Modeling and Simulation in Social Sciences

Georgiy Bobashev, Ph.D.

Dec. 14th, 2016

3040 Cornwallis Road ■ P.O. Box 12194 ■ Research Triangle Park, North Carolina, USA 27709
Phone 919-541-6167 Fax 919-541-6722 e-mail bobashev@rti.org

Elements of a Model

ODD Protocol

1. Why it is critical to follow the protocol
2. ODD
    1. Overview
    2. Design Concepts
    3. Details
3. Other rules and good practices of modeling

Complex Systems and Protocols

"Flying Fortress" B-17 the largest and the most sophisticated aircraft of the time based on the Boeing Model 299.

On 30 October 1935, Army Air Corps test-pilot Major Ploer Peter Hill and Boeing employee Les Tower took the aircraft on a second evaluation flight; however, the crew forgot to disengage the airplane's "gust lock," a device that held the bomber's movable control surfaces in place while the aircraft was parked on the ground. Having taken off, the aircraft entered a steep climb, stalled, nosed over and crashed, killing both, Hill and Tower.

Complex Systems and Protocols

Checklist

ODD "Standard" Protocol
(Grimm et al. 2006, Railsback and Grimm, 2011)

	Purpose
Overview	Entities, state variables and scales
	Process overview and scheduling
	Emergence
	Objectives
	Adaptation
Design	Learning
concepts	Prediction
	Sensing
	Interaction
	Stochasticity
	Collectives
	Observation
	Initialization
Detail	Input data
	Submodels

Purpose

- **Why are we developing the model?**
- Clear and concise statement of the question or problem **addressed by the model**
- It is better to get an approximate answer to the right question than a precise answer to the wrong question
- Refer to the objectives when you need to:
  - Know when to stop
  - What should be and what should not be in the model
  - Is my model reasonable?

Entities, State variables, and Scales

- **What are the model components?**
- "A model should be as simple as possible but not simpler " A. Einstein
- Agents (**Who?**)
- Environments **(Where?)**
- Major variables (global and state variables) (**What?**)
- Major rules (**How?**)

Entities, State variables, and Scales

- Temporal and spatial scales (discrete and continuous)
- Populations and boundaries (time, space, population)

Process Overview and Scheduling

- **How are the components functioning**?
- Model building process (top-down vs. bottom up?)
- Empirical vs. theoretical
- Linear vs. non-linear
- What are agents doing? (e.g., searching, mating, fighting)

Process Overview and Scheduling

- How is environment changing? (e.g., providing resource, restricting activities)
- What is the observer doing? (e.g., collects the data)

Except for very simple schedules, one should use pseudo-code to describe the schedule in every detail, so that the model can be re-implemented from this code. Ideally, the pseudo-code corresponds fully to the actual code used in the program implementing the ABM.

Time and Synchronization

- Discrete
  - Synchronous sequential
  - Synchronous simultaneous
  - Asynchronous sequential
- Event-driven continuous time
- Time scales

Design concepts

Design Concepts

Emergence
What emerges from the model (rather than being imposed)?

Adaptation
How do the agents adapt to improve their fitness? (Directly and indirectly)

Fitness
What are the goals of the agents? What determines their survival?

Prediction
How do agents predict the consequences of their decisions?
Use of learning, memory, environmental cues, embedded assumptions

Sensing
What are agents assumed to know or perceive when making decisions?
Is the sensing process itself explicitly modelled?

Interaction
What forms of interaction among agents are there?

Stochasticity
Justification for any stochasticity in the model

Collectives
Grouping of individuals

Observation
How are data collected from the model for analysis?

