A Gentle Introduction to Mixed-Effects Models for Repeated Measures Designs

Lesa Hoffman
Professor, Educational Measurement and Statistics Program,
University of Iowa College of Education

Presented 12/3/21 to the Cognitive Science Program,
Department of Psychology, Mississippi State University

Slides available at: https://www.lesahoffman.com/Workshops/index.html

How could anyone be confused???

Btw, I prefer “multilevel model” (but never “multi-level” model)

And regularized regression is clearly the outlier here…

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 weighted function of its values of the predictor variables
    - Fixed slopes are **estimated constants** that multiply predictors

- **Model for the Variance:**
  - *Aka* **Random Effects and Residuals**, Stochastic Part of Model
  - What you are used to **making assumptions about** instead
  - How residuals are distributed and related across observations (persons, stimuli, occasions, etc.) → these relationships are called "dependency" and ***this is the primary way that mixed-effects models differ from general linear models*** *(e.g., regression)*

The Two Sides of a General Linear Model

$$y_i = \beta_0 + \beta_1(x_i) + \beta_2(z_i) + \cdots + e_i$$

Our focus

- **Model for the Means → Predicted Values:**
  - Each person's expected (predicted) outcome is a weighted linear function of his/her values on predictor variables x_i and z_i, each measured once per person (i.e., this is a univariate model)
  - **Estimated parameters are called fixed effects** (here, β_0, β_1, and β_2)

- **Model for the Variance:**
  - $e_i \sim \mathrm{N}(0, \sigma_e^2)$ → ONE source of residual (unexplained) error
  - In the GLM, e_i has a mean of 0 with a single estimated constant variance σ_e^2, is normally distributed, is unrelated to x_i and z_i, and is unrelated across all observations (which is just one per person here)
  - **Estimated parameter is ONE variance across persons** (not each e_i)
  - Proportion of variance reduced relative to *empty means* model = $\boldsymbol{R^2}$

Dimensions for Organizing Models

- Outcome type: General (normal) vs. General*ized* (not normal)
- Sampling dimensions: One (so one variance term per outcome) vs. **Multiple** (so multiple variance terms per outcome) → **OUR WORLD**

- **General Linear Models (regression, ANOVA):** conditionally normal outcome distribution, **fixed effects only** (identity link; only one sampling dimension)
- **General*ized* Linear Models: any conditional outcome distribution**, **fixed effects only** through **link functions**, no random effects (one dimension)
- **General Linear Mixed-Effects Models:** conditionally normal outcome distribution, **fixed + random effects** (identity link, for multiple sampling dimensions)
- **General*ized* Linear Mixed-Effects Models: any conditional outcome distribution**, **fixed + random effects** through **link functions** (for multiple sampling dimensions)

- "Linear" means fixed effects predict the *link-transformed* conditional outcome mean in a linear combination of (slope*predictor) + (slope*predictor)…
- "Nonlinear" could mean a deviation from that form OR a generalized model

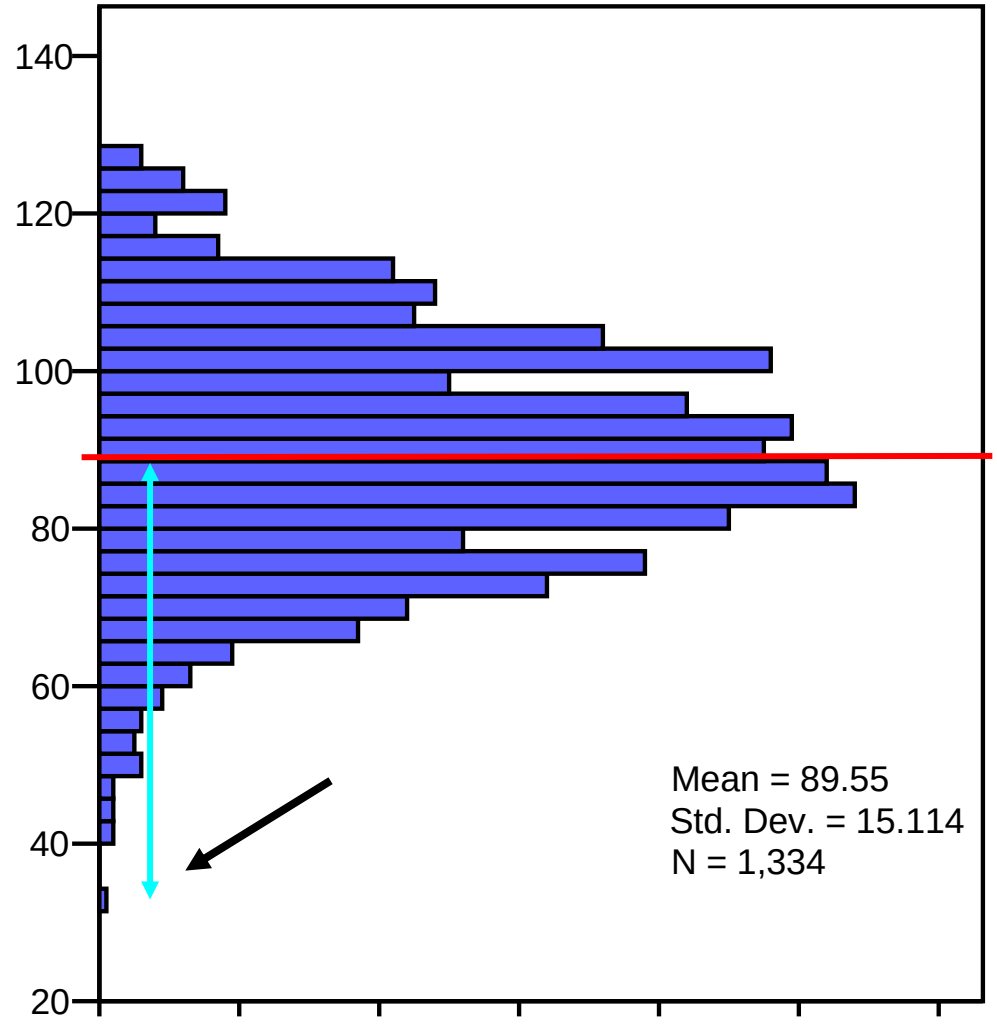
Who Run the World? Mixed-Effects Models!

