A (Brief) Introduction to Crossed Random Effects Models for Repeated Measures Data

- Today's Class:
  - Review of concepts in multivariate data
  - Introduction to random intercepts
  - Crossed random effects models for repeated measures

The Two Sides of *Any* Model

- **Model for the Means:**
  - *Aka* **Fixed Effects**, Structural Part of Model
  - What you are used to **caring about for testing hypotheses**
  - How the expected outcome for a given observation varies as a function of values on predictor variables

- **Model for the Variance:**
  - *Aka* **Random Effects and Residuals**, Stochastic Part of Model
  - How residuals are distributed and related across observations
  - What you are used to **making assumptions about** instead…
  - For general linear models, that residuals come from a **normal** distribution, are **independent** across persons, and have **constant variance** across persons and predictors ("identically distributed")

The Two Sides of a *General Linear* Model

$$\boxed{y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i \dots} + \boxed{e_i}$$

Our new focus

- **Model for the Variance:**
  - $e_i \sim N(0, \sigma_e^2)$ → ONE residual (unexplained) deviation
  - e_i has a mean of 0 with some estimated constant variance σ_e^2, is normally distributed, is unrelated to predictors, and is unrelated across observations (across all people here)
  - **Estimated parameter is residual variance** (not each e_i)
  - What happens when each person has more than one y_i? A single independent e_i will not be sufficient because:
    - Each outcome may have a different amount of residual variance
    - Residuals of outcomes from the same person will be correlated
    - So we need multivariate models with a new model for the variance

Comparing Models for the Variance

- **Relative model fit** is indexed by 2*sum of individual LL values = **−2LL**
  - **−2LL indicates BADNESS of fit (shortness), so smaller values = better models**
  - **LL indicates GOODNESS of fit (tallness), so larger values = better models**

- **Nested variance models are compared using −2LL values: −2ΔLL Test** (aka, "χ^2 test" in SEM; "deviance difference test" in MLM)

"fewer" = from model with fewer parameters
"more" = from model with more parameters

Results of 1. and 2. must be positive values!

1. Calculate $-2\Delta LL = (-2LL_{fewer}) - (-2LL_{more})$ OR $-2\Delta LL = -2 * (LL_{fewer} - LL_{more})$
2. Calculate Δdf: $(\#\ Parms_{more}) - (\#\ Parms_{fewer})$
3. Compare $-2\Delta LL$ to χ^2 distribution with df = Δdf
   CHIDIST in excel will give exact p-values for the difference test; so will STATA lrtest

- Nested or non-nested models can also be compared by **<u>Information Criteria</u>** that reflect **−2LL** AND # parameters used and/or sample size
  - **AIC** = Akaike IC = **−2LL** + 2 *(#parameters)
  - **BIC** = Bayesian IC = **−2LL** + log(N)*(#parameters) → penalty for complexity
  - No significance tests or critical values, just "smaller is better"

Types of Multivariate Models

When $\mathbf{y_i}$ is still a single outcome conceptually, but:

- You have 2+ outcomes per person as created by multiple conditions (e.g., longitudinal or repeated measures designs)
  - <u>If there really is only one outcome per condition</u>, then "ANOVA" models are potentially problematic restrictions of more general multivariate models in which there is a "right answer" for the residual variance and covariance across conditions (as shown in Lecture 5 and Example 5)
  - <u>If each condition has more than one outcome (e.g., per trial)</u>, do **NOT** aggregate them into a condition mean outcome! Up next is what to do instead, although there will not be a "right answer" of variance and covariance against which to judge the fit of your model for the variance

- When your y_i comes from people nested/clustered in groups (e.g., children nested in teachers, people nested in families)
  - You really have multivariate outcomes of a group, and there also won't be a single "right answer" for the model for the variance (up next time)

From a “Multivariate” to “Stacked” Data

New data structure so that y_i is still a single outcome….

RM ANOVA uses “wide” **multivariate** data structure:

ID	Girl	T1	T2	T3	T4
100	0	5	6	8	12
101	1	4	7	.	11

A row = a case = a person

So people missing any data are excluded (data from ID 101 are not included at all)

ML/REML in MIXED uses “long” or **stacked** data structure instead:

ID	Girl	Time	Y
100	0	1	5
100	0	2	6
100	0	3	8
100	0	4	12
101	1	1	4
101	1	2	7
101	1	3	.
101	1	4	11

A case is now one outcome per person

Only cases missing data are excluded

ID 100 uses 4 cases

ID 101 uses 3 cases

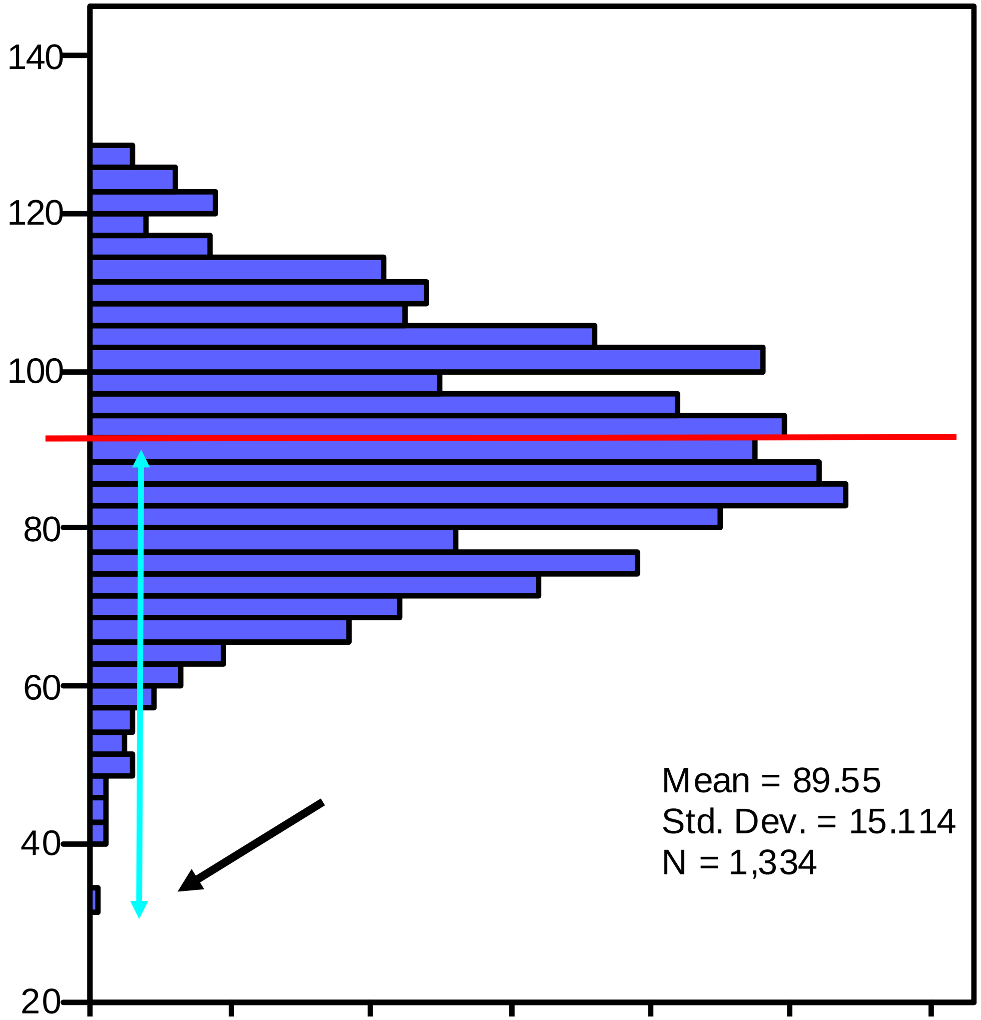
Time can also be **unbalanced** across people such that each person can have his or her own measurement schedule: Time “0.9” “1.4” “3.5” “4.2”…

