsymbol{\alpha}_k$ and $\boldsymbol{\Sigma}_k$, such that:

$$\begin{aligned} \boldsymbol{\alpha}_k(\boldsymbol{\theta}_k) &= \boldsymbol{\Lambda}_k \boldsymbol{\alpha}_k \\ \boldsymbol{\Sigma}_k(\boldsymbol{\theta}_k) &= \boldsymbol{\Lambda}_k \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k \end{aligned}$$

and π_k is the unconditional probability of membership in the k^{th} latent class, with $\sum_{k=1}^{K} \pi_k = 1$.

Latent Growth Mixture Models

Meaningful use of LGMMs requires having sufficient confidence about four inter-related problems in these models...

1. The **absence of a local solution** when fitting the model to data.
2. Identifying the **correct number of latent groupings**.
3. **Accuracy of parameter estimates**.
4. **Accuracy in classification** of individuals to their latent grouping.

Non-trivial problems, for which conditions influencing failure of occurrence are not well understood.

Failure in any one of these four *may* (or *will*) obviously affect one or more others.

Illustrative Application of Framework to LGMMs

A Monte Carlo study investigating plausible explanations of local solutions in LGMMs.

The focal outcome measure is *number of local solutions* from multiple sets of random starting values used to estimate LGMMs.

- Local solution is where, for a given set of random start values for model parameters, ML estimation converges to a lower value than that obtained from a different set of random starting values

The independent variables in the meta-model are the various factors being manipulated within and across the three phases of the Monte Carlo simulation study.

Illustrative Application of Framework to LGMMs (cont.)

The frequency of local solutions may be influences by:

1. Particular aspects the data when fitting a LGMM
2. Particular kinds of misspecification in the LGMM itself
3. Particular choices of model estimation options

The meta-model therefore reflect the following experimental designs at each phase:

1. **[Phase 1]**: A 2^{9-2}_{VI} fractional factorial design
2. **[Phase 2]**: A 3-way fully-cross factorial design
3. **[Phase 3]**: A one-way between-subjects design

Phase 1: Design for *Data Generating Process*

The 9 factors, and their levels, in 2^{9-2}_{VI} fractional factorial design were:

[A] Between-Class Variance Structure of Growth Factors

- $(-)$ *Heterogeneous*
- $(+)$ *Homogeneous*

[B] Class Separation (Steinley & Hanson, 2004)

- $(-)$ *Close* (p = 0.35 for joint class overlap)
- $(+)$ *Wide* (p = 0.05 for joint class overlap)

[C] Number of Latent Classes

- $(-)$ *K = 3*
- $(+)$ *K = 4*

Phase 1: Design for *Data Generating Process* (cont.)

[D] Sample Size

(−) *N = 300*
(+) *N = 1000*

[E] Class Probabilities

(−) *1 Big* ([0.71 0.16 0.13] or [.52 .19 .16 .13])
(+) *2 Big* ([0.45 0.42 0.13] or [.37 .34 .16 .13])

[F] Reliability of Outcome Measures

(−) *Low* (α = 0.90)
(+) *High* (α = 0.70)

[G] Growth Process

(−) *Linear* growth in all classes
(+) *Quadratic* growth in last two classes

Phase 1: Design for *Data Generating Process* (cont.)

[H] Alignment of Classes & Probabilities

- (−) *Positive* (largest prob. in first class)
- (+) *Negative* (largest prob. in last class)