Because **random effects** are the same thing as **latent variables**, these are all special cases of a more general system:

- Random Effects ANOVA or **Repeated Measures ANOVA**
- (Latent) Growth Curve Model *(where "Latent" implies the use of SEM software to estimate the same model)*
- Within-Person Fluctuation Model *(e.g., for EMA data or daily diary data or "intensive longitudinal data")*

- Clustered/Nested Observations Model *(e.g., for kids in schools)*
- **Crossed Random Effects**, Cross-Classified, or Multiple Membership Models *(e.g., for students who move, for teacher "value-added" status)*
- Psychometric Models *(e.g., for items nested in people)*
  - Yes, even measurement models! Confirmatory factor analysis, item factor analysis, item response theory, structural equation modeling…

An "Empty Means" General Linear Model for a **Single** Sampling Dimension (People)



$$y_i = \beta_0 + e_i$$

Filling in values:

$$32 = \underbrace{90}_{\hat{y}_i} + -58$$

$\hat{y}_i$ = "y-hat" model-predicted outcome

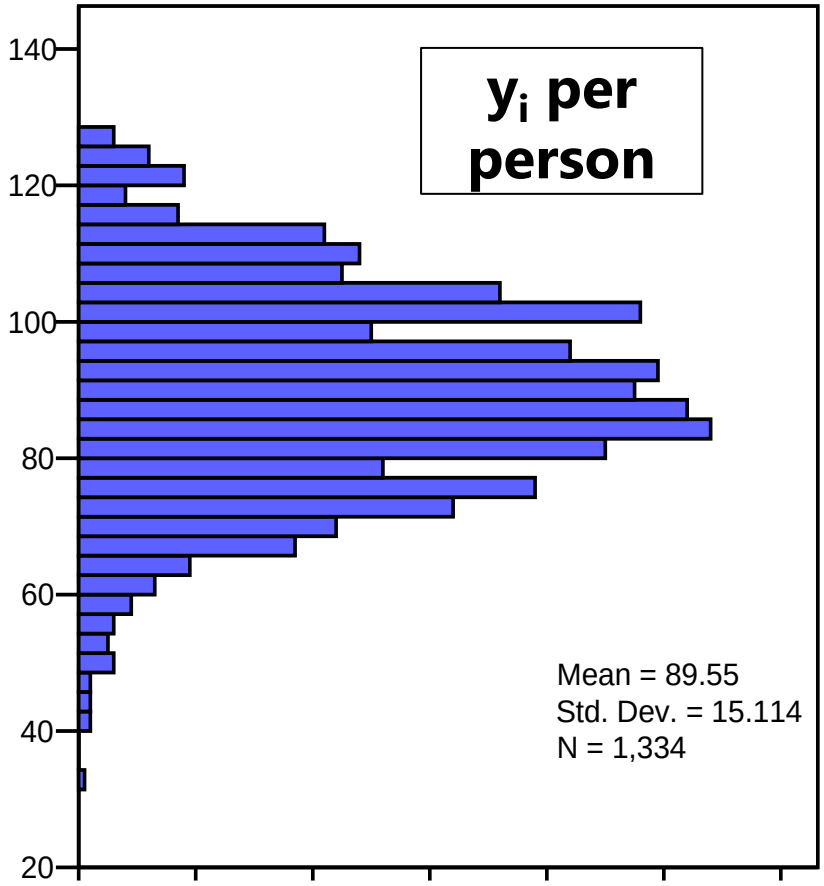
Model for the Means

y_i residual ("error") variance:

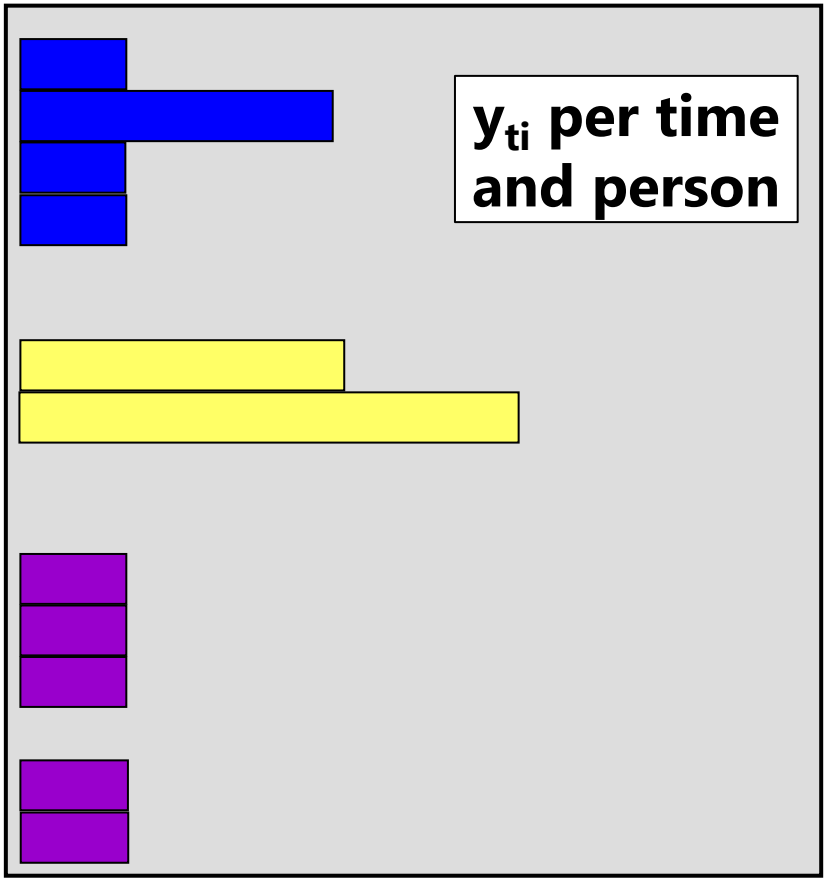
$$\frac{\sum(y_i - \hat{y}_i)^2}{N - 1}$$