Multivariate = Multilevel Models

- When $\mathbf{y_i}$ is still a single outcome conceptually, but you have more than one $\mathbf{y_i}$ per person or per group, the models (for the variance) used for these data are usually referred to as “multilevel” models
  - aka, hierarchical linear models, general linear mixed models

- They are based on the idea of separating what was just a single “residual variance” into multiple “kinds” of variance that arise from different dimensions of sampling, each of which can be explained by predictors of that same kind
  - e.g., between-person, between-item, between-group variances
  - A “level” is a set of variances that are unrelated to the other sets of variances, but we won’t worry about this notation for now…

An Empty Between-Person Model (i.e., Single-Level)



$$y_i = \beta_0 + e_i$$

Filling in values:

$$32 = \underbrace{90}_{y_{pred}} + -58$$

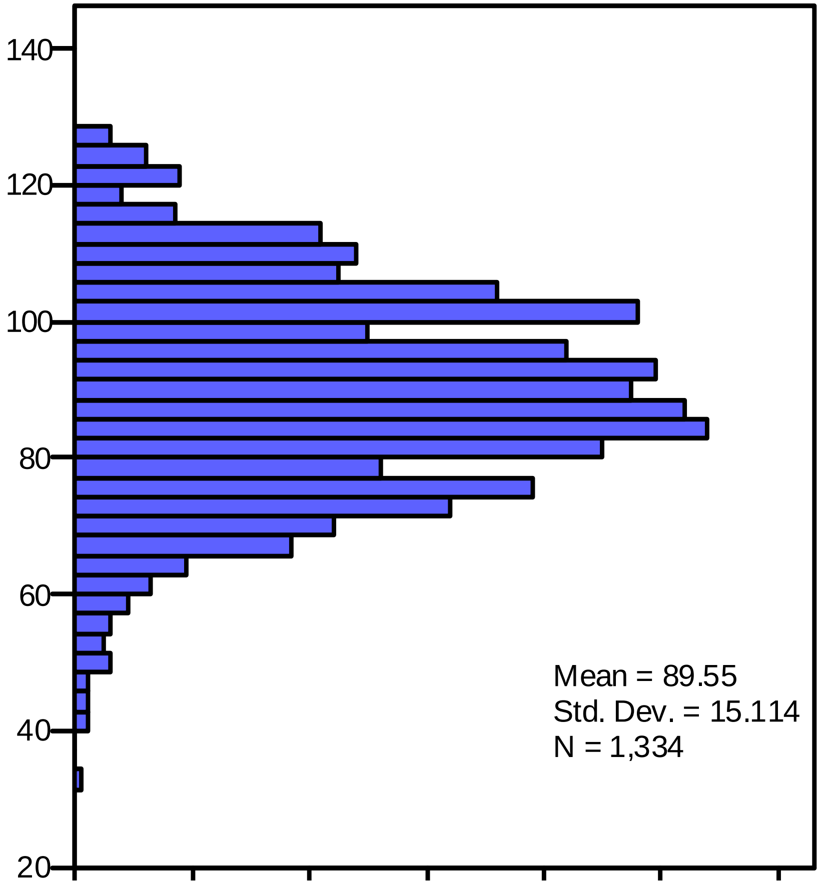
Model for the Means

y_i error variance:

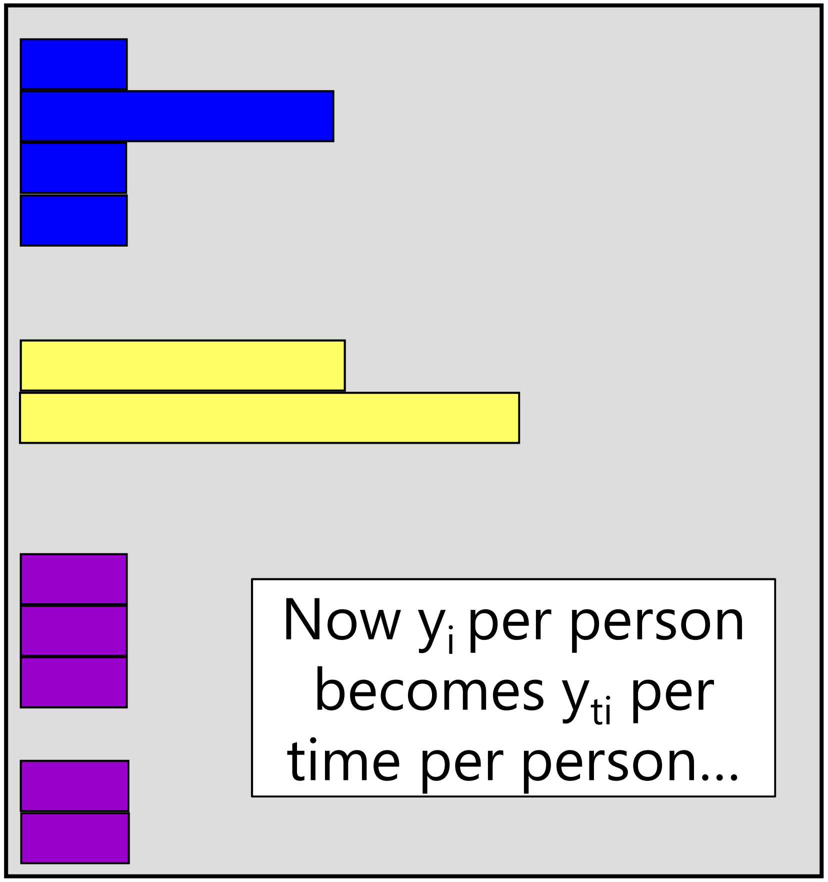
$$\frac{\sum (y_i - y_{pred})^2}{N - 1}$$

Adding Within-Person Information… (i.e., to become a Multilevel Model)