RTI INTERNATIONAL

Design Concepts

- **How the concepts are translated into a model?**
- How are the entities modeled?
- How are the rules modeled?
- What are the measures for the variables and parameters?
- How are interactions modeled?
- How are choices and decisions modeled?
- Inputs and outputs

Environments

- Natural environments (roads, buffers)
- Degradation and re-growth
- Sources of food, survival, etc.
- Spatial scales, multilevel models

Randomness

- Probabilistic decisions
- Random connections between agents
- Communication error
- What does and does not need to be randomized?
- Random activation order

Mathematical models

- First principles: Fundamental laws.
- First principles + data to estimate parameters
- Previous models: If it worked for somebody else, it might work for you, too.
- “Inverse engineering”

Functions

- Linear vs. non-linear
- Shapes and extremes
- Empirical vs. theory based

Observer

- What is monitored?
- What outputs are needed?
- Individuals vs. collectives?
- Graphics, tables, outputs
- Experiments

Example
Gusset et al. 2009

2.2.1. Purpose

The model was designed to predict the probability of small re-introduced populations of wild dogs establishing themselves and persisting in the release area under various scenarios, including regular translocation of disperser groups.

Example
Gusset et al. 2009

2.2.2. State variables and scales

The three entities included in the model were individuals, packs and disperser groups. Individuals were characterized by their state variables sex, age, social status and pack or disperser group membership. A pack was defined as a reproductive unit (either newly formed or established, see below) that contained a dominant pair, potentially also including pups as well as subordinate yearlings and adults of both sexes. Pups were less than one, yearlings between one and two, and adults more than 2 years of age. A disperser group consisted of one or more same-sexed individuals originating from the same pack. Time proceeded in discrete steps of 1 year. The model was not spatially explicit to make it more generally applicable and because disperser groups are highly mobile; however, space was indirectly included in the model by considering the ecological capacity for wild dogs in HiP (see below).

Example
Gusset et al. 2009

2.2.3. Process overview and scheduling

The fate of each individual in the population was traced from birth to death. Within each year, the following processes were simulated in the given (biologically meaningful and computationally practical) order for each of the given entities: ageing (individuals), reproduction (packs), dispersal (individuals), pack formation (disperser groups), mortality (individuals), catastrophes (individuals), management interventions (packs and disperser groups) and dominance (packs). Individuals, packs and disperser groups were processed in a randomized sequence every year. The rules defining the above processes are described in Section 2.2.7 below.

2.2.4. Design concepts

2.2.4.1. Emergence. Wild dog population and pack dynamics emerged from the behaviour of individuals, but individual behaviour was entirely imposed by probabilistic empirical rules. No Allee effects at the pack level were imposed onto the model, as no such effects were observed in the population modelled here (Somers et al., 2008). However, possible Allee effects were allowed to emerge from the model.

2.2.4.2. Interaction. Four types of interaction were modelled

Example
Gusset et al. 2009

2.2.5. Initialization

Simulations started with a specified number of packs and individuals per pack, but no disperser groups. One male and female per pack were randomly selected as dominants. Sex and age of individuals in initial packs was random: the probability of being male was 0.50 and age was uniformly distributed from 1 to 6 years.

2.2.6. Input

The model did not include any environmental variables as driving the population, as competitor density, amount of rainfall and prey availability did not significantly influence the population modelled here (Somers et al., 2008). Environmental variation was represented by environmental stochasticity and random catastrophic events.

Example

2.2.7. Submodels

2.2.7.1. Ageing. The age of all individuals increased by 1 year. All individuals that reached their observed maximum age of 9 years died (Somers et al., 2008).

2.2.7.2. Reproduction. Both males and females could theoretically become dominant and reproduce from 1 to 8 years of age, with only packs that contained a dominant pair potentially reproducing (Somers et al., 2008). The probability of a pack reproducing in a given year was piecewise density-dependent, which best matched the observed linear negative density dependence in population growth rate (Somers et al., 2008). HiP's ecological capacity for wild dogs, based on the availability of the most important prey species, was estimated to be at $N = 62$ (Lindsey et al., 2004), with N being the total number of all adults and yearlings plus half the number of pups. If N was smaller than half of the ecological capacity, a litter was added annually with an observed probability 0.33 to newly formed packs (i.e. in the first breeding season after formation)

Calibration and Analysis

Adapted from S.F Railsback and V. Grimm, *Agent-Based and Individual-Based Modeling. A practical introduction*. Princeton University Press, 2011

Calibration

- **Difference between**
  - ***Parameterization*** **(finding reasonable parameter values) and**
  - ***Calibration*** **(parameterization that reproduces specific behavior)**
- **Calibration could be used when we don't know what is the right parameter values are**

Calibration Strategies

- **Identify a few good calibration parameters (not too many to get lost)**
- **Choose categorical vs. "best-fit" calibration**
- **Transitions vs. Steady states, responses to shocks**
- **How to fit time series? (difference in mean, difference in variation, maximum error, mean squared error, pseudo R^2)**
- **Identify calibration criteria**
- **Experimental designs**

Useful Model Statistics

1. **Summary Statistics**

   Contrasting scenarios. Are we looking for the means or for the extreme values?