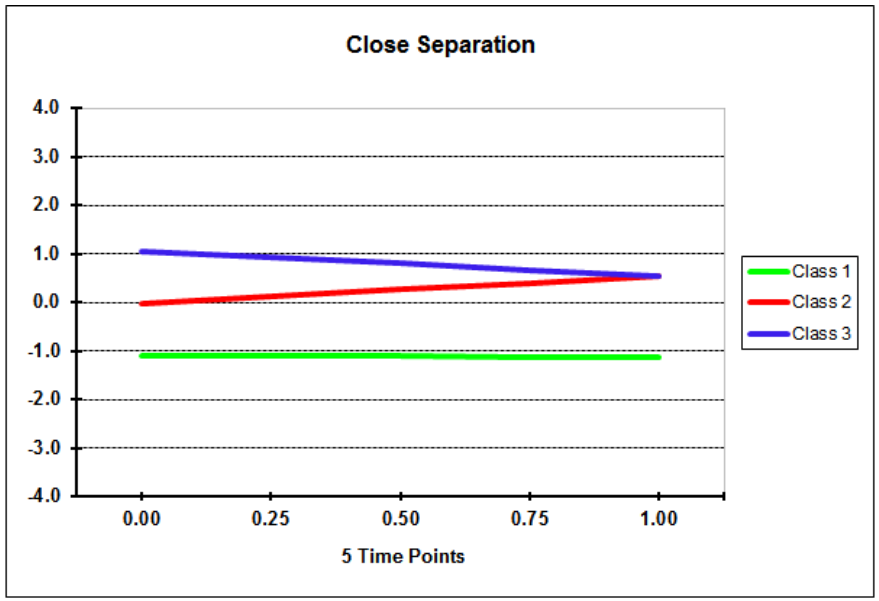
[J] Missing Data

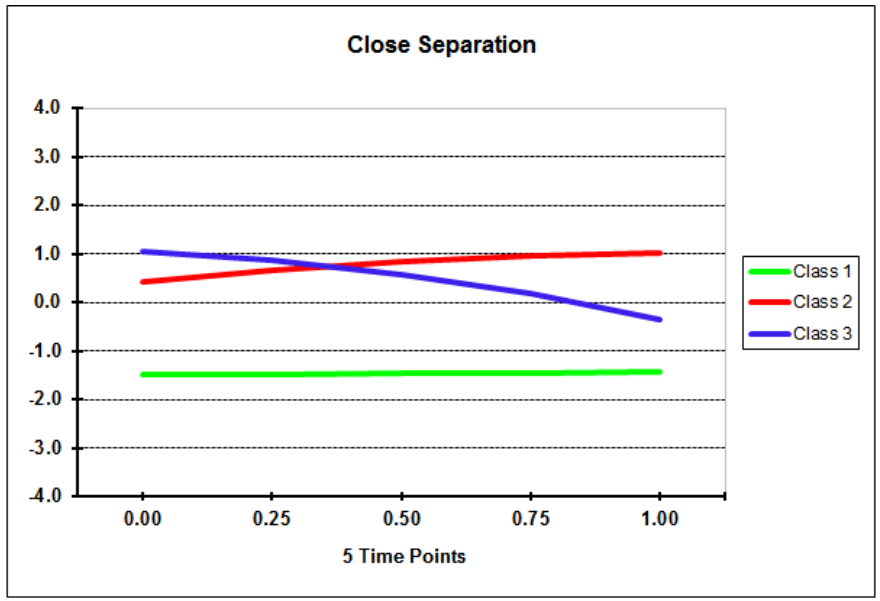
- (−) *None*
- (+) *MAR* ([100% 95% 90% 85% 80%])

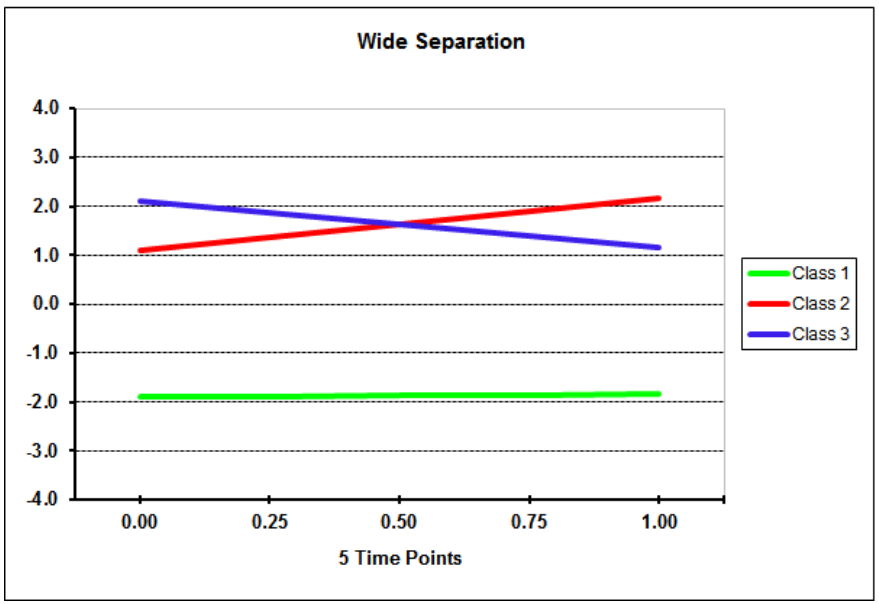
Notes:

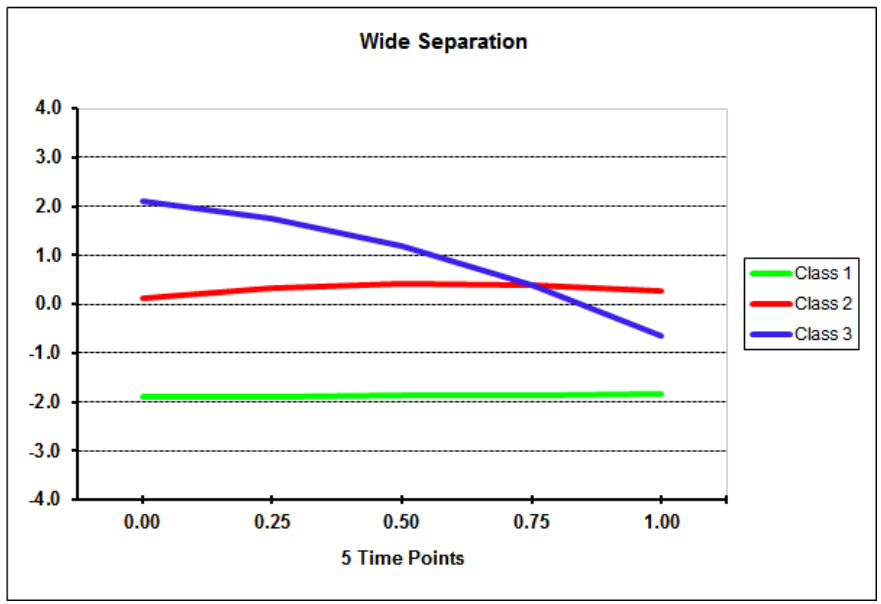
- Number of time points always equalled 5.
- All factors, except *Reliability* and *Missing Data*, were informed by random sample of 40 published LGMM studies.
- Linear growth factor variance set at 25% of intercept factor variance.
- Variance of quadratic growth factor was set at 0.
- Data were generated by "building from inside out" to maintain correct class separation (i.e., number of classes → growth structure → variance structure → reliability)