Adding a **second** sampling dimension: Repeated measures occasions

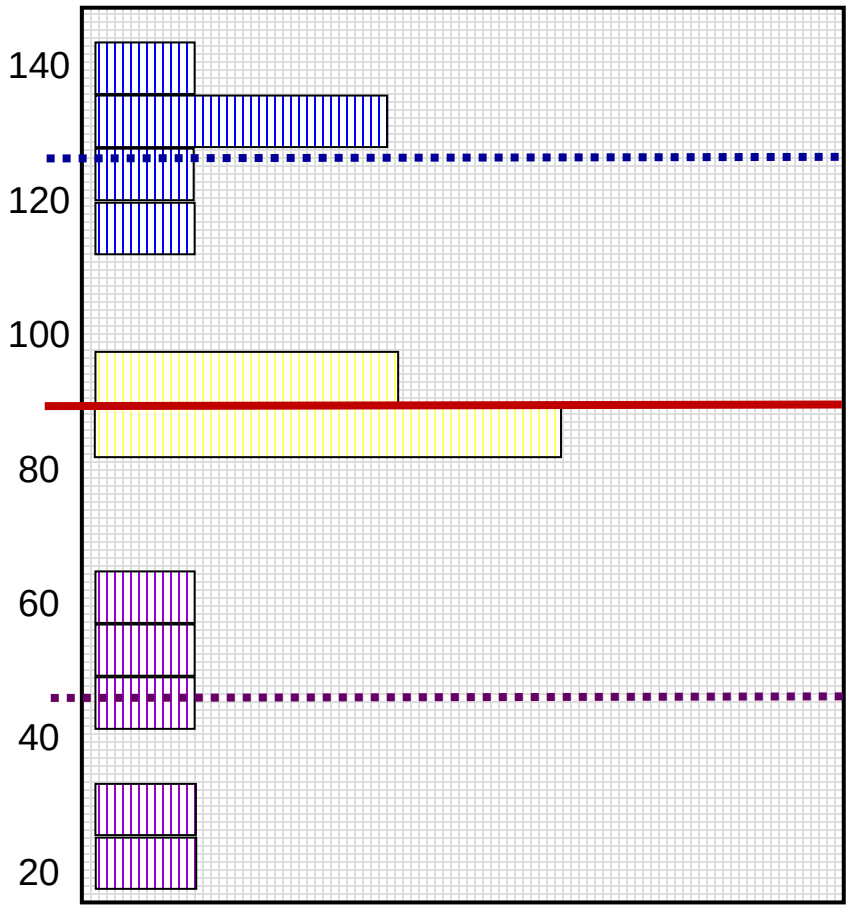
Full Sample Distribution



5 Occasions (t); 3 People (i)



Empty Means, Two-Level Model for the Variance (for Repeated Measures)