   95% trajectory bands

2. **Scenarios**

   Qualitative differences (e.g. population decay vs. sustainability)

   Quantitative differences

3. **Quantifying correlative relationships**

   Regression models on parameters, initial values.

4. **Comparison to empirical data qualitative and quantitative**

Uncertainty in Agent-Based Models

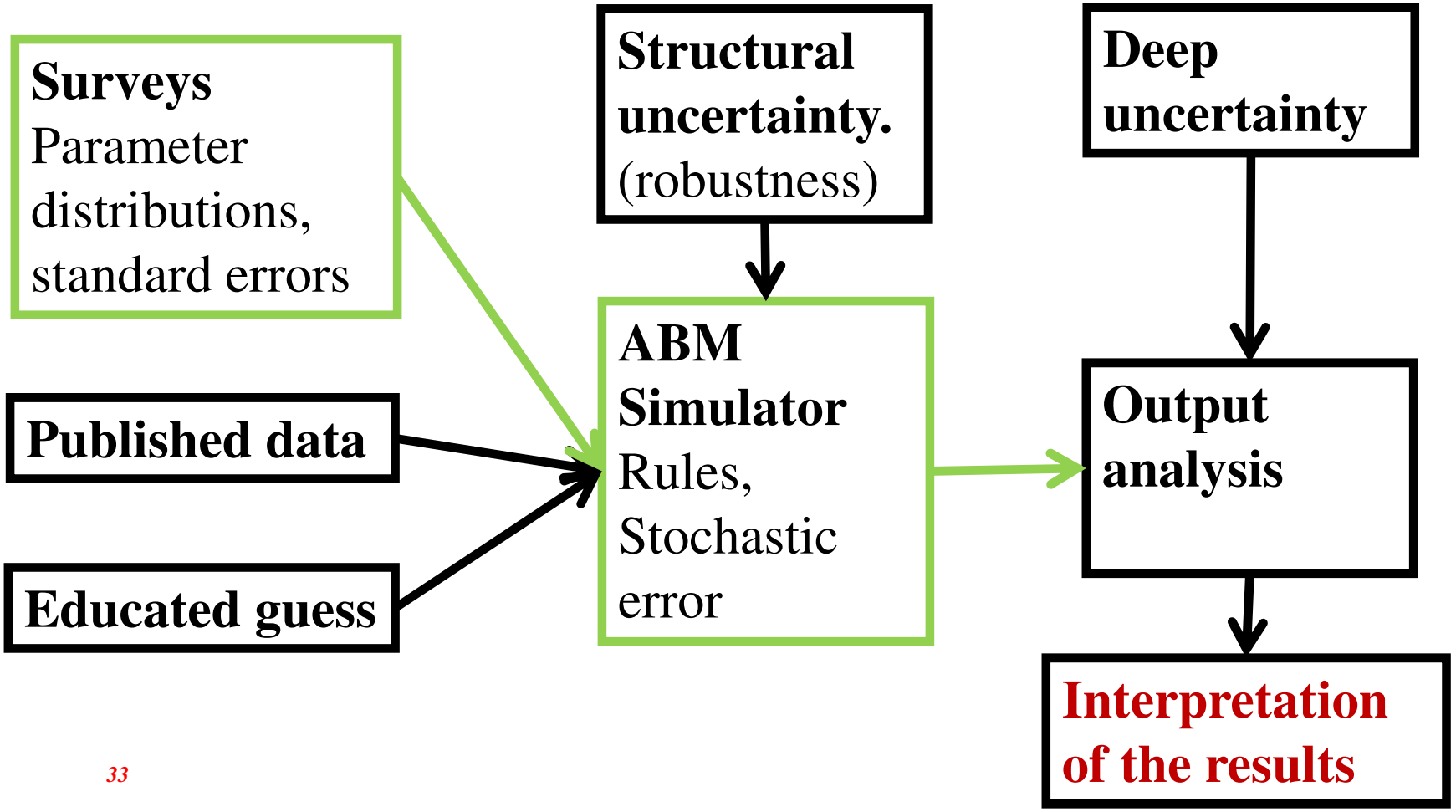
Sensitivity Uncertainty Robustness

1. **Sensitivity explores how sensitive the model to small change in parameters and initial conditions.**
2. **Uncertainty translates the uncertainty in parameters into reliability of model results**
3. **Robustness analysis explores the robustness of results to change in structure**

Multiple Parameters

- Sensitivity analysis in joint parameter space? Latin Hypercube
- Given the diverse nature of parameter values (continuous, categorical, ordinal) and the power of modern computing conduct global Monte-Carlo
- Independence vs. internal (model-based) correlation vs. external correlation