Full Sample Distribution

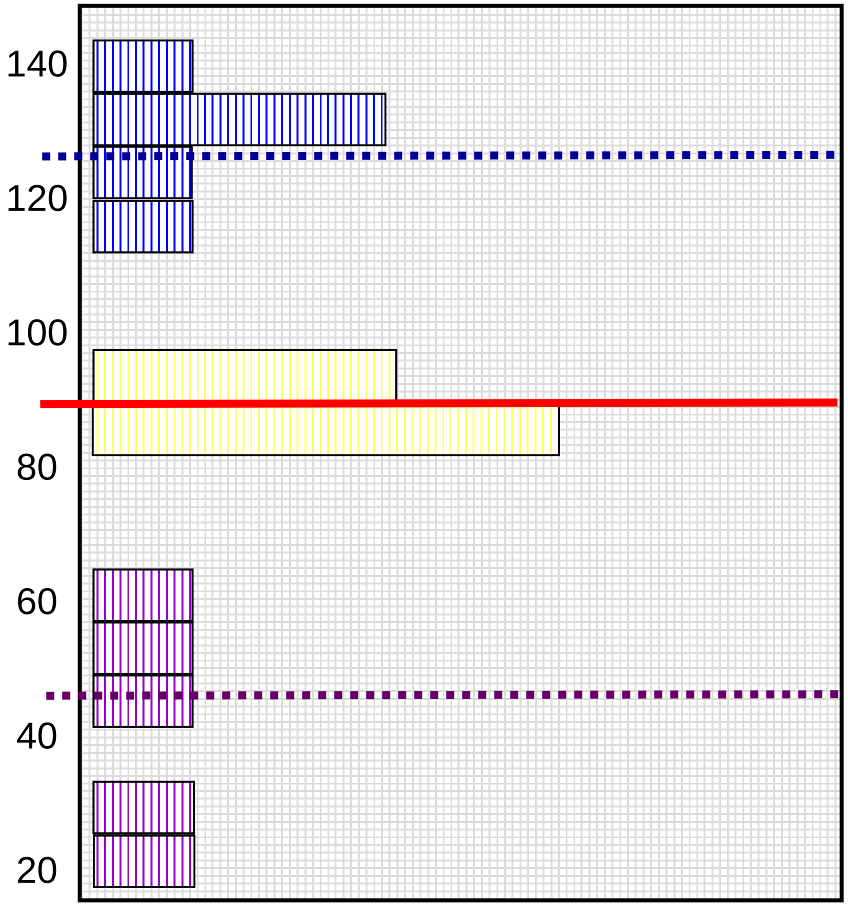


3 People, 5 Occasions each



Now y_i per person becomes y_{ti} per time per person…

Empty +Within-Person Model for y_{ti}



Start off with mean of y_{ti} as "best guess" for any value:

= Grand Mean

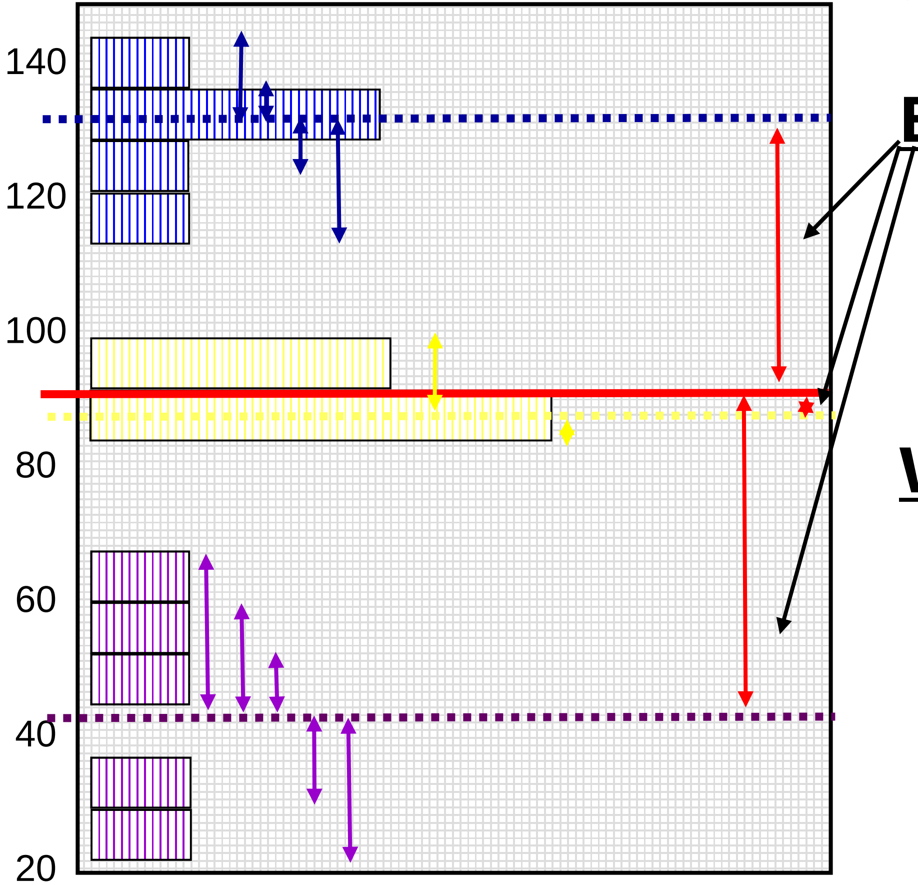
= Fixed Intercept

Can make better guess by taking advantage of repeated observations:

= Person Mean

→ Random Intercept

Empty +Within-Person Model



y_{ti} variance → 2 sources:

Between-Person (BP) Variance:

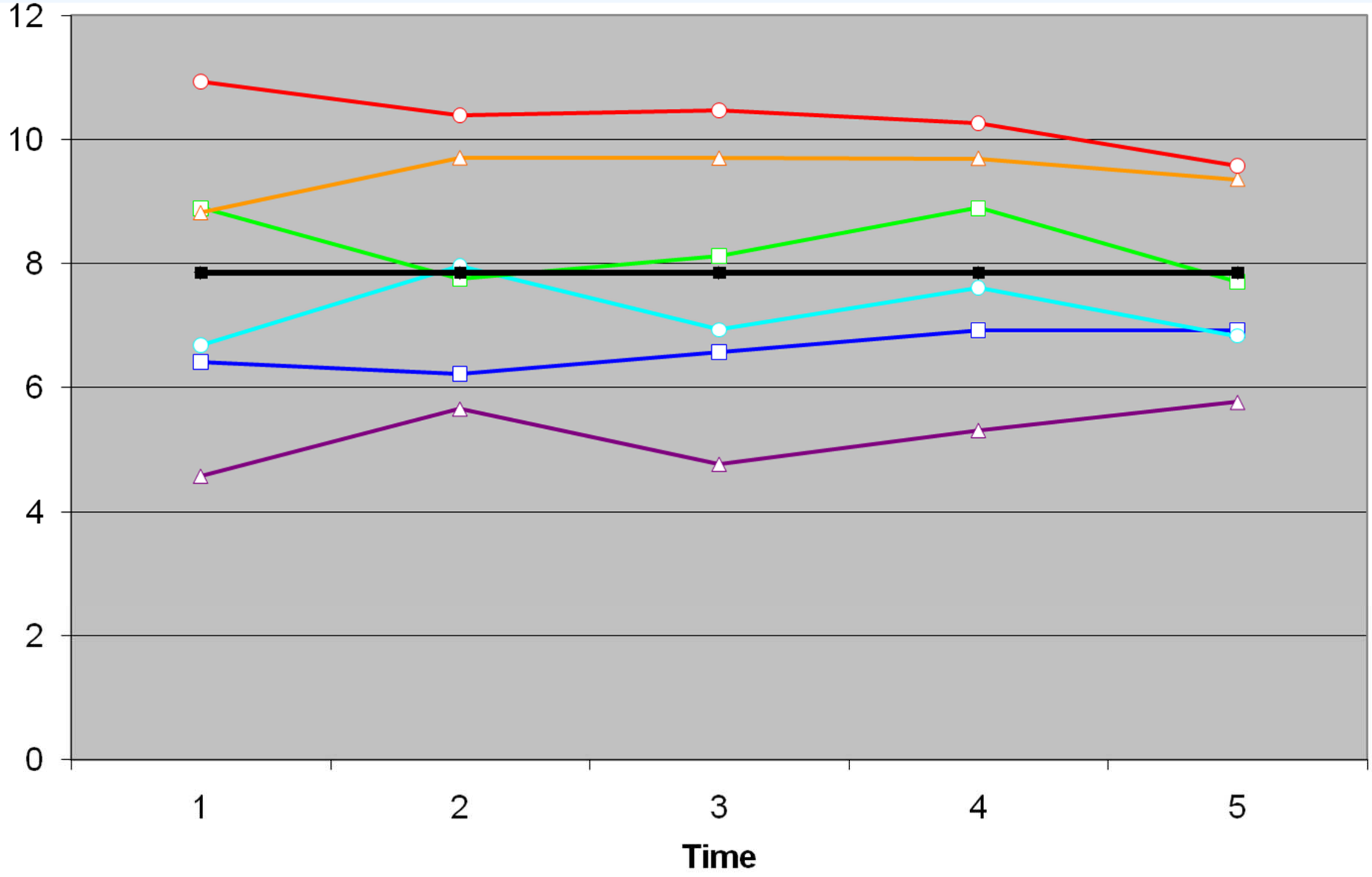
- → Differences from **GRAND** mean
- → **INTER**-Individual Differences

Within-Person (WP) Variance:

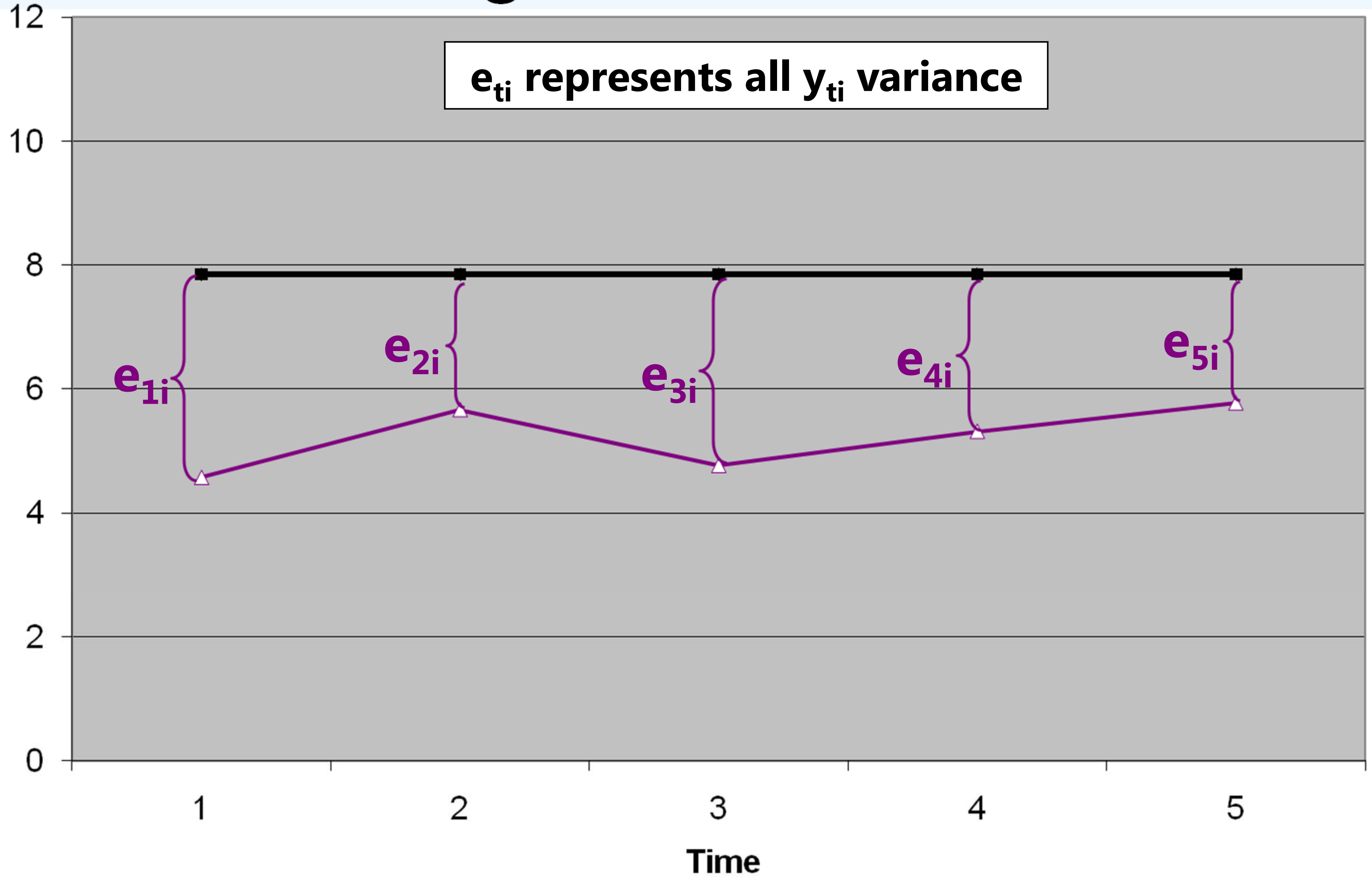
- → Differences from **OWN** mean
- → **INTRA**-Individual Differences
- → This part is only observable through longitudinal data.

Now we have 2 piles of variance in y_{ti} to predict.