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

= Grand Mean

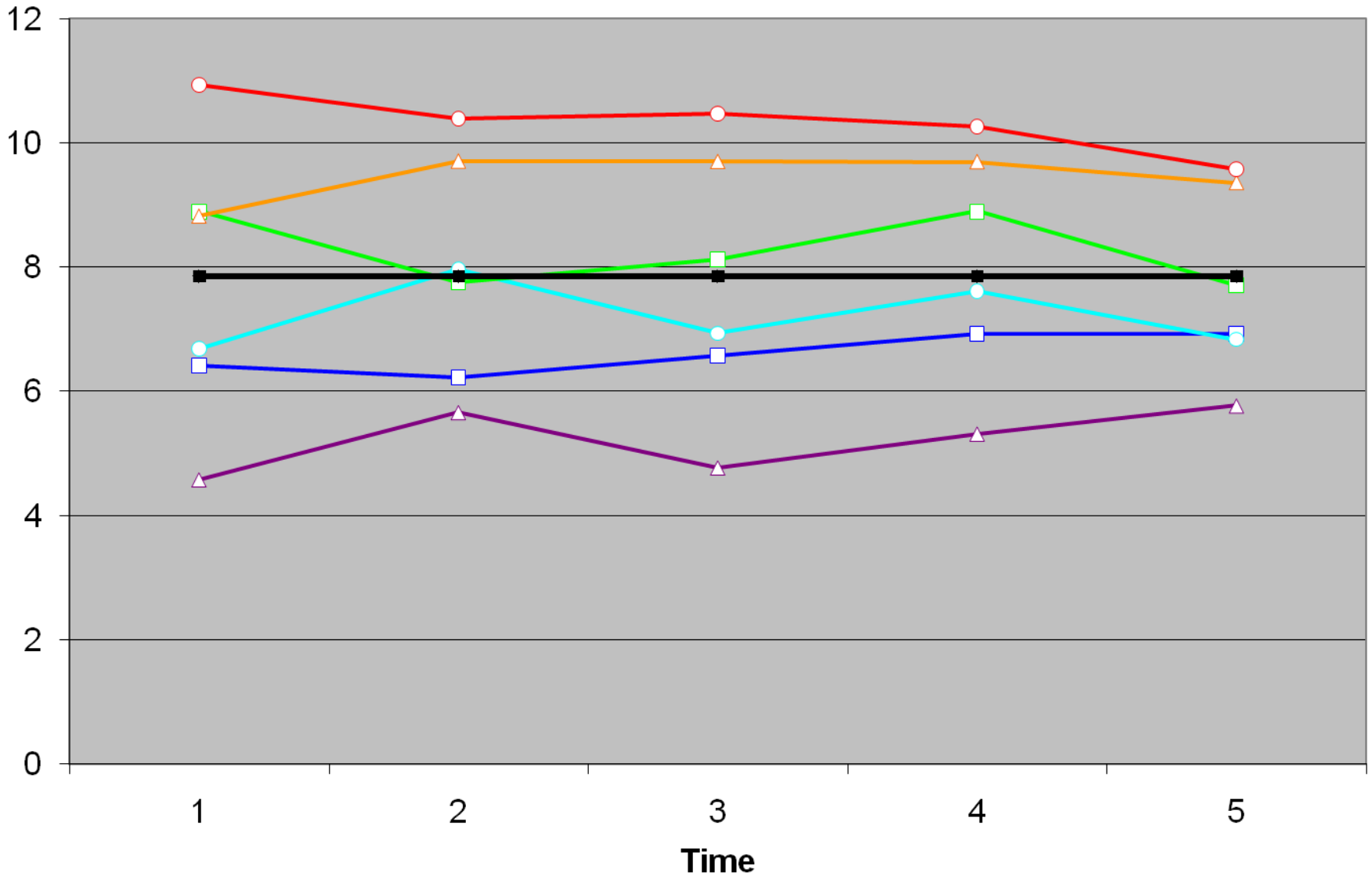
= Fixed Intercept

Can make better guess by taking advantage of repeated observations:

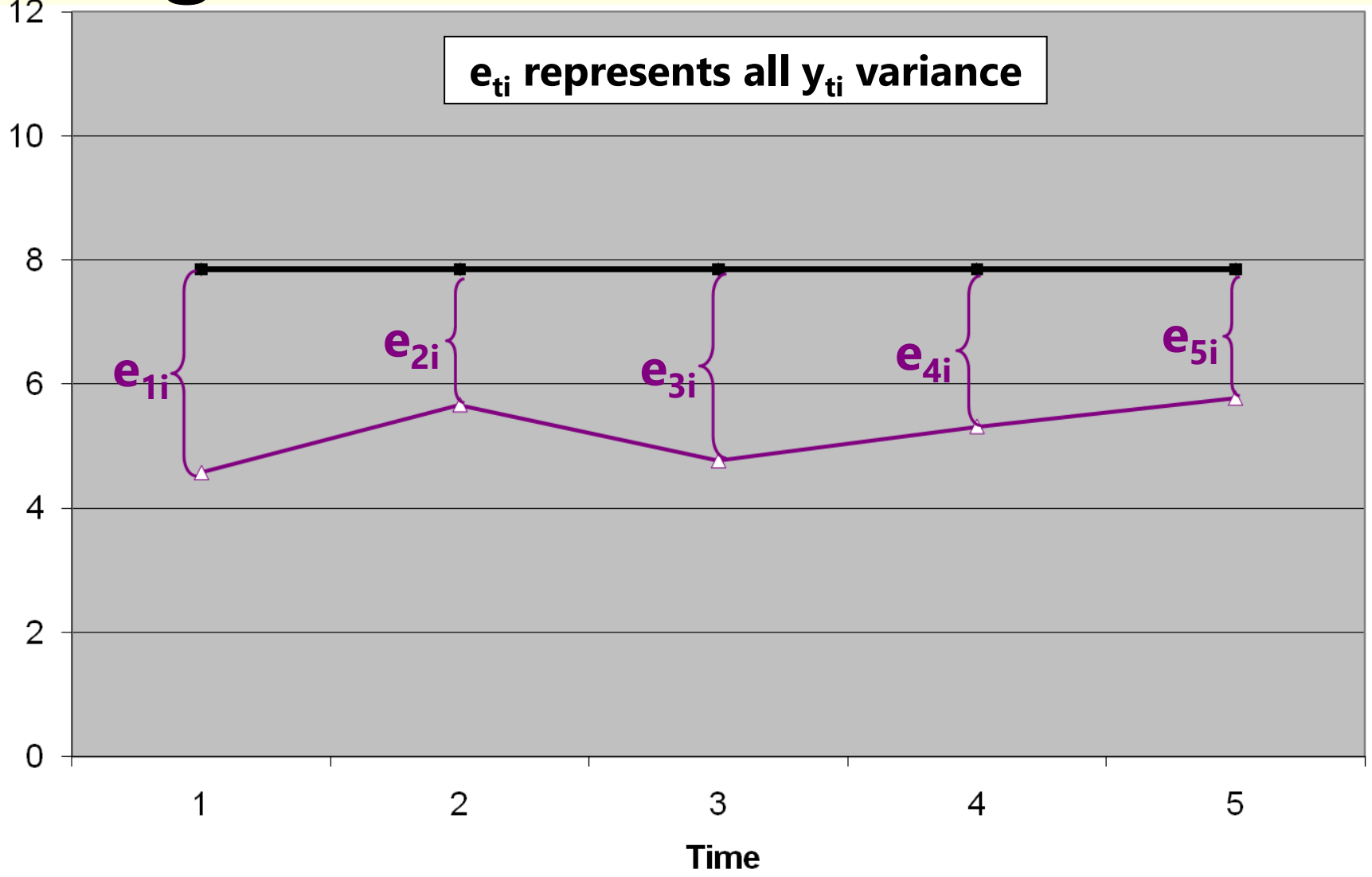
= Person Mean

→ Random Intercept

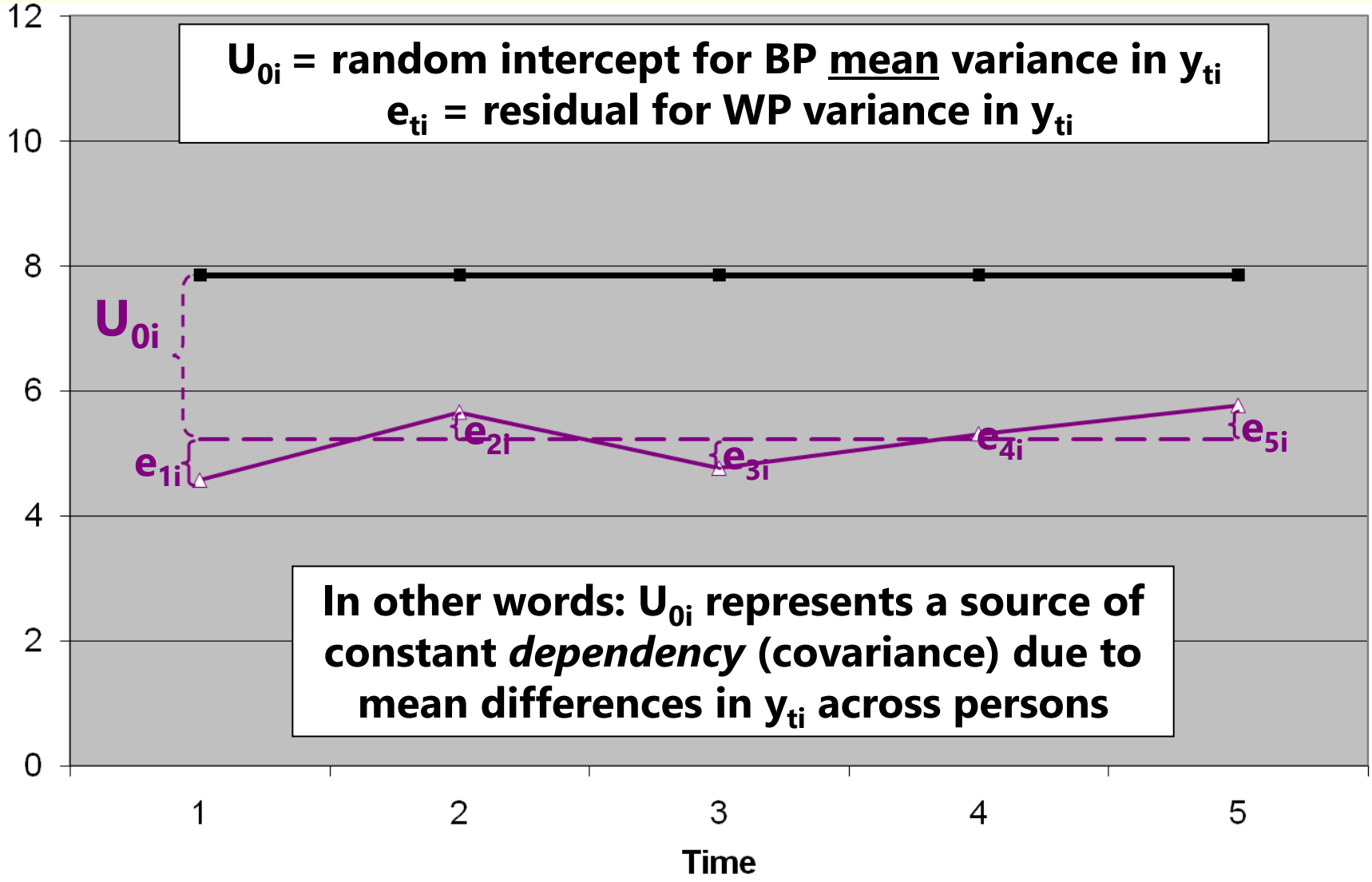
Hypothetical Repeated Measures Data



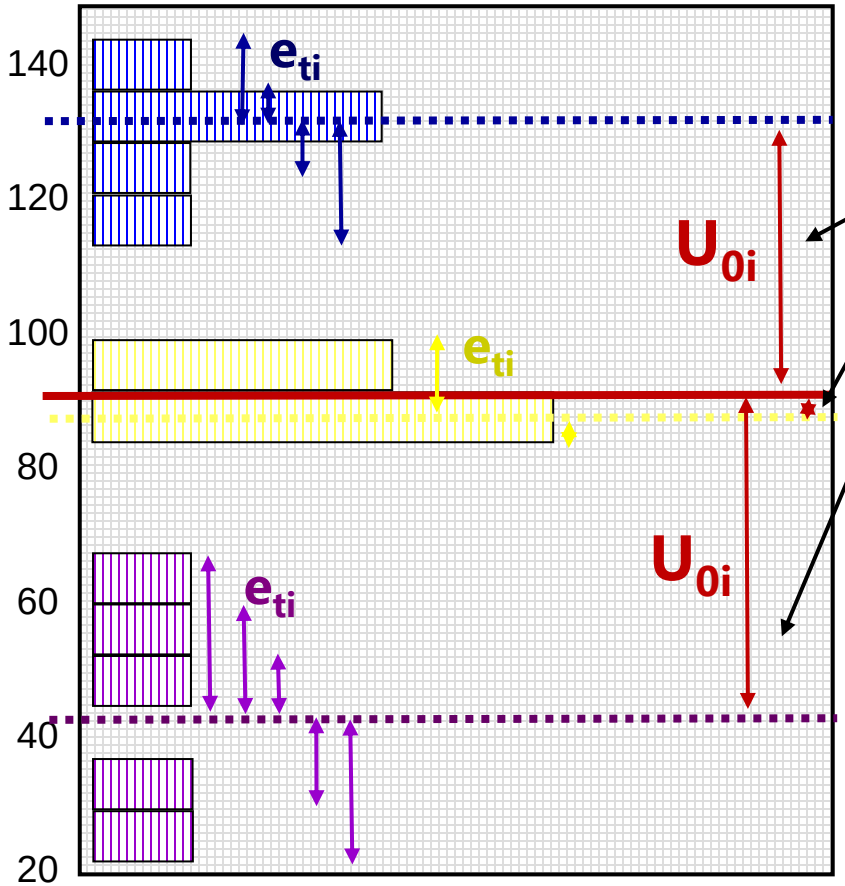
Comparison: "Error" in a **Single-Level** Model for the Variance



Comparison: "Error" in a **Two-Level** Model for the Variance



Empty Means, Two-Level Model (for the Variance)



y_{ti} variance → 2 sources:

Level 2 Random Intercept Variance (of U_{0i}, as τ^2_{U0}):

- → **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

Single- vs. Two-Level Conditional Models

- Univariate **Between-Subjects** ANOVA: **1 variance**
  - $y_i = (\beta_0 + \beta_1 x_i + \beta_2 z_i + \ldots) + e_i$
  - e_i → ONE residual, assumed uncorrelated with equal variance across observations (here, just persons) → "**BP (all) variation**"

- Univariate **Repeated Measures** ANOVA: **2 variances**
  - $y_{ti} = (\beta_0 + \beta_1 x_i + \beta_2 z_i + \ldots) + U_{0i} + e_{ti}$
  - U_{0i} → A random intercept for differences in person means, assumed uncorrelated with equal variance across persons → "**BP (mean) variation**"= $\tau^2_{U_0}$ is now "leftover" after predictors
  - 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**" = σ^2_e is also now "leftover" after predictors

The Curse of Non-Exchangeable Items

Jim Bovaird, University of Nebraska-Lincoln

Larry Locker, Georgia Southern University

Downtown Lawrence, KS

- Psycholinguistic research (trials are words and non-words)
  - Common persons, common trials designs
  - 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… F1 or F2 (or both)?

Two Kinds of ANOVAs using Summary Data

By convention, fixed effects change from β to γ

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

"F1" RM ANOVA on *N* persons:

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

"F2" Between-Subjects ANOVA on *T* trials:

$$RT_t = \gamma_0 + \gamma_1 A_t + \gamma_2 B_t + \gamma_3 A_t B_t + e_t$$

Person Summary Data

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

Trial Summary Data

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

Choosing Amongst ANOVA Models

- F1 Repeated Measures ANOVA on **person** summary data:
  - Assumes trials are fixed—within-condition ***trial*** variability is gone
- F2 Between-Subjects ANOVA on **trial** summary data:
  - Assumes persons are fixed—within-trial ***person*** variability is gone

- Proposed ANOVA-based resolutions:
  - **F'** → quasi-F test that treats both trials and persons 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
  - **F1 x F2 criterion** → effects are only "real" if they are significant in **both F1 and F2 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 (stay tuned for more):

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

Cons:

- **Is still wrong**

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

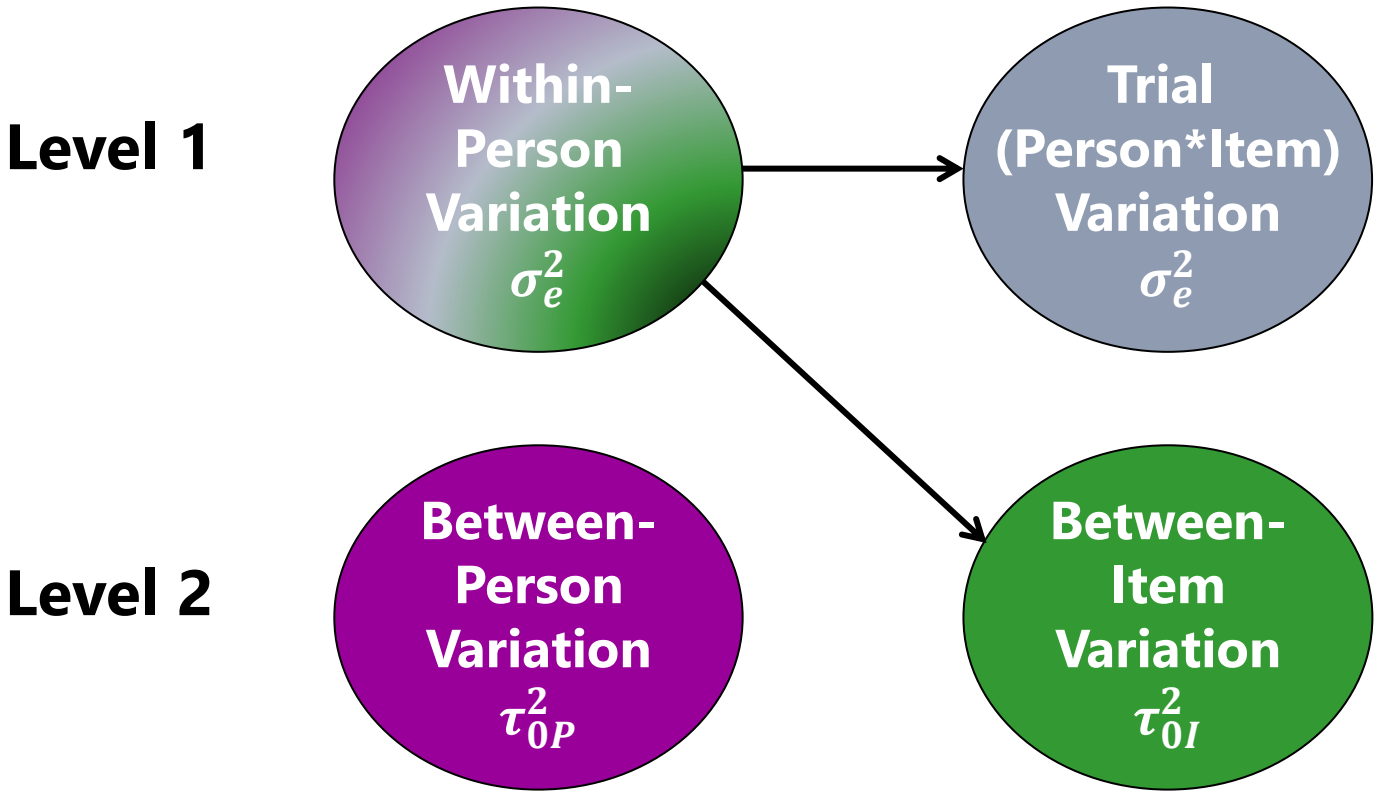
Level 2:

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

Level 1 = Within-Person Variation (Across Trials)

Level 2 = Between-Person Variation

Multilevel Models: A New Way of Life?



Empty Means, Crossed Random Effects Models

- **Residual-only model:**
  - $RT_{tpi} = \gamma_{000} + e_{tpi}$
  - Assumes no effects (dependency) of subjects or items

- **Random persons (or "subjects") <u>nested</u> model:**
  - $RT_{tpi} = \gamma_{000} + U_{0p0} + e_{tpi}$
  - Adds systematic mean differences **between persons**

- **Random persons and items <u>crossed</u> model:**
  - $RT_{tpi} = \gamma_{000} + U_{0p0} + U_{00i} + e_{tpi}$
  - Adds systematic mean differences **between items**

A Better Way of (Multilevel) Life

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

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

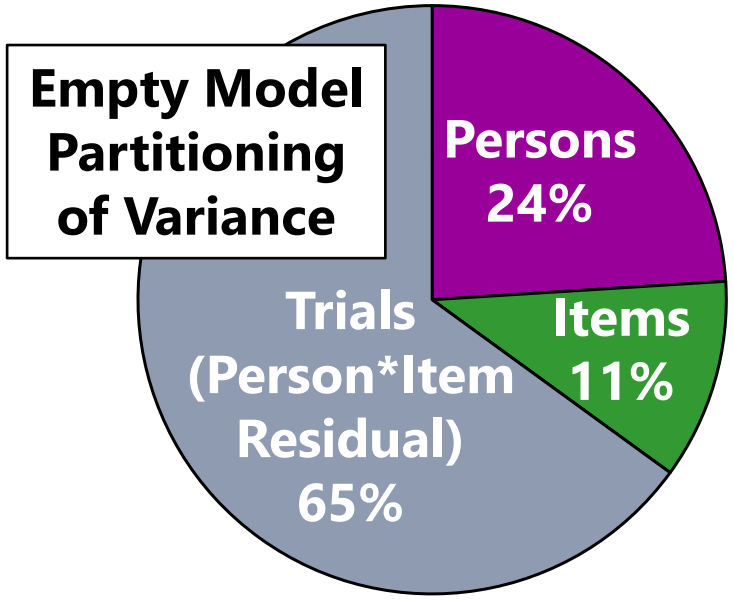
$$RT_{tpi} = \gamma_{000} + \gamma_{001}A_i + \gamma_{002}B_i + \gamma_{003}A_iB_i + \boldsymbol{U_{0p0}} + \boldsymbol{U_{00i}} + \boldsymbol{e_{tpi}}$$

t trial
p person
i item

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

Example 1: Larry's Psycholinguistics Data

- Crossed design: 38 persons see 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



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:

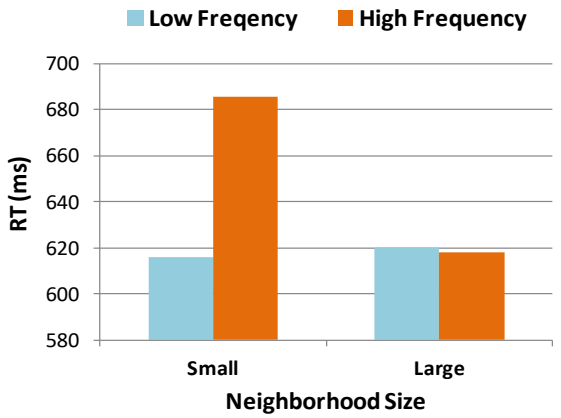
Residual ≈ 0%

Person ≈ 0%

Items ≈ 30%*

Total R^2 ≈ 3.3%

***Significant item 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 Person 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	F (1, 32) = 5.4 p = .0272	F (1, 32) = 4.6 p = .0393	F (1, 32) = 6.0 p = .0199

So when does it matter???