Sources of Uncertainty in ABMs



Example of an Epidemic

- Infectious disease in a population
- Proportion of infected individuals is θ
- Contact rate between individuals is λ
- Probability of infection given the contact is γ
- Then a probability of catching the disease by a susceptible individual is

$$P \approx \lambda\gamma\theta = \beta\theta$$

Epidemics Example

- Assume that the disease transmission rate β is estimated from a sample of 1000 individuals and has a mean of $\hat{\beta}$ with a standard error of S_{β}.

- Assume parameter distribution (Bayesian approach)

- Estimate for β has the mean of 0.1 per day and the standard error of 0.02

- Fix β = 0.1, time = 30 days. Assume homogeneous mixing. The exact solution gives the overall proportion of infected individuals is θ=0.075.

Epidemics Example

The overall proportion of infected individuals is $\theta=0.075$

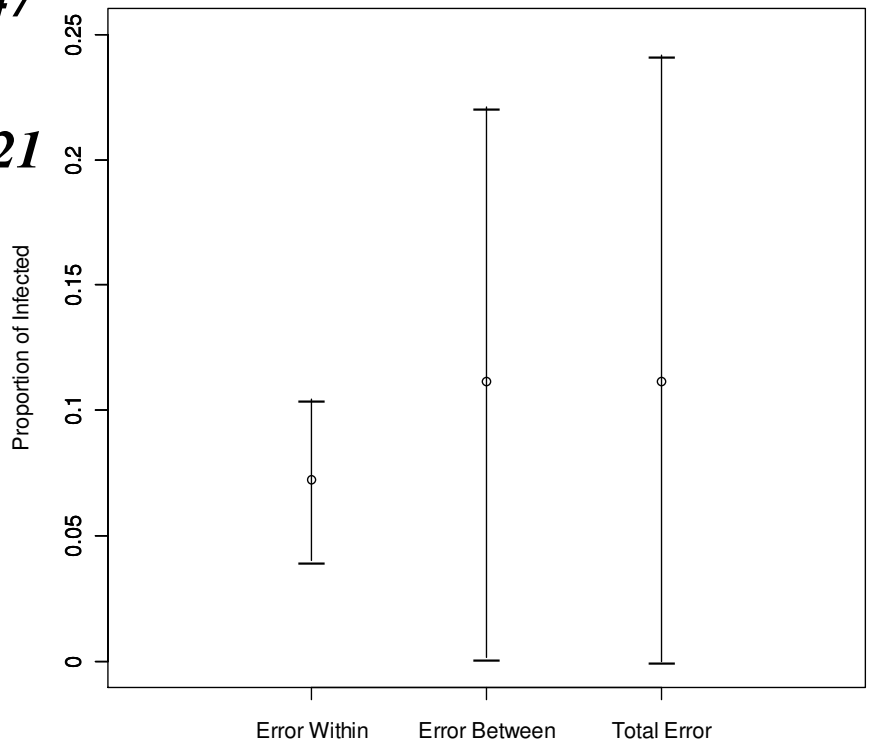
$Var_{over\ U}(E_{over\ j}\{\hat{\theta}_U \mid Y_{ij}, X_{ij}, U, n \})=0.0147$