Mean Growth Trajectories Used in Study

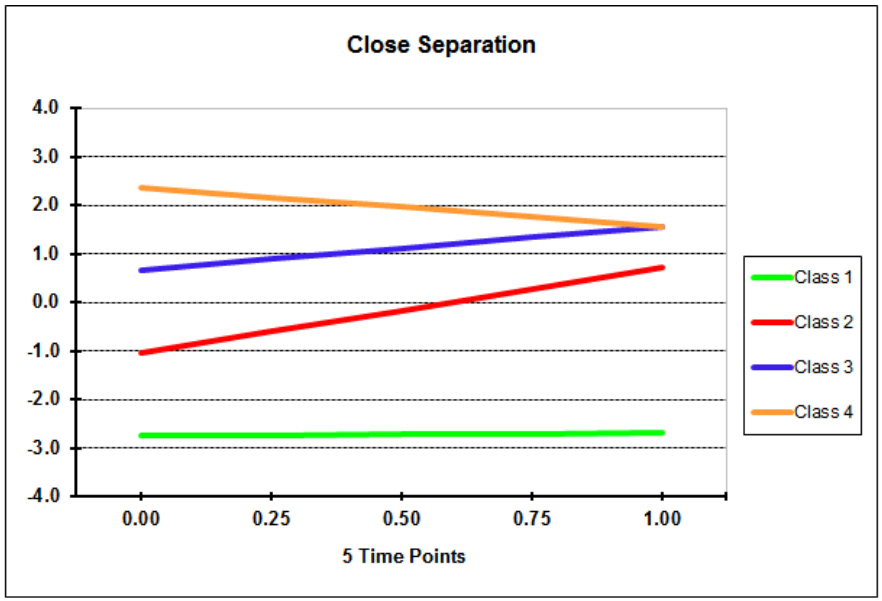


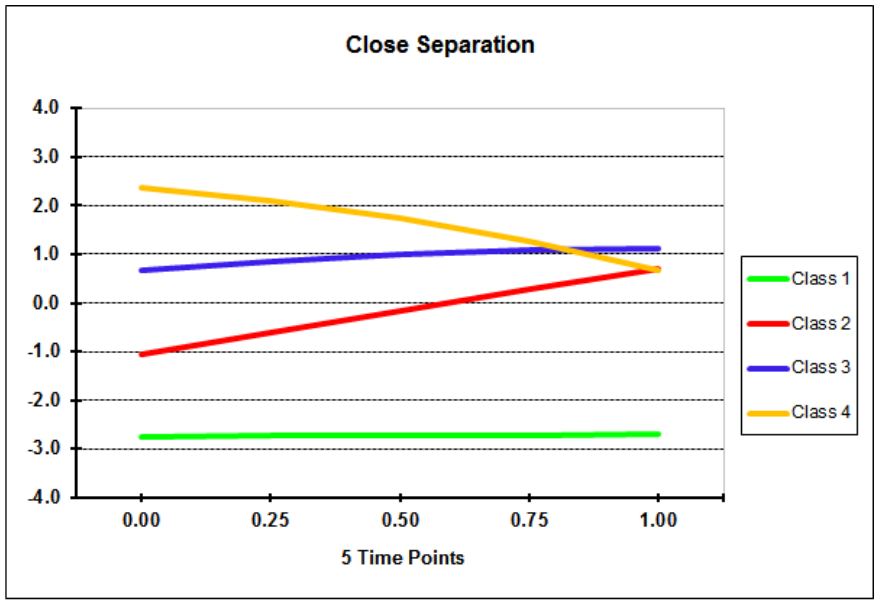


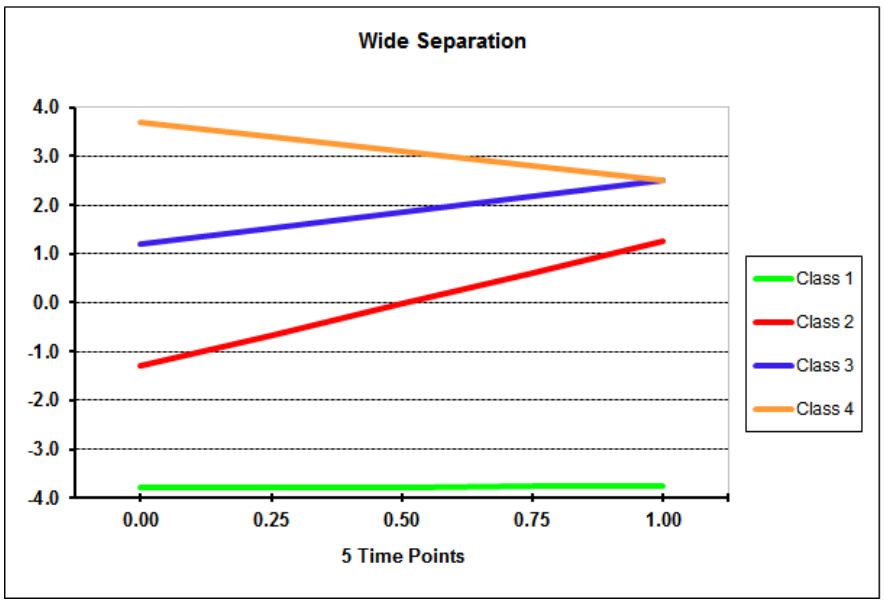


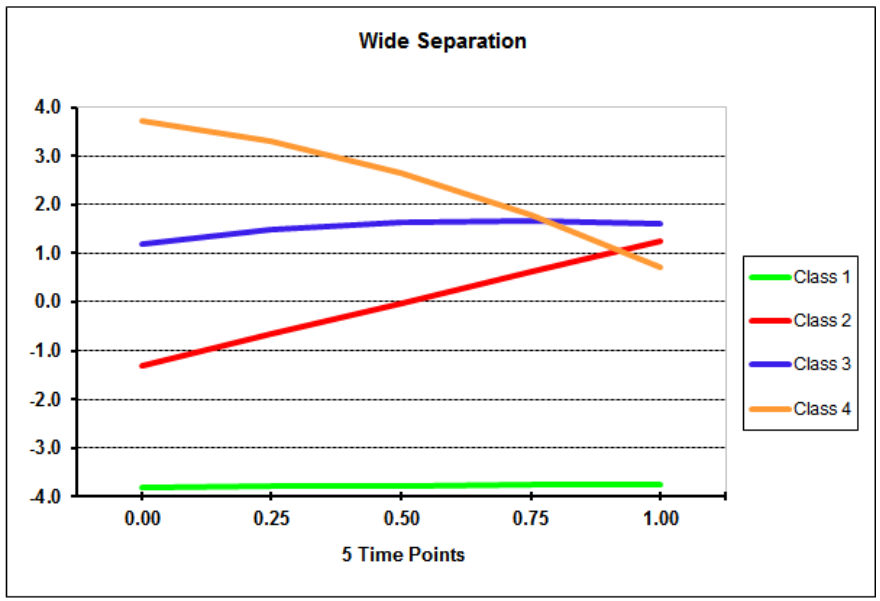


Mean Growth Trajectories Used in Study (cont.)









Some More Definitions

Run

A particular combination of levels of all factors in the fractional factorial design used in the data generating conditions in Phase 1. Skrondal (2000) refers to this as a *treatment*.

Profile

A particular combination of levels of the factors in the full factorial designs used in both the model specification (Phase 2) and model estimation (Phase 3).

Replication

A single, simulated data set in a particular run. The number of runs $\times$ the number of profiles $\times$ the number of replications equals the total number of simulated data sets generated for the Monte Carlo investigation.

Resolution VI Fractional Factorial Designs

For any Resolution VI fractional factorial design:

- All main effects are independent
- All 2-way interactions are independent
- Some 3-way interactions aliased with subset of other 3-way interactions
- Contains 128 runs (i.e., different combinations of the 9 factors)
- Represents 1/4 fraction of a full factorial design containing $2^9 = 512$ runs.
- Takes advantage of assuming the *sparsity-of-effects* principle

Phase 2: Design for *Model Specification*

Phase 2 used a ($2 \times 3 \times 3$) fully-crossed factorial design to investigate model specification threats.

[P] Model-imposed Variance Structure

