Multilevel Models for Subjects Crossed with Items (i.e., *Explanatory* Item Response Theory or Latent Trait Models)

- Topics:
  - MLM as an alternative to multiple types of ANOVAs
  - IRT models as "items as fixed effects" MLMs
  - Explanatory IRT models as "items as random effects" MLMs

Prelude: The Hofflin Lego-Based View of Quantitative Methods

I want to illustrate how thinking this way has shaped my teaching and research for the better.

Big Picture Idea:
If you understand the elemental building blocks of statistical models, then you can build **anything**!

The 4 Lego Building Blocks

1. **Linear models** (for **answering questions** of prediction)

2. **Estimation** (for iterative ways of **finding the answers**)

3. **Link functions** (for predicting **any type of outcome**)

4. (a) **Random effects /**
   (b) **Latent traits / factors / variables**
   - (a) for modeling multivariate **"correlation/dependency"**
   - (b) for modeling relations of **"unobserved constructs"**

How the Blocks Fit Together

1. **Linear models** answer research questions, and are the first building block of every more complex analysis
   - *Is there an effect? Is this effect the same for everyone? Is the effect still there after considering something else?*

To add more blocks, we need iterative **estimation**

2. Maximum likelihood or Bayesian (e.g., MCMC)

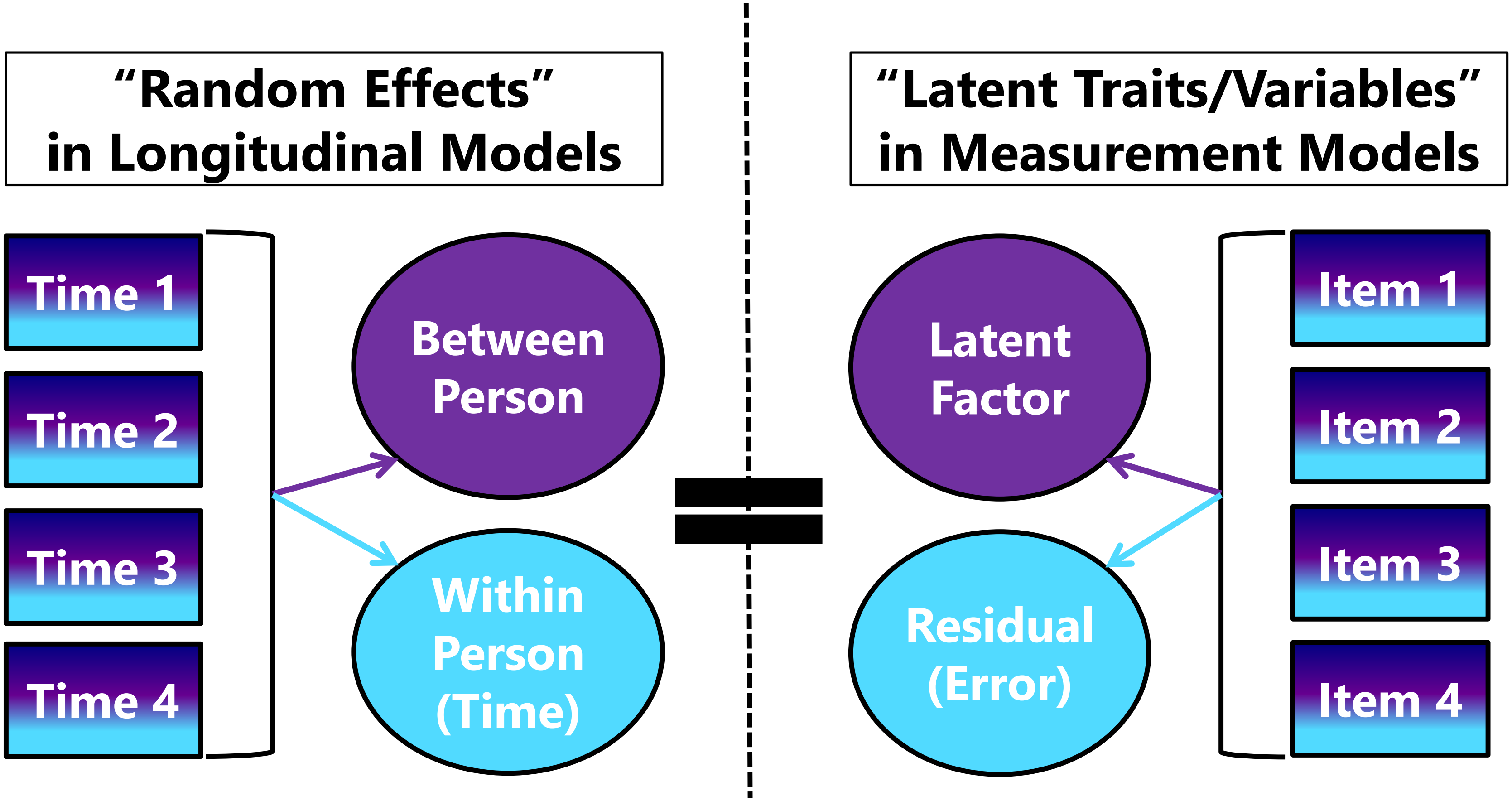
What other blocks you will need is determined by:

3. How your outcome is measured → link functions

4. Your dimensions of sampling → random/latent effects

From One to Many Outcomes…

- Most designs have more than one outcome per person…
  - e.g., multiple outcomes, occasions, items, trials … per person
  - Multiple dimensions of **sampling** → multiple kinds of **variability**



4. Random Effects / Latent Variables

- **Random effects** are for "handling dependency" that arises because multiple dimensions of sampling → multiple variances
  - Occasions within children (need 1+ random effect)
  - Children within classrooms within schools (need 2+ random effects)
  - *aka*, multilevel, mixed, or hierarchical linear models
- **Latent <traits/factors/variables>** are for representing "error-free true construct variance" within observed variables
  - Normal outcomes + latent variables = factor analysis (CFA; SEM)
  - Categorical outcomes + latent variables = item response theory (IRT)
- Random effects / latent variables are **mechanisms** by which:
  - Make best use of all the data; avoid list-wise deletion of incomplete data
  - Quantify and predict distinct sources of variation… *cue story-time*…

The Curse of Non-Exchangeable Items

Jim Bovaird, University of Nebraska-Lincoln

Larry Locker, Georgia Southern University

Downtown Lawrence, KS

- Psycholinguistic research (items are words and non-words)
  - Common persons, common items designs
  - Contentious fights with reviewers about adequacy of experimental control when using real words as stimuli
  - Long history of debate as to how data should be analyzed: **F1 ANOVA**, **F2 ANOVA**, **or both**?

Larry's Kinds of ANOVAs

Original Data per Person

	B1	B2
A1	Item 001 Item 002 .... Item 100	Item 101 Item 102 .... Item 200
A2	Item 201 Item 202 .... Item 300	Item 301 Item 302 .... Item 400

"F1" Within-Persons ANOVA on *N* persons:

$$\mathrm{RT_{cp}} = \gamma_0 + \gamma_1 \mathrm{A_c} + \gamma_2 \mathrm{B_c} + \gamma_3 \mathrm{A_c B_c} + \mathbf{U_{0p}} + \mathrm{e_{cp}}$$

"F2" Between-Items ANOVA on *I* items:

$$\mathrm{RT_i} = \gamma_0 + \gamma_1 \mathrm{A_i} + \gamma_2 \mathrm{B_i} + \gamma_3 \mathrm{A_i B_i} + \mathrm{e_i}$$

Person Summary Data

	B1	B2
A1	Mean (A1, B1)	Mean (A1, B2)
A2	Mean (A2, B1)	Mean (A2, B2)

Item Summary Data

	B1
A1, B1	Item 001 = Mean(Person 1, Person 2,... Person *N*) Item 002 = Mean(Person 1, Person 2,... Person *N*) .... Item 100
A1, B2	Item 101 = Mean(Person 1, Person 2,... Person *N*) Item 102 = Mean(Person 1, Person 2,... Person *N*) .... Item 200
A2, B1	Item 201 = Mean(Person 1, Person 2,... Person *N*) Item 202 = Mean(Person 1, Person 2,... Person *N*) .... Item 300
A2, B2	Item 301 = Mean(Person 1, Person 2,... Person *N*) Item 302 = Mean(Person 1, Person 2,... Person *N*) .... Item 400

Multilevel Models: A New Way of Life?