$E_{over\ U}(Var_{over\ j}\{\hat{\theta}_j \mid Y_{ij}, X_{ij}, U, n\})=0.0021$

$Var_{total}=0.0168$

If it were a longitudinal study the total variance estimate would be

$Var(\hat{\theta})= p(1-p)/1000=0.000069$



Approaches to Dealing with Uncertainty

- Robust decisions under deep uncertainty and model simplification (Klein et al. 2010)
- Risk vs. Uncertainty (Ben Haim 2003, Yemshanov et al. 2010)
- Optimization under uncertainty (Marecki, 2010)
- Standard errors, p-values, and simulation stochasticity (Bobashev et al. 2010)
- Within-model vs. Within-system?
- Other approaches??

NetLogo Helpful Hints

- **Save program often**
- **Keep version control**
- **Find and fix mistakes as you go. Hierarchical compartmental structure helps**
- **NB.**
  - **Choosers switchers and sliders work only on global variables**
  - **When moving a variable to interface remove/comment it in the global variable declaration**
  - **Remove/comment it in the setup section**

Debugging the Program

Adapted from S.F Railsback and V. Grimm, *Agent-Based and Individual-Based Modeling. A practical introduction*. Princeton University Press, 2011

Debugging

- **Writing code and testing it is a simultaneous process. No program is bug free**
- **Common errors**
  - **Typographical errors**
  - **Copy and paste errors**
  - **Misunderstanding primitives**
  - **Wrong settings (e.g. wrap up vs. boundaries)**
  - **Runtime errors**
  - **Logic errors**
  - **Formulation errors**

Remedies

- Syntax checking
- Visual checking
- Print statements
- Agent monitor
- Stress test
- Using test programs
- Code reviews
- Statistical analysis

Issues in Proposals and Papers

Some Terms

- Internal testing (first-order validation, reliability, verification): are the equations programmed correctly, with no bugs?
- Internal (second-order) validation (calibration): are inputs and outputs consistent with data used to create the model?
- External (third-order) validation: does model match other data, not explicitly used to create the model?
- Cross-validation: does the model reach the same conclusions as other models?
- Predictive validity: does the model make accurate predictions about future events?
- Face validity: does it pass the smell test, intuition?

Proposal Arguments that are Commonly Attacked

" Inform policymakers about possible consequences of various interventions"

Who are the policymakers? How would one inform them?

"Inform the scientific community about the consequences of … "

Who are the scientific community? Who cares? Have anyone asked your opinion?

"Generate new hypotheses?"

We already have enough existing hypotheses that need to get tested before considering new questions.

"Increase understanding of…"

What is understanding? How does one measure whether we (better) understand a phenomenon or not?

Problems With Validation of ABMs in Social Science

Triple curse:

1. Statistical models

   Likelihood concept. P-value

2. System dynamics models in physics (known laws and measurements)

   Increased precision of measurements

3. Process models

   Processes are repeating many times, so one can get solid statistics

Are we looking for the mean, most likely outcome, or for the extreme values?

Common Sense Issues

- Has the model been validated *before* it was built?
- If a theoretical model follows common sense it is useless
- If the model produces an unusual outcome how do we know whether we deal with model artifact or discovery?
- Numeric values and uncertainty estimates. Do we need/care about uncertainty? If yes then what additional types if uncertainty we need to consider?

Practical Suggestions for Model Validation

- Objectives, design concepts, details: ODD (Grimm and Railsback)
- SME validation of the conceptual model and the results. Comparison to data is not always philosophically correct. (MacKerrow)
- Sensitivity analysis (Gilbert)
- Compare with observed data. ML approach to show plausibility (Whitney)
- Visual inspection. The pattern looks right (Epstein)
- Model accreditation (DoD)
- The criterion of truth is practice (Lenin)

Good Practices in Modeling

Three Commandments for Modelers

(S.P. Ellner and J. Guckenheimer, *Dynamic Models in Biology* 2006)

The principles of model development can be summarized as three important rules:

Lie

Cheat

Steal

Three Commandments for Modelers (continued)

Lie

- A good model can include incorrect assumptions. Practical models have to be simple enough that the number of parameters does not outstrip the available data. Theoretical models have to be simple enough that you can figure out what they're doing and why. The real world, unfortunately, lacks these properties. So in order to be useful, a model must ignore some known real life details, and replace these with simpler assumptions that sometimes are literally false.