(1) *Correct for data generating process*
(2) *Misspecified*

[Q] Model-imposed Growth Process Structure

(1) *Correct for data generating process*
(2) *Misspecified linear*
(3) *Misspecified quadratic*

[R] Model-imposed Number of Latent Classes

(1) *Correct for response process*
(2) *Underspecified*
(3) *Overspecified*

Phase 2: Design for *Model Specification* (cont.)

These 18 profiles represent different model specifications compared to the true data generating process defined by the FFD in Phase 1.

E.g., Data may have been generated in Phase 1 in one run by:

[A] Heterogenous variance structure

[B] Close class separation

[C] 3 latent classes

[D] $N = 300$

[E] 1 big class probability

[F] Low reliability

[G] Linear growth process

[H] Largest class prob. aligned with 1st class

[J] No missing data

But a profile in Phase 2 may have instead fitted a model containing:

- Homogenous variance structure (misspecified)
- 4 latent classes (overspecified)
- Linear growth process (correctly specified)

Phase 3: Design for *Model Estimation Options*

Phase 3 used a 2-level one-way design to investigate potential threats from one kind of model estimation option.

[V] Inequality Constraints on Variance Structure

- $(-)$ *No*.
- $(+)$ *Yes*.

Rationale being that one potential reason for the occurrence of local solutions is that parameter values may become inadmissible during model estimation when using poor standing values.

In summary, there were $18 \times 2 = 36$ different model specifications being fitted in the study.