Original Data per Person

	B1	B2
A1	Item 001 Item 002 .... Item 100	Item 101 Item 102 .... Item 200
A2	Item 201 Item 202 .... Item 300	Item 301 Item 302 .... Item 400

Pros:

- Use all original data, not summaries
- Responses can be missing at random
- Can include continuous predictors

Cons:

- **Is still wrong (is ~F1 ANOVA)**

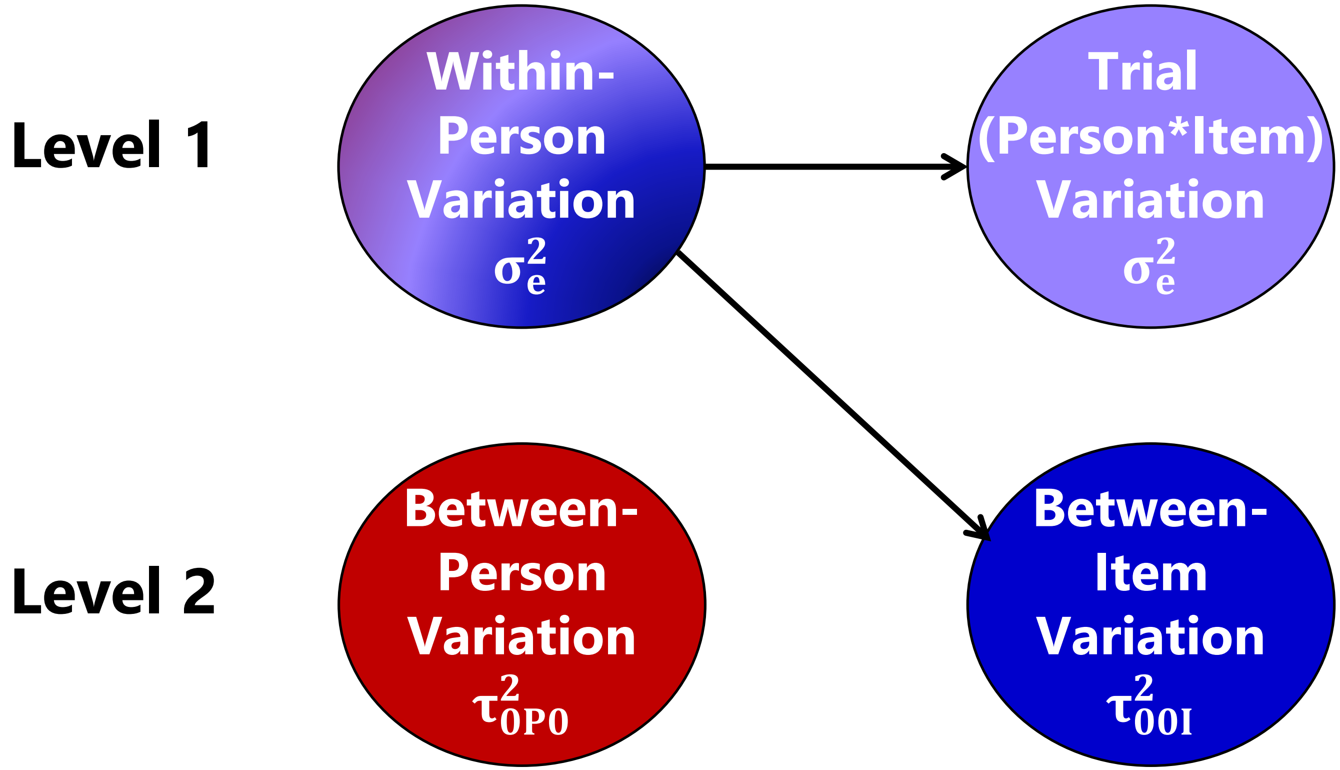
Level 1: $y_{ip} = \beta_{0p} + \beta_{1p}A_{ip} + \beta_{2p}B_{ip} + \beta_{3p}A_{ip}B_{ip} + e_{ip}$

Level 2:
$$\beta_{0p} = \gamma_{00} + U_{0p}$$
$$\beta_{1p} = \gamma_{10}$$
$$\beta_{2p} = \gamma_{20}$$
$$\beta_{3p} = \gamma_{30}$$

Level 1 = Within-Person Variation (Across Items)

Level 2 = Between-Person Variation

Multilevel Models: A New Way of Life?