Hypothetical Longitudinal Data (black line = sample mean)

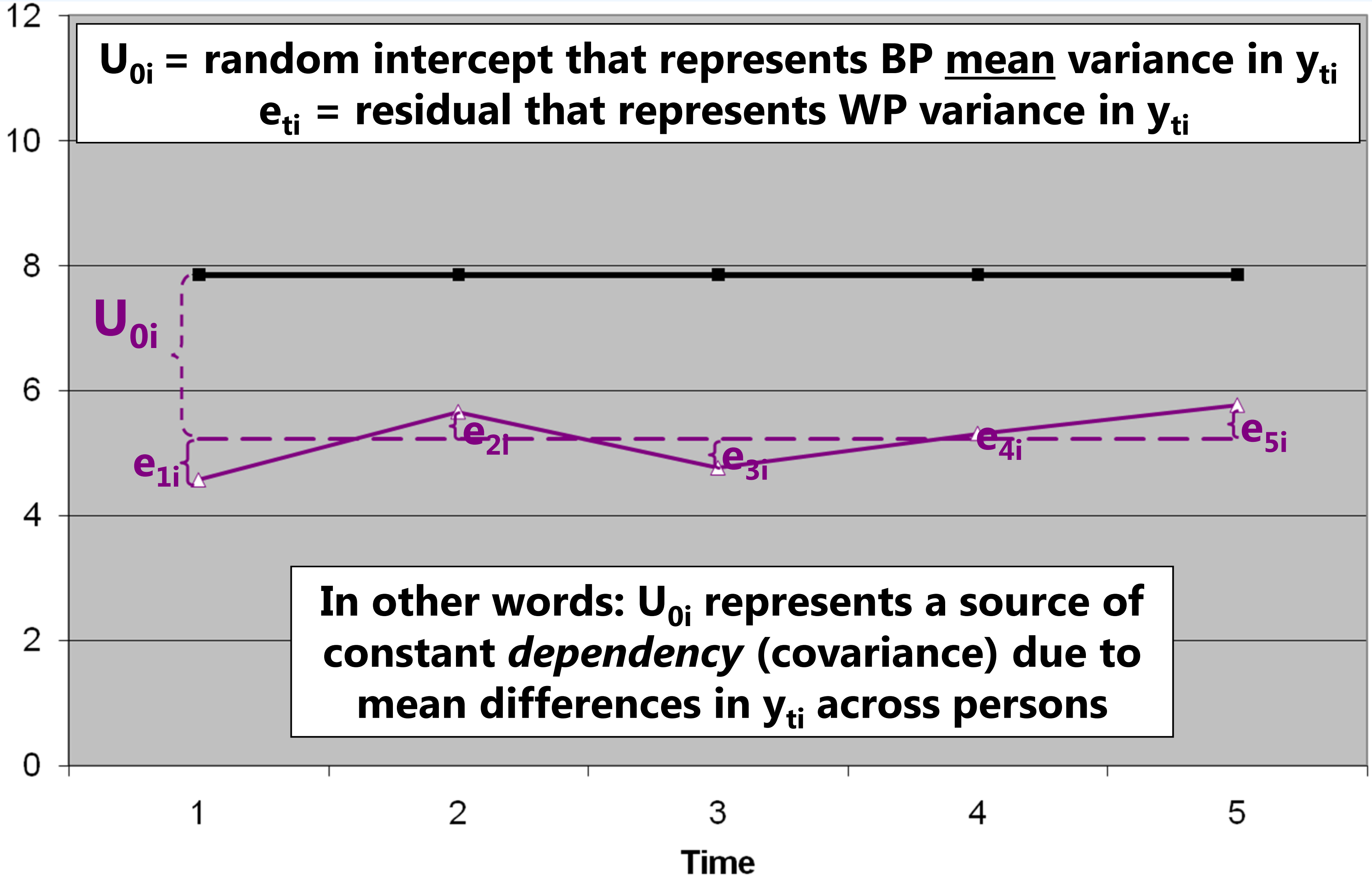


"Error" in a BP Model for the Variance: Single-Level Model



e_{ti} represents all y_{ti} variance

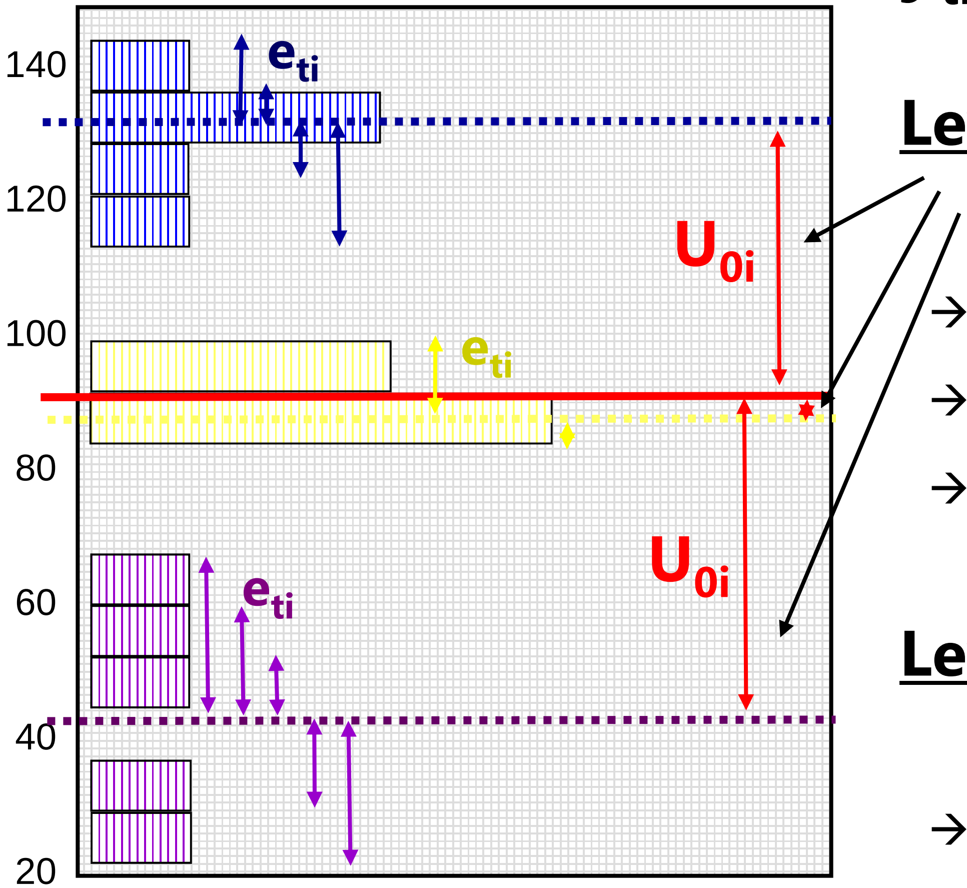
"Error" in a +WP Model for the Variance: Multilevel Model



U_{0i} = random intercept that represents BP <u>mean</u> variance in y_{ti}
e_{ti} = residual that represents WP variance in y_{ti}

In other words: U_{0i} represents a source of constant *dependency* (covariance) due to mean differences in y_{ti} across persons

Empty +Within-Person Model



y_{ti} variance → 2 sources:

Level 2 Random Intercept Variance (of U_{0i}, as $\tau^2_{U_0}$):

- → **Between**-Person Variance
- → Differences from **GRAND** mean
- → **INTER**-Individual Differences

Level 1 Residual Variance (of e_{ti}, as σ^2_e):

- → **Within**-Person Variance
- → Differences from **OWN** mean
- → **INTRA**-Individual Differences

BP vs. +WP Empty Models

- Empty **Between-Person** Model (used for 1 occasion):

$$y_i \quad = \quad \beta_0 \quad + \quad e_i$$

  - β_0 = fixed intercept = grand mean
  - e_i = residual deviation from GRAND mean