Three Commandments for Modelers (continued)

Cheat

- Do things with data that would make a statistician nervous, such as using univariate data to fit a multivariate rate equation by multiplication of limiting factors. Sometime make choices based on the intuition. Statisticians like to let data “speak for themselves.” Often the data are only one input into decisions about model structure, the rest coming from modelers’ experience and subject-area knowledge.

Three Commandments for Modelers (continued)

Steal

- Take ideas from other modelers and models, regardless of discipline. Cutting-edge original science is often done with conventional kinds of models using conventional functional forms for rate equations. If somebody else has developed a sensible-looking model for a process that appears in your model, try it. If somebody else invested time and effort to estimate a parameter in a reasonable way, use it. Of course you need to be critical, and don't hesitate to throw out what you've stolen if it doesn't fit what you know about your system.

Before Modeling

- Identify objectives (prediction, concept, risk evaluation, decision)
- Identify model type (level of rules => SD, DE, AB)
- Lifetime of a model (will there be future development?)
- Exploratory, production, education, prototype, sharing?
- Budget, Timeline, Human resources
- Software

8 Modeling Stages

(Adapted form A. Law *Simulation Modeling and Analysis*, 2007)

1. **Formulate the problem.**
2. **Collect information and the data.**
3. **Write down and critically examine all the assumptions. Build a conceptual model. Design verification experiments for each of the assumptions. Design a calibration plan.**
4. **Program the model.**

8 Generic Modeling Stages (continued)

5. Verify the model.

6. Calibrate the model. Conduct sensitivity analysis.

7. Design and conduct validation experiments. Face validity. Do the results make sense?

8. Document and present the model. Document stages 5-7, especially sensitivity analysis.

Stage 4 usually takes about 20-25% of the whole project. Most work is done at stage 3.

Within-model Components of Variance for Simulation Results

Consider a simulated cohort of n subjects indexed by i
Consider a parameter estimate (e.g. odds ratio) for j'th realizations.
Assume the model is based on a parameter U with known distribution $F(U)$. Given the fixed value of input parameter U

$$\hat{\theta}_j = f(Y_{ij}, X_{ij}, U, n)$$

and the expectation over all stochastic realizations is

$$E_{over\ j}\left(\hat{\theta}_j \mid Y_{ij}, X_{ij}, U, n\right)$$

The expectation over all values of U is

$$E_{over\ U}\left(\hat{\theta}_U \mid Y_i(j), X_i(j), n\right)$$

Within-model Components of Variance for Simulation Results

The variance over all realizations for fixed U is

$$Var_{over\ j}\left(\hat{\theta}_j \mid Y_{ij}, X_{ij}, U, n\right)$$

The variance over U is

$$Var_{over\ U}\left(\hat{\theta}_U \mid Y_{ij}(j), X_{ij}(j), n\right)$$

The total variance Var_{total} is

$$Var_{over\ U}\left(E_{over\ j}\,\hat{\theta}_j \mid Y_{ij}, X_{ij}, U, n\right) + E_{over\ U}\left(Var_{over\ j}\,\hat{\theta}_j \mid Y_{ij}, X_{ij}, U, n\right)$$

$$Var_{over\ j}\left(E_{over\ U}\hat{\theta}_U \mid Y_{ij}, X_{ij}, U, n\right) + E_{over\ j}\left(Var_{over\ U}\,\hat{\theta}_U \mid Y_{ij}, X_{ij}, U, n\right)$$

Within-model Components of Variance for Simulation Results

If $Var_{over\ U}(Var_{over\ j}\hat{\theta}_j)$ **and** $Var_{over\ j}(Var_{over\ U}\hat{\theta}_j)$ **are small then the formula for the total variance can be simplified:**

$$Var_{total} \approx$$

$$Var_{over\ U}(\hat{\theta}_U \mid Y_i(j), X_i(j), n) + Var_{over\ j}(\hat{\theta}_j \mid Y_{ij}, X_{ij}, U, n)$$

For example, rather than running 10000 (100*100) simulations we can run only 200 (100+100)