A Better Way of (Multilevel) Life

Between-Person Variation L2 τ^2_{0P0}	Between-Item Variation L2 τ^2_{00I}	Trial (Person*Item) Variation L1 σ^2_e

Random effects over **persons** of **item** or **trial** predictors can also be tested and predicted.

- **Multilevel Model with Crossed Random Effects:**

$$\mathrm{RT}_{tpi} = \gamma_{000} + \gamma_{001}\mathrm{A}_i + \gamma_{002}\mathrm{B}_i + \gamma_{003}\mathrm{A}_i\mathrm{B}_i + \mathbf{U}_{0p0} + \mathbf{U}_{00i} + \mathbf{e}_{tpi}$$

t **trial**
p **person**
i **item**

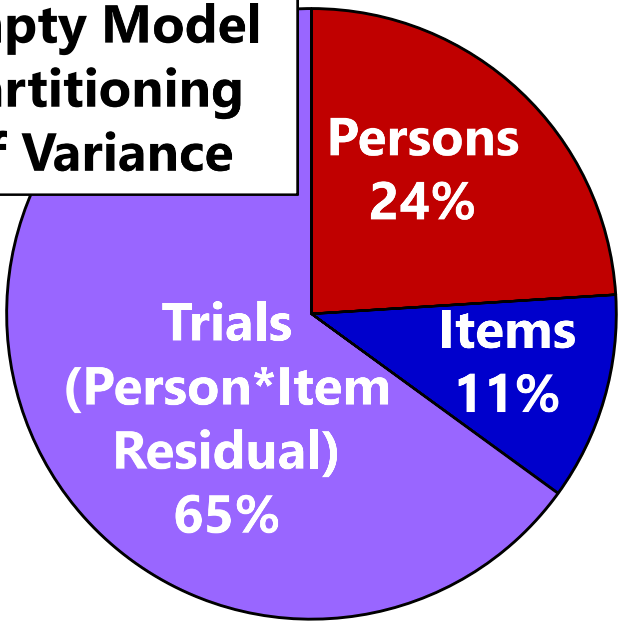
- Explicitly test **persons** and **items** as random effects:
  - Person predictors capture between-person mean variation: $\boldsymbol{\tau^2_{0P0}}$
  - Item predictors capture between-item mean variation: $\boldsymbol{\tau^2_{00I}}$
  - Trial predictors capture trial-specific residual variation: $\boldsymbol{\sigma^2_e}$

Larry's Data: See bonus posted on 10/31

- Crossed design: 38 persons by 39 items (words or nonwords)
- Lexical decision task: Response Time to decide if word or nonword
- 2 word-specific predictors of interest:
  - A: Low/High Phonological Neighborhood Frequency
  - B: Small/Large Semantic Neighborhood Size

***F2 ANOVA was closest to MLM results**

Empty Model Partitioning of Variance



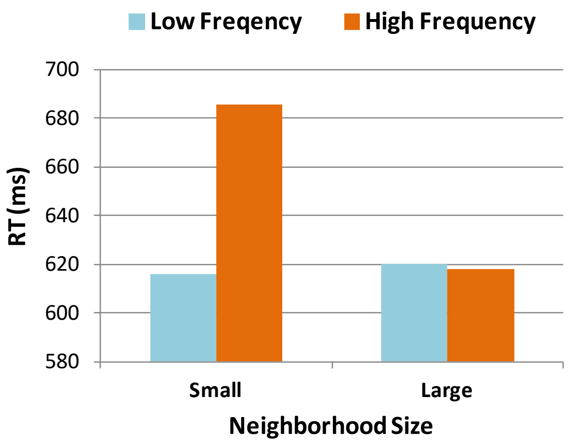
Model and Results

$$RT_{tpi} = \gamma_{000} + \gamma_{001}A_i + \gamma_{002}B_i + \gamma_{003}A_iB_i + U_{0p0} + U_{00i} + e_{tpi}$$

Pseudo-R^2:

Residual ≈ 0%
Person ≈ 0%
Items ≈ 30%*
Total R^2 ≈ 3.3%

***Significant item variance remained**



Not Just in Larry's Example Data…

- Generality of results examined via simulation study of Type I error rates for person or item predictor effects

- **Testing person effects in common persons design?**
  - Need **person** variance to exist in model (so not **F2 ANOVA**)
  - Need random effect for **persons** (in **MLM** or in **F1 ANOVA**), so that **person** predictors can explain that **person** variance