Data Set Generation and Model Fitting

For each of the 18 profiles in the $2 \times 3 \times 3$ factorial design in Phase 2:

- Separate data sets generated according to the 2^{9-2}_{VI} FFD
- 10 replications generated in each run

Implies that $10 \times 128 \times 18 = 23040$ data sets generated.

- Each run within each profile used a different random seed to simulate the data
- Data generation was done separately in Phase 1 to fitting each of the model specification profiles in Phases 2 and 3

In Phase 3, dealing with the inequality constraints on variance parameters, the same 23040 data sets were used in each condition.

Model Estimation Set-Up

Details about the fitting of models is as follows:

- All data generation and model fitting was done in *M*plus version 7.0
- 1000 initial sets of random starring values used
- Best 101 from the initial set were then estimated to convergence
  - STARTS = 1000 101;
- All default options for mixture modelling in *M*plus were used, except the scaling of the random staring values was reduced from 10 to 5.
  - STSCALE = 5.0;
- The population parameter values were *not* used as default starting values in the model specification.
- Generation of *M*plus syntax files and parsing of output files done using MATLAB scripts.

Model Estimation Set-Up (cont.)

The outcome measure (Y) was the number of local solutions obtained out of the best 100 random starting values.

- $Y = 0$: the highest maximum likelihood value was replicated from *every* set of starting values
- $Y > 0$: the highest maximum likelihood value was replicated on a partial set of starting values
- $Y = 100$: the highest maximum likelihood value was *never* replicated from any set of starting values

Finally, Y was dichotomized to form a final outcome measure on which a logistic regression model was fitted using the factors from Phase 1 and Phase 2 to investigate replication of the maximum likelihood value.

- $1 = 0 < Y < 50$
- $2 = 51 < Y < 50$

Presentation of Results

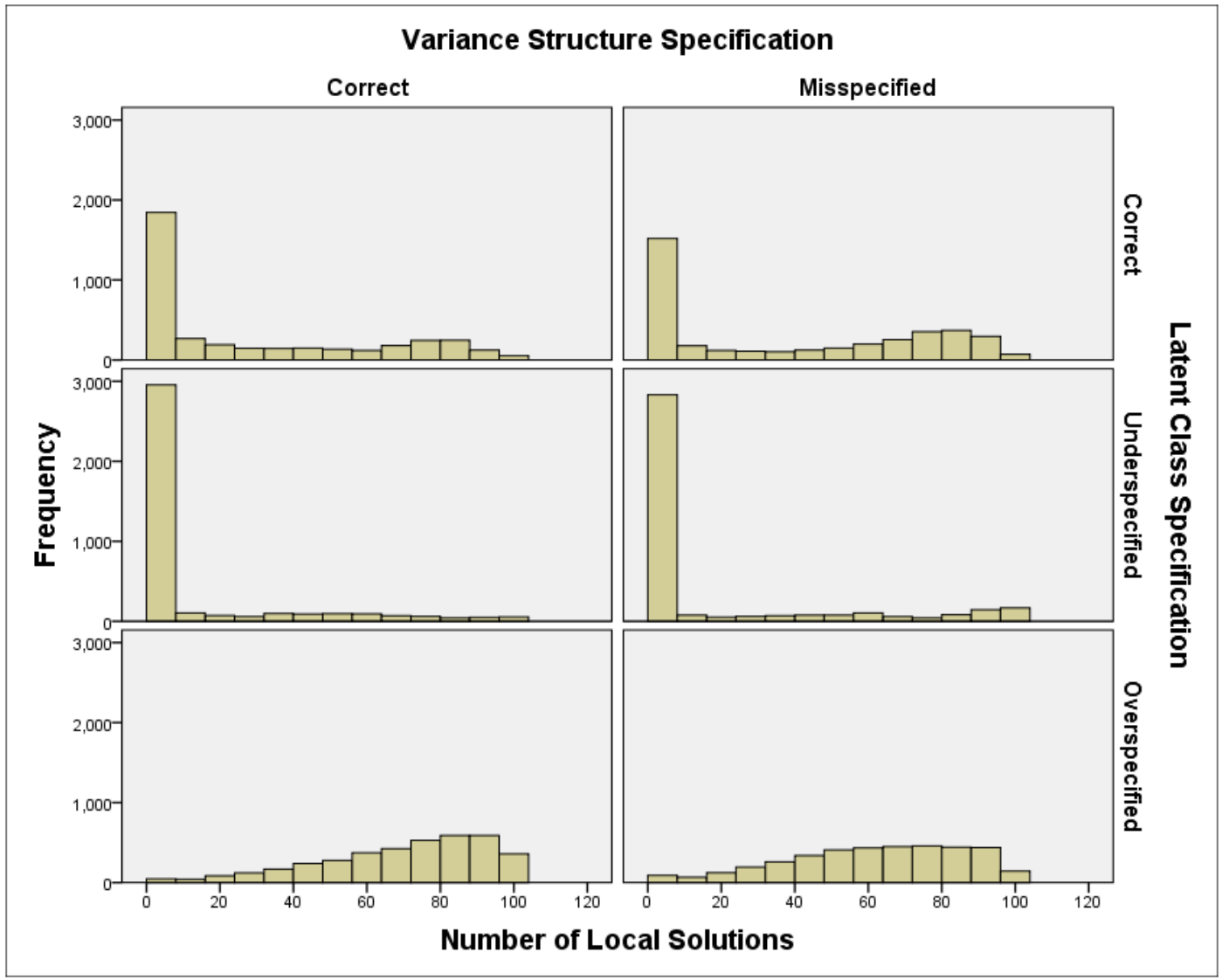
Results will be considered in two stages:

1. Initially trellis histograms of Y for different profiles in Phases 2 and 3
   - Broken down for (i) variance misspecification and (ii) growth misspecification within each of number of latent classes specified in Phase 2
   - Separately for the two conditions in Phase 3

2. Hierarchical logistic regression meta-model for dichotomized Y separately for variance constraints condition in Phase 3.
   - Main effects for data generating conditions (Phase 1)
   - Interaction effects for data generating conditions (Phase 1)
   - Main effects for model specification conditions (Phase 2)
   - Interaction effects for model specification conditions (Phase 2)
   - Interaction effects between Phase 1 and Phase 2 conditions

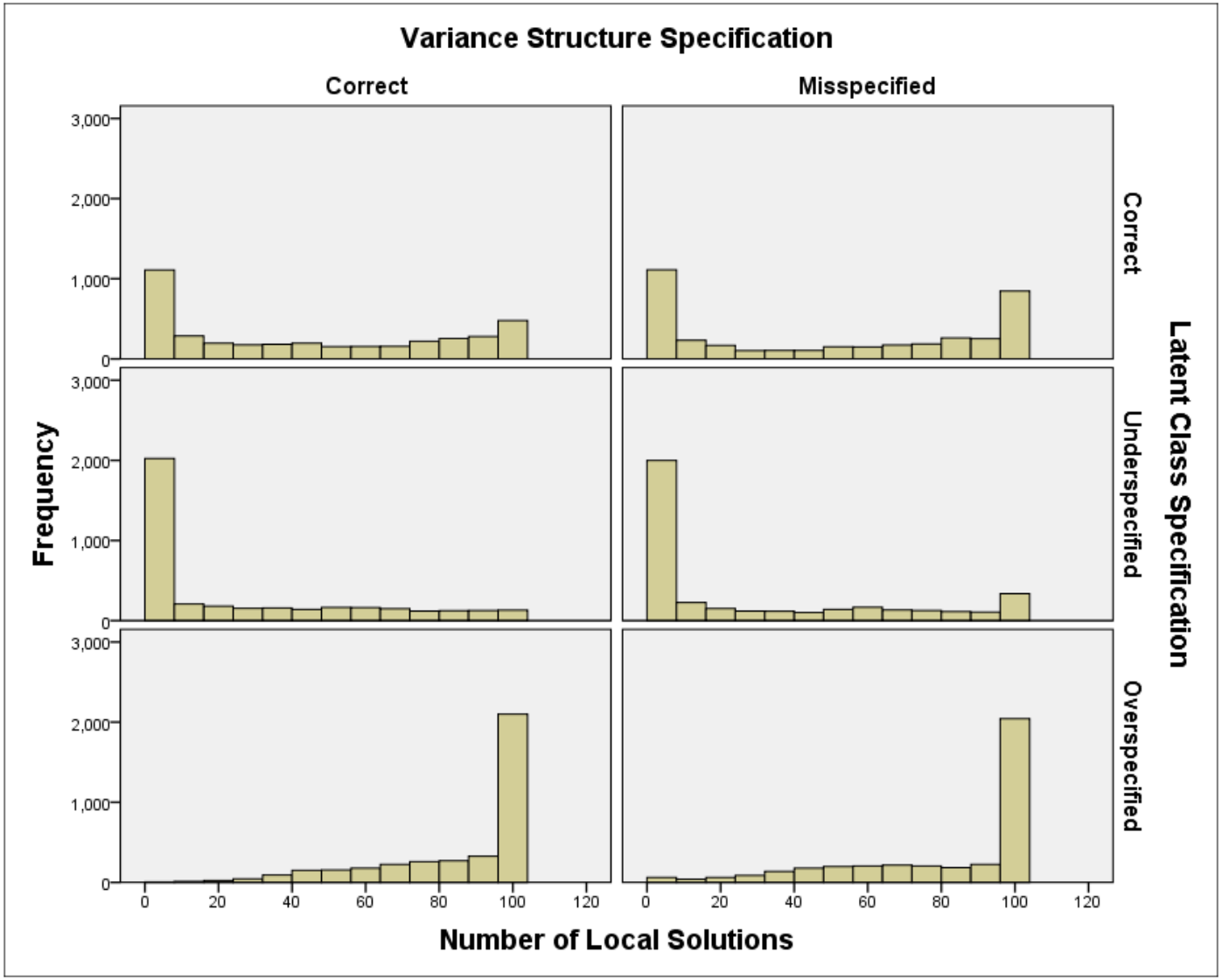
Phase 3: Model Estimation Options

No Constraints on Variance Parameters



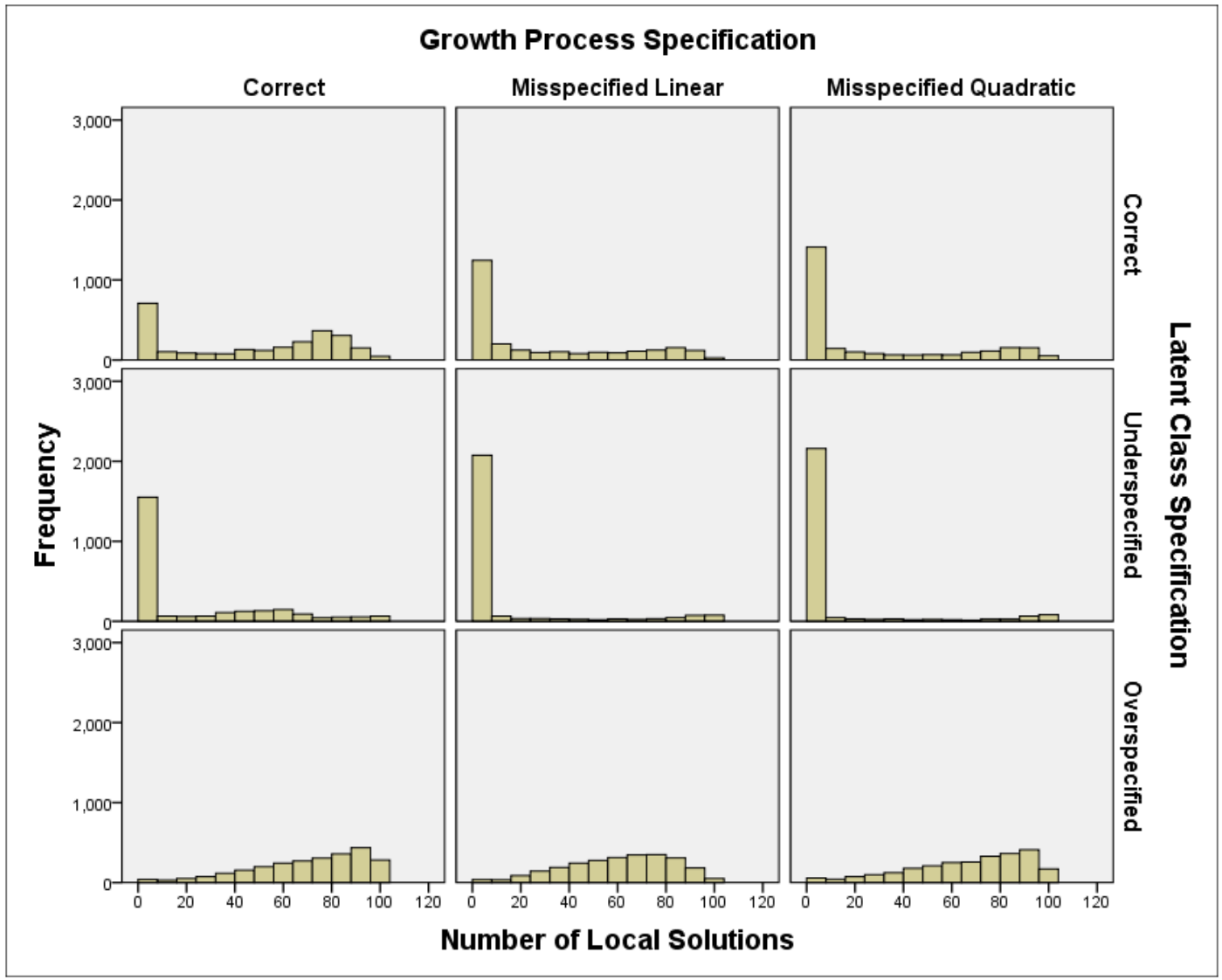
Phase 3: Model Estimation Options (cont.)