- Empty **+Within-Person** Model (for >1 occasions):

$$y_{ti} \quad = \quad \beta_0 \quad + \quad U_{0i} \quad + \quad e_{ti}$$

  - β_0 = fixed intercept = grand mean
  - U_{0i} = random intercept = individual deviation from GRAND mean
  - e_{ti} = time-specific residual deviation from OWN mean

BP and +WP Conditional Models

- Multiple Regression, **Between-Person** ANOVA: **1 PILE**
  - $\mathbf{y_i = (\beta_0 + \beta_1 X_i + \beta_2 Z_i \ldots) + e_i}$
  - $\mathbf{e_i}$ → ONE residual, assumed uncorrelated with equal variance across observations (here, just persons) → "**BP (all) variation**"

- Repeated Measures, **Within-Person** ANOVA: **2 PILES**
  - $\mathbf{y_{ti} = (\beta_0 + \beta_1 X_i + \beta_2 Z_i \ldots) + U_{0i} + e_{ti}}$
  - $\mathbf{U_{0i}}$ → A random intercept for differences in person means, assumed uncorrelated with equal variance across persons → "**BP (mean) variation**"= $\boldsymbol{\tau^2_{U_0}}$ is "leftover" after BP predictors
  - $\mathbf{e_{ti}}$ → A residual that represents remaining time-to-time variation, usually assumed uncorrelated with equal variance across observations (now, persons and time) → "**WP variation**" = $\boldsymbol{\sigma^2_e}$ is also now "leftover" after WP predictors

ANOVA works well when…

- Experimental stimuli are ***controlled*** and ***exchangeable***
  - Controlled → Constructed, not sampled from a population
  - Exchangeable → Stimuli vary only in dimensions of interest
  - …What to do with non-exchangeable stimuli (e.g., words, scenes)?

- Experimental manipulations create ***discrete conditions***
  - e.g., set size of 3 vs. 6 vs. 9 items
  - e.g., response compatible vs. incompatible distractors
  - …What to do with *continuous* item predictors (e.g., time, salience)?

- One has ***complete data***
  - e.g., if outcome is RT and accuracy is near ceiling
  - e.g., if responses are missing for no systematic reason
  - …What if data are not missing completely at random (e.g., inaccuracy)?

Motivating Example: Psycholinguistic Study Designs

- Word Recognition Tasks (e.g., Lexical Decision)
  - Word lists are constructed based on targeted dimensions while controlling for other relevant dimensions
  - Outcome = RT to decide if the stimulus is a word or non-word (accuracy is usually near ceiling)

- Tests of effects of experimental treatment are typically conducted with the person as the unit of analysis…
  - Average the responses over words within conditions
    - Contentious fights with reviewers about adequacy of experimental control when using real words as stimuli
    - Long history of debate as to how words as experimental stimuli should be analyzed… F_1 ANOVA or F_2 ANOVA (or both)?
    - F_1 only creates a “Language-as-Fixed-Effects Fallacy” (Clark, 1973)

ANOVAs on Summary Data

Original Data per Subject

	B1	B2
A1	Trial 001 Trial 002 .... Trial 100	Trial 101 Trial102 .... Trial 200
A2	Trial 201 Trial 202 .... Trial 300	Trial 301 Trial302 .... Trial 400

Subject Summary Data

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

"F_1" Repeated Measures ANOVA on *N* subjects:

$$\mathrm{RT}_{cs} = \gamma_0 + \gamma_1 A_c + \gamma_2 B_c + \gamma_3 A_c B_c + \mathbf{U_{0s}} + e_{cs}$$

"F_2" Between-Groups ANOVA on *T* trials:

$$\mathrm{RT}_t = \gamma_0 + \gamma_1 A_t + \gamma_2 B_t + \gamma_3 A_t B_t + e_t$$

Trial Summary Data

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

Choosing Amongst ANOVA Models

- F_1 RM ANOVA on **subject** summary data:
  - Assumes trials are fixed—within-condition ***trial*** variability is gone
- F_2 ANOVA on **trial** summary data:
  - Assumes persons are fixed—within-trial ***subject*** variability is gone

- Proposed ANOVA-based resolutions:
  - **F′** → quasi-F test that treats both trials and subjects as random (Clark, 1973), but requires complete data (least squares)
  - **Min F′** → lower-bound of F′ derived from F1 and F2 results, which does not require complete data, but is (too) conservative
  - **F_1 x F_2 criterion** → effects are only "real" if they are significant in **both F_1 and F_2 models** (aka, death knell for psycholinguists)

  - But neither model is complete (two wrongs don't make a right)…

Multilevel Models to the Rescue?

Original Data per Person

	B1	B2
A1	Trial 001 Trial 002 ……… Trial 100	Trial 101 Trial102 ……… Trial 200
A2	Trial 201 Trial 202 ……… Trial 300	Trial 301 Trial302 ……… Trial 400

Pros:

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

Cons:

- **Is still wrong**

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

Level 2: $\beta_{0s} = \gamma_{00} + U_{0s}$

$\beta_{1s} = \gamma_{10}$

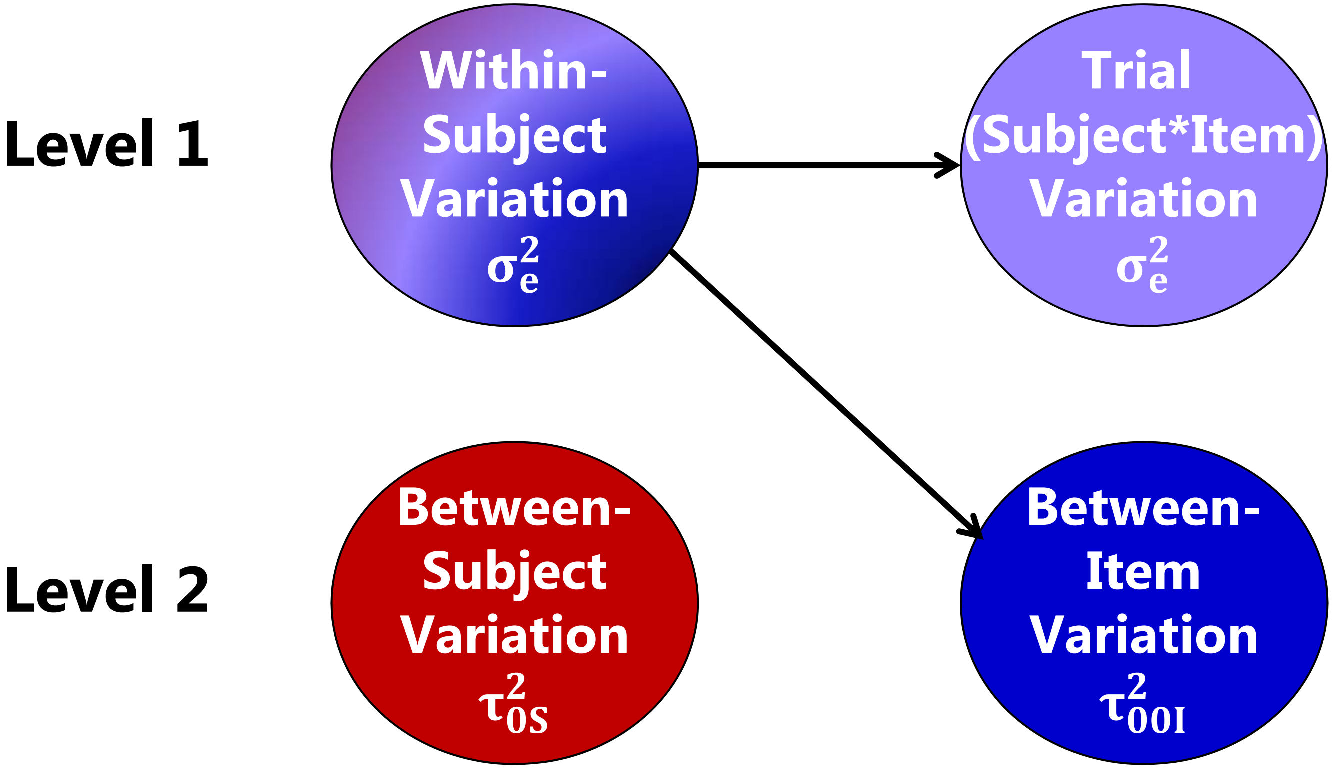
$\beta_{2s} = \gamma_{20}$

$\beta_{3s} = \gamma_{30}$

Level 1 = Within-Subject Variation
(Across Trials)

Level 2 = Between-Subject Variation

Multilevel Models to the Rescue?



Level 1

Within-Subject Variation σ_e^2

Trial (Subject*Item) Variation σ_e^2

Level 2

Between-Subject Variation τ_{0S}^2

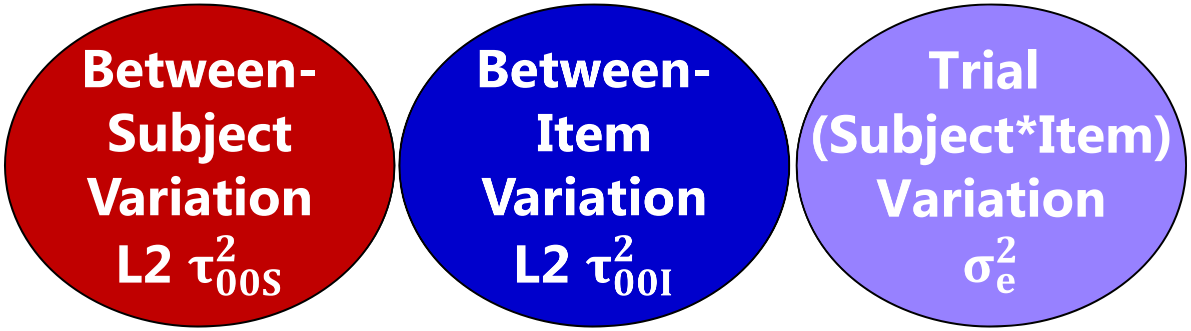
Between-Item Variation τ_{00I}^2

Empty Means, Crossed Random Effects Models

Note the new symbol for a fixed effect: now γ **(gamma)** instead of β **(beta)** to follow traditional multilevel model notation…

- **Residual-only model:**
  - $RT_{tis} = \gamma_{000} + e_{tis}$
  - Assumes no effects (dependency) of subjects or items
- **Random subjects model:**
  - $RT_{tis} = \gamma_{000} + \mathbf{U_{00s}} + e_{tis}$
  - Models systematic mean differences **between subjects**
- **Random subjects and items model:**
  - $RT_{tis} = \gamma_{000} + U_{00s} + \mathbf{U_{0i0}} + e_{tis}$
  - **Also** models systematic mean differences **between items**

A Better Way of (Multilevel) Life



Between-Subject Variation
L2 τ^2_{00S}

Between-Item Variation
L2 τ^2_{00I}

Trial (Subject*Item) Variation
σ^2_e

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

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

$$\mathrm{RT}_{tis} = \gamma_{000} + \gamma_{010}A_i + \gamma_{020}B_i + \gamma_{030}A_iB_i$$
$$+\mathbf{U_{00s}} + \mathbf{U_{0i0}} + \mathbf{e_{tis}}$$

t trial
i item
s subject

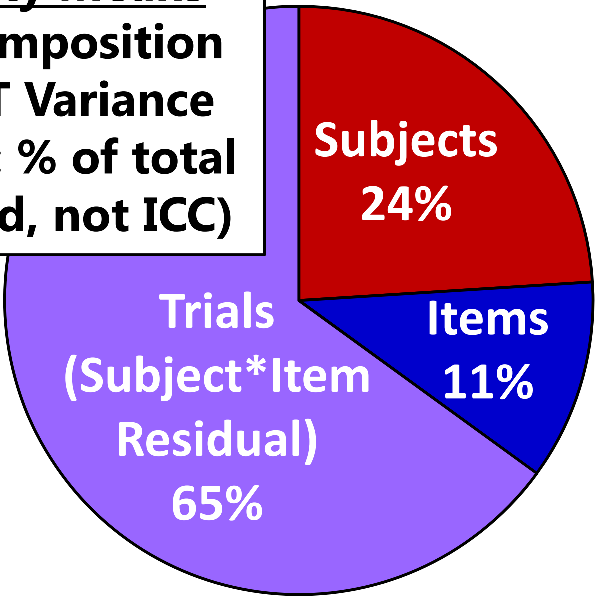
- Both **subjects** and **items** as random effects:
  - Subject predictors explain between-subject mean variation: $\boldsymbol{\tau^2_{00S}}$
  - Item predictors explain between-item mean variation: $\boldsymbol{\tau^2_{00I}}$
  - Trial predictors explain trial-specific residual variation: $\boldsymbol{\sigma^2_e}$

Example Psycholinguistic Study

(Locker, Hoffman, & Bovaird, 2007)

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

Empty Means Decomposition of RT Variance (note: % of total is used, not ICC)



Model and Results

$$\mathrm{RT}_{tis} = \gamma_{000} + \gamma_{010}A_i + \gamma_{020}B_i + \gamma_{030}A_iB_i + \mathbf{U_{00s}} + \mathbf{U_{0i0}} + \mathbf{e_{tis}}$$

Pseudo-R²:

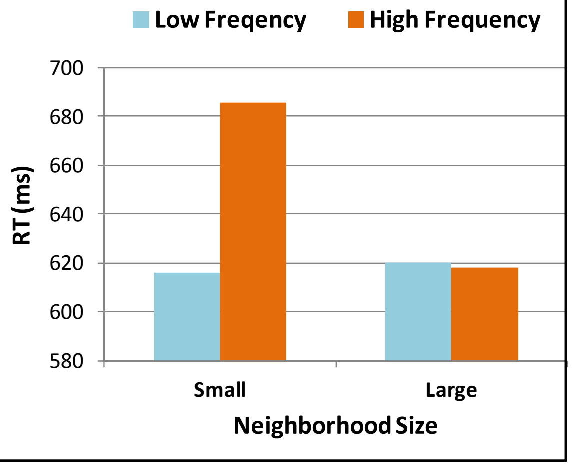
Residual ≈ 0%

Subjects ≈ 0%

Items ≈ 30%*

Total R² ≈ 3.3%

***Significant item variability remained**



Tests of Fixed Effects by Model

	A: Frequency Marginal Main Effect	B: Size Marginal Main Effect	A*B: Interaction of Frequency by Size
F_1 Subjects ANOVA	$F\,(1,37) = 16.1$ $p = .0003$	$F\,(1,37) = 14.9$ $p = .0004$	$F\,(1,37) = 38.2$ $p < .0001$
F_2 Words ANOVA	$F\,(1,35) = 5.3$ $p = .0278$	$F\,(1,35) = 4.5$ $p = .0415$	$F\,(1,35) = 5.7$ $p = .0225$
F′ min (via ANOVA)	$F\,(1,56) = 4.0$ $p = .0530$	$F\,(1,55) = 3.5$ $p = .0710$	$F\,(1,45) = 5.0$ $p = .0310$
Crossed MLM (via REML)	$F\,(1,32) = 5.4$ $p = .0272$	$F\,(1,32) = 4.6$ $p = .0393$	$F\,(1,32) = 6.0$ $p = .0199$

Simulation: Type 1 Error Rates

Condition			Models					
Item Variance	Subject Variance		**1: Both Random Effects**	2: Random Subjects Only	3: Random Items Only	4: No Random Effects	5: F1 Subjects ANOVA	6: F2 Item ANOVA
Item Effect:								
2	2		**0.03**	0.09	0.03	0.09	0.09	0.03
2	10		**0.05**	0.14	0.05	0.12	0.15	0.05
10	2		**0.04**	0.32	0.04	0.31	0.32	0.04
10	10		**0.05**	0.31	0.05	0.29	0.33	0.05
Subject Effect:								
2	2		**0.04**	0.04	0.12	0.11	0.04	0.12
2	10		**0.05**	0.05	0.34	0.34	0.05	0.36
10	2		**0.04**	0.03	0.12	0.09	0.03	0.12
10	10		**0.06**	0.06	0.34	0.31	0.05	0.37

Model Items as Fixed → Wrong Item Effect

Condition			Models					
Item Variance	Subject Variance		1: Both Random Effects	**2: Random Subjects Only**	3: Random Items Only	4: No Random Effects	**5: F1 Subjects ANOVA**	6: F2 Item ANOVA
Item Effect:								
2	2		0.03	**0.09**	0.03	0.09	**0.09**	0.03
2	10		0.05	**0.14**	0.05	0.12	**0.15**	0.05
10	2		0.04	**0.32**	0.04	0.31	**0.32**	0.04
10	10		0.05	**0.31**	0.05	0.29	**0.33**	0.05
Subject Effect:								
2	2		0.04	0.04	0.12	0.11	0.04	0.12
2	10		0.05	0.05	0.34	0.34	0.05	0.36
10	2		0.04	0.03	0.12	0.09	0.03	0.12
10	10		0.06	0.06	0.34	0.31	0.05	0.37

Model Subjects as Fixed → Wrong Subject Effect

Condition			Models					
Item Variance	Subject Variance		1: Both Random Effects	2: Random Subjects Only	**3: Random Items Only**	4: No Random Effects	5: F1 Subjects ANOVA	**6: F2 Item ANOVA**
Item Effect:								
2	2		0.03	0.09	0.03	0.09	0.09	0.03
2	10		0.05	0.14	0.05	0.12	0.15	0.05
10	2		0.04	0.32	0.04	0.31	0.32	0.04
10	10		0.05	0.31	0.05	0.29	0.33	0.05
Subject Effect:								
2	2		0.04	0.04	**0.12**	0.11	0.04	**0.12**
2	10		0.05	0.05	**0.34**	0.34	0.05	**0.36**
10	2		0.04	0.03	**0.12**	0.09	0.03	**0.12**
10	10		0.06	0.06	**0.34**	0.31	0.05	**0.37**

Random Slopes

- In addition to allowing each subject his or her own intercept for a mean difference, we can also test (using a −2LL LRT) whether subjects show individual differences in their effect of an item predictor → **random slope**

- For example: $\mathrm{RT_{tis}} = \gamma_{000} + \gamma_{010}\mathrm{A_i} + \gamma_{020}\mathrm{B_i} + \gamma_{030}\mathrm{A_i B_i}$
  $+\mathbf{U_{00s}} + \boxed{\mathbf{U_{01s}}\mathrm{A_i}} + \mathbf{U_{0i0}} + \mathbf{e_{tis}}$

  - The new $\boxed{\mathbf{U_{01s}}\mathrm{A_i}}$ term is a subject-specific deviation that creates a **subject-specific effect of item predictor A**
  - As with all random effects, we estimate its **variance** (as τ^2_{U01}) instead of the separate subject values—this variance can then **be predicted via interactions of A by subject predictors**, allowing us to test why some subjects show a stronger effect of the item predictor
  - It also creates heterogeneity of variance and covariance across outcomes as a function of the levels of the A predictor

- Random slopes of predictor effects over people are also technically possible (but harder to envision in practice)

Explanation of Random Effects Variances

- We can test the significance of a random intercept or slope variance, but the variances do not have inherent meaning
  - e.g., "I have a significant fixed effect of item predictor A of $\boldsymbol{\gamma_{010}}$ **= 70**, so the slope for predictor A is 70 on average. I also have a significant random slope variance of $\boldsymbol{\tau^2_{U_{01}}}$ **= 372**, so people need their own slopes for the effect of A. But how much is a variance of **372**, really?"
- **95% Random Effects Confidence Intervals** can tell you
  - Can be calculated for each effect <u>that is random</u> in your model
  - Provide range around the fixed effect within which 95% of your sample is predicted to fall, based on your random effect variance:

$$\text{Random Effect 95\% CI} = \text{fixed effect} \pm \left(1.96 * \sqrt{\text{Random Variance}}\right)$$

$$\text{Slope for A 95\% CI} = \gamma_{010} \pm \left(1.96 * \sqrt{\tau^2_{U_{10}}}\right) \rightarrow 70 \pm \left(1.96 * \sqrt{372}\right) = 32 \text{ to } 107$$

  - Predictor A has a positive slope = 70 on average, and people's individual slopes for A are predicted to range from 32 to 107 (the A effect varies)

Conclusions

- A RM ANOVA model may be less than ideal when:
  - Stimuli are not completely controlled or exchangeable
  - Experimental conditions are not strictly discrete
  - Missing data may result in bias, a loss of power, or both

- RM ANOVA is a special case of a more general family of multivariate/multilevel models (with nested or crossed effects as needed) that can offer additional flexibility:
  - Useful in addressing statistical problems →
    - Dependency, heterogeneity of variance, unbalanced or missing data
    - Examine predictor effects pertaining to each source of variation more accurately given that all variation is properly represented in the model
  - Useful in addressing substantive hypotheses →
    - Examining individual differences in effects of experimental manipulations