- **Testing item effects in common items design?**
  - Need **item** variance to exist in model (so not **F1 ANOVA**)
  - Need random effect for **items** (in **MLM** or in **F2 ANOVA**), so that **item** predictors can explain that **item** variance

Nested vs. Crossed Items in Multilevel Designs

- When should **items** be a separate level-2 **random effect**?
  - Items are clearly nested within persons if the model **fixed effects explain all** of the item variation (so no item variation remains)
    - e.g., via item-specific indicators (CFA, IRT; stay tuned)
    - e.g., by item design features given only one item per condition
  - Items are clearly nested within persons if they are **endogenous**
    - e.g., autobiographical memories, eye movements, speech utterances
  - More ambiguous if items are **randomly generated** per person
    - If items are truly unique per person, then there are no common items… but items are usually constructed systematically
    - Modeling items as **nested (no variance) assumes exchangeability**
- When does this matter? When turning **experimental tasks** into instruments in which the outcome is non-normal, and we want to predict sources of item difficulty

Latent Variables = Random Effects

- **1PL model** predicts accuracy via fixed item effects and random person effects (i.e., n items are nested in persons)
- **1PL model**:
  - $\text{Probability}\left(\mathrm{y_{pi}} = 1 \middle| \boldsymbol{\theta}_{\mathbf{p}}\right) = \frac{\exp\left(\boldsymbol{\theta}_{\mathbf{p}} - \mathbf{b_i}\right)}{1+\exp\left(\boldsymbol{\theta}_{\mathbf{p}} - \mathbf{b_i}\right)}$
  - $\text{Logit}\left(\mathrm{y_{pi}} = 1 \middle| \boldsymbol{\theta}_{\mathbf{p}}\right) = \boldsymbol{\theta}_{\mathbf{p}} - \mathbf{b_i}$

$\mathbf{b_i}$ is fixed effect of <u>difficulty</u> per item

$\boldsymbol{\theta}_{\mathbf{p}}$ is random person ability (variance τ^2_{θ})

- **1PL is also a generalized multilevel model (t = trial)**:
  - $\text{Logit}\left(\mathrm{y_{tpi}} = 1 \middle| \mathbf{U_{0p0}}\right) = \boldsymbol{\gamma}_{\mathbf{001}}\mathbf{I_1} + \boldsymbol{\gamma}_{\mathbf{002}}\mathbf{I_2} + \cdots + \boldsymbol{\gamma}_{\mathbf{00}\boldsymbol{n}}\mathbf{I}_{\boldsymbol{n}} + \mathbf{U_{0p0}}$
  - Because item difficulty/easiness is perfectly predicted by the ***I*** indicator variables, here **items do not need a level-2 crossed random effect**

$\boldsymbol{\gamma}_{\mathbf{00i}}$ is fixed effect of <u>easiness</u> per item

$\mathbf{U_{0p0}}$ is random person ability (variance $\tau^2_{\mathbf{0P0}}$)

Latent Variables = Random Effects

- **1PL model identification**:
  - $\text{Logit}\left(y_{pi} = 1 \middle| \boldsymbol{\theta}_\mathbf{p}\right) = \boldsymbol{\theta}_\mathbf{p} - \mathbf{b}_\mathbf{i}$
  - On means side, fix one of these to 0:
    - One item difficulty, sum of item difficulties, or theta mean
  - One variance side, fix one of these to 1:
    - Item discrimination, or theta variance

$\mathbf{b}_\mathbf{i}$ is fixed effect of <u>difficulty</u> per item

$\boldsymbol{\theta}_\mathbf{p}$ is random person ability (variance τ^2_θ)

γ_{00i} is fixed effect of <u>easiness</u> per item

$\mathbf{U}_{\mathbf{p0}}$ is random person ability (variance τ^2_{0P0})

- **1PL as Generalized MLM**:
  - $\text{Logit}\left(y_{tpi} = 1 \middle| \mathbf{U}_{\mathbf{0p0}}\right) = \boldsymbol{\gamma}_{\mathbf{001}}\mathbf{I}_\mathbf{1} + \boldsymbol{\gamma}_{\mathbf{002}}\mathbf{I}_\mathbf{2} + \cdots + \boldsymbol{\gamma}_{\mathbf{00}n}\mathbf{I}_n + \mathbf{U}_{\mathbf{0p0}}$
  - Will be on the same scale as 1PL (“Rasch”) model when theta mean = 0 and item discrimination is fixed to 1 so that theta’s variance can be estimated