Inequality Constraints on Variance Parameters



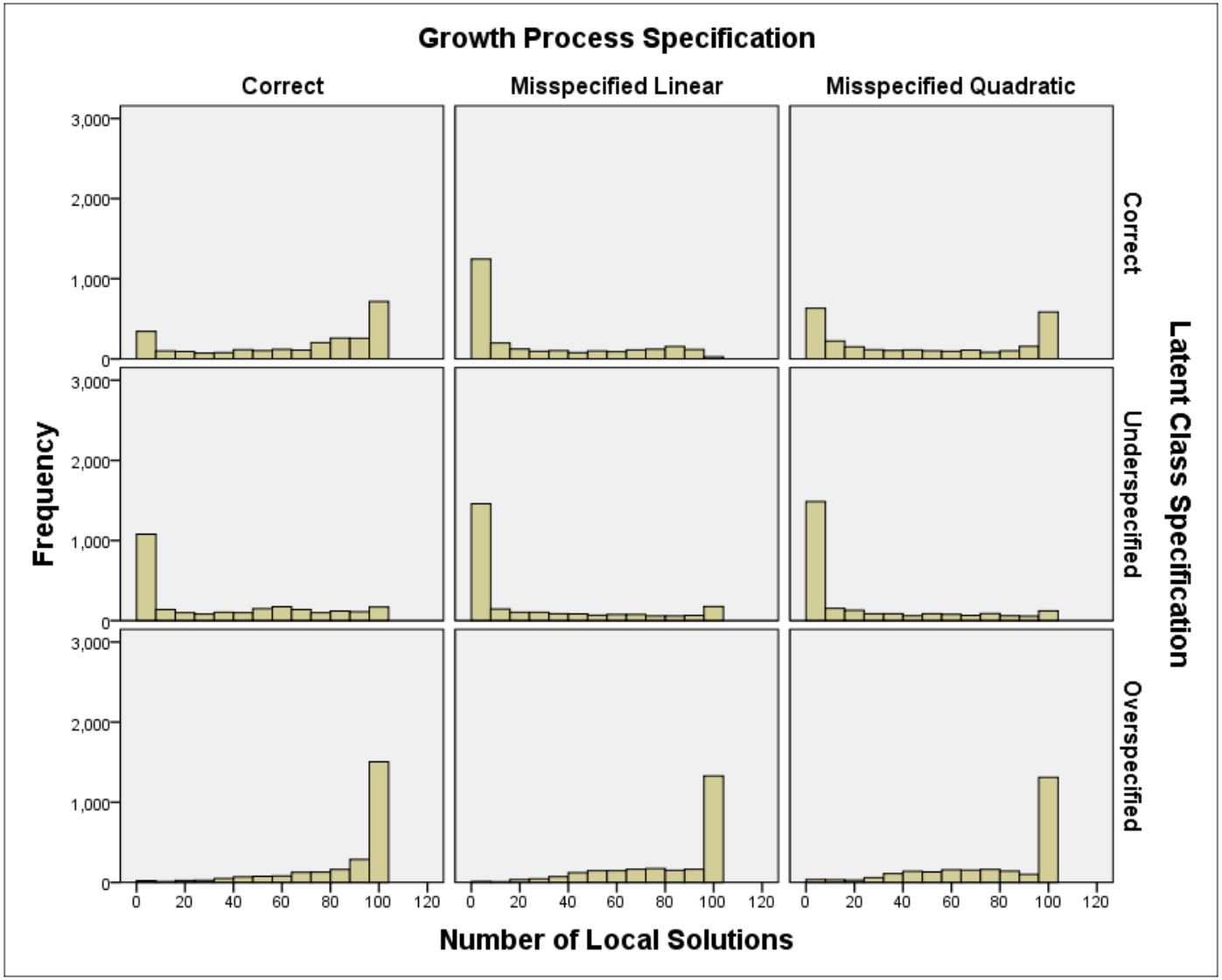
Phase 3: Model Estimation Options (cont.)

No Constraints on Variance Parameters



Phase 3: Model Estimation Options (cont.)

Inequality Constraints on Variance Parameters



Logistic Regression Meta-Model Results

Stage	*Effects Entered*	*df*	Inequality Constraints	
			Yes	*No*
Phase 1	Main Effects	9	1121	1161
	2-way Interactions	36	681	492
	3-way interactions	55	130	111
Phase 2	Main Effects	5	8209	10665
	2-way Interactions	6	272	231
	3-way Interactions	6	54	26
Phase 1 × 2	2-way Interactions	45	8809	6629

Top 10 Effects for Unconstrained Variance Parameters

By Phase:

1. Homogenous $\times$ Variance Misspecified
2. Quadratic Growth $\times$ Misspecified Linear
3. Quadratic Growth $\times$ Misspecified Quadratic
4. Overspecified No. Classes
5. Quadratic Growth
6. Homogenous Classes $\times$ High Reliability
7. Underspecified No. Classes
8. Large Sample Size
9. Misspecified Variance $\times$ Overspecified No. Classes
10. Wide Class Separation $\times$ Overspecified No. Classes

Note: *Blue* indicates Phase 1 condition.

Top 10 Effects with Unconstrained Variance Parameters (cont.)

By Direction of Effect:

1. Homogenous $\times$ Variance Misspecified
2. Quadratic Growth $\times$ Misspecified Linear
3. Quadratic Growth $\times$ Misspecified Quadratic
4. Overspecified No. Classes
5. Quadratic Growth
6. Homogenous Classes $\times$ High Reliability
7. Underspecified No. Classes
8. Large Sample Size
9. Misspecified Variance $\times$ Overspecified No. Classes
10. Wide Class Separation $\times$ Overspecified No. Classes

Note: *Red* indicates increase in probability of local solutions.

Some Observations from the Omnibus Results and Top 10

- Complex set of conditions lead to chance of local solutions in estimating LGMMs
- Model specification, especially, getting the number of latent groups correct, is critical
- Depends upon both the particular data generating process and whether or not the LGMM is correctly specified
- Getting the variance structure correctly specified is most important individually
- However, getting the growth process correctly specified interactions more often with other conditions
- Interactions more common than main effects

More Generally...

- The framework enables a more focused investigation that delineates competing threats to the use of LGMMs
- Identifies which particular conditions in both the data and the model specification are most important
- At the same time, it can identify many conditions in both phases that are not important
  - Extends the external validity... reduced "limitations to this research were..."
  - Though should state these findings may not necessarily reflect results using LatentGOLD :)
- Potentially extend Phase 3 to compare smaller number of random starts
- Framework is focused, systematic, broad in scope, intensive—yet tractable

Finally...

Thank You for Listening.