Adding Lego #1: Linear Models

- 1PL can be extended to **predict item difficulty** via the LLTM

- **LLTM →** k item features predict b_i; random persons ($\boldsymbol{\theta}_\mathbf{p}$):
  - $\text{Logit}(y_{pi} = 1 | \boldsymbol{\theta}_\mathbf{p}) = \boldsymbol{\theta}_\mathbf{p} - \mathbf{b_i}$
  - $\mathbf{b_i} = \boldsymbol{\gamma}_\mathbf{0} + \boldsymbol{\gamma}_\mathbf{1}\mathbf{X_{1i}} + \boldsymbol{\gamma}_\mathbf{2}\mathbf{X_{2i}} + \cdots + \boldsymbol{\gamma}_\mathbf{k}\mathbf{X_{ki}}$

Item difficulty =linear model of k item features (of X*γ fixed effects); $\boldsymbol{\theta}_\mathbf{p}$ **is random person ability (variance $\boldsymbol{\tau}^2_\boldsymbol{\theta}$)**

- **LLTM written as a generalized multilevel model**:
  - $\text{Logit}(y_{tpi} = 1 | \mathbf{U_{p0}}) = \boldsymbol{\gamma}_\mathbf{000} + \boldsymbol{\gamma}_\mathbf{001}\mathbf{X_{1i}} + \boldsymbol{\gamma}_\mathbf{002}\mathbf{X_{2i}} + \cdots + \boldsymbol{\gamma}_\mathbf{00k}\mathbf{X_{ki}} + \mathbf{U_{0p0}}$
  - Because there is no random item effect, the model says that items are still just nested within persons—that item difficulty or easiness is ***perfectly*** predicted by the X item features (no item differences remain)

Item easiness = a linear model of k item features (of X*γ fixed effects); $\mathbf{U_{0p0}}$ **is random person ability (variance $\boldsymbol{\tau}^2_\mathbf{0P0}$)**

Example: Measuring Visual Search Ability

Blank 80 ms | Blank 80 ms | Blank 80 ms | Blank 80 ms

A 280 ms | A 280 ms | A′ 280 ms | A′ 280 ms

cycle continues until response for max of 45 sec

Change detection task using the "flicker paradigm"

<u>Rated Item Design Features:</u>

- Visual clutter of the scene
- Relevance of the change to driving
- Brightness of the change
- Change made to legible sign

- 155 persons, 46 items retained, DV = accuracy (for now)

Proof of Concept: Random Items Matters

Item re-analysis predicting accuracy in dissertation data using SAS PROC GLIMMIX (Laplace estimation)

Effect	Items Treated as Fixed			Items Treated as Random		
	Est	SE	*p* <	Est	SE	*p* <
Intercept	0.862	0.153	.0001	1.311	0.635	.0474
Clutter	-0.268	0.055	.0001	-0.324	0.242	.1809
Relevance	0.220	0.099	.0266	0.037	0.426	.9305
Brightness	0.474	0.113	.0001	0.790	0.499	.1136
Legible Sign	0.662	0.082	.0001	0.739	0.337	.0283

Putting It All Together…

- Experimental tasks can become psychometric instruments via **explanatory IRT (generalized multilevel) models** in which **items** and **persons** have crossed random effects at level 2

$$\text{Logit}(y_{tpi}=1)=\gamma_{000}+\gamma_{001}X_{1i}+\gamma_{002}X_{2i}+\cdots+\mathbf{U_{0p0}}+\mathbf{U_{00i}}$$

  - $\mathbf{U_{0p0}}$ is person ability with random (unpredicted) variance of $\boldsymbol{\tau^2_{0P0}}$
  - $\mathbf{U_{00i}}$ is item easiness is predicted from a linear model of the X item features, with random (leftover) variance of $\boldsymbol{\tau^2_{00I}}$
  - Can add person predictors to explain $\boldsymbol{\tau^2_{0P0}}$
  - Can examine random effects across persons of X item features (i.e., differential susceptibility to item manipulations)

- Let's try to estimate these models using SAS and STATA…