Example 2: Visual Search for Change

- Outcome (DV)
  - Natural log RT to detect a change (up to 60 seconds)
  - 51 out of 80 natural scenes (as items) with > 90% accuracy
- Between-Subjects IV
  - Age: Younger (n = 96) vs. Older (n = 57) Adults
- Within-Subjects IVs
  - Change Relevance to Driving (Low vs. High)
  - Change Salience (Low vs. High)
- Original Analysis Plan
  - 2 x 2 x 2 mixed design (split-plot) ANOVA predicting RT

Analysis Plan, Reconsidered
Issue #1: Systematic Item Differences

Can you find the change?

- Collapsing across scenes into condition means ignores systematic differences between scenes
- Treats scenes as fixed effects → F_1 ANOVA problem
  - Scenes will still vary in difficulty due to uncontrolled factors
  - Results may be optimistic if that variability is not included
- ANOVA via least squares requires complete data to include random variation across persons and scenes simultaneously…

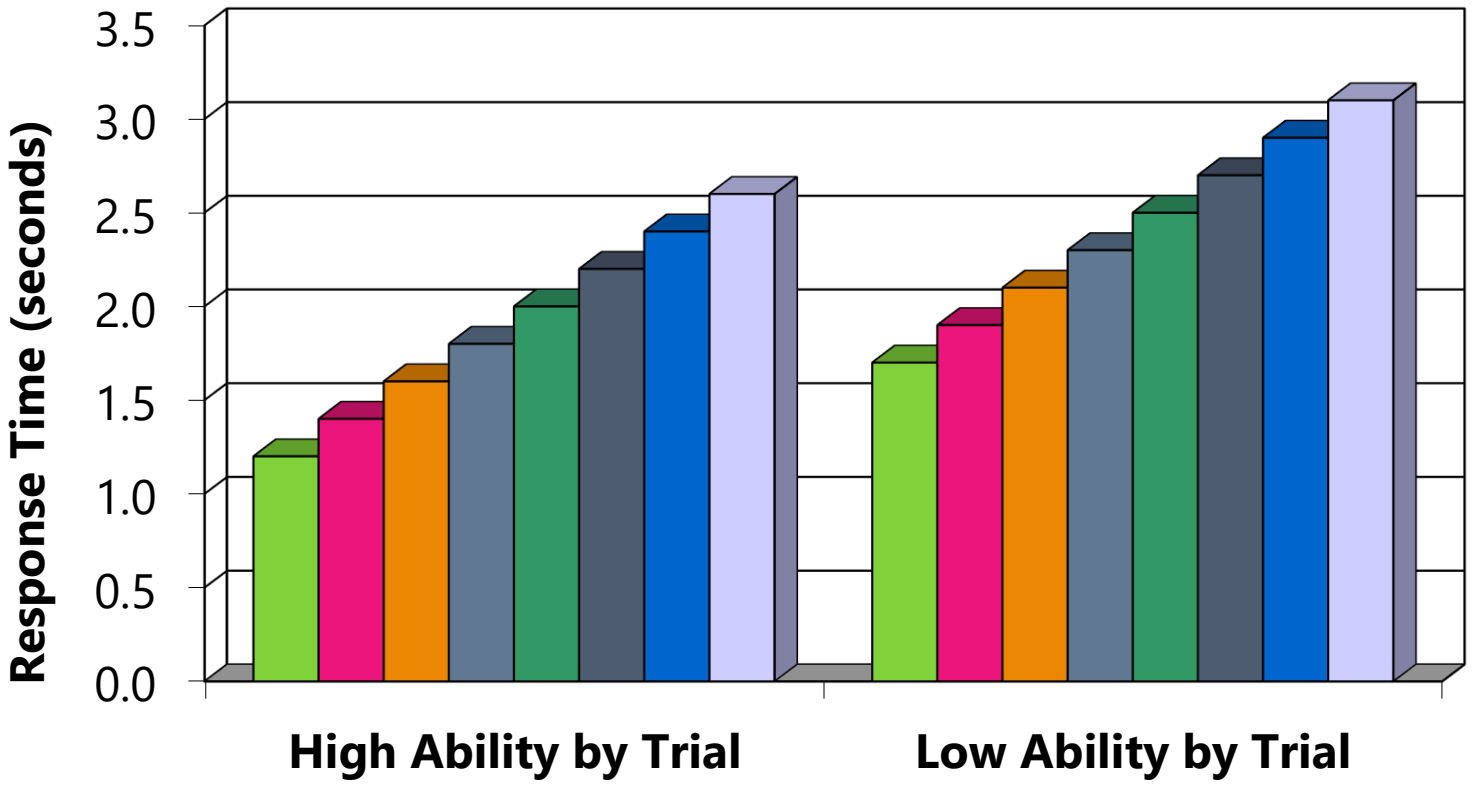
Analysis Plan, Reconsidered
Issue #2: Missing RTs for Incorrect Trials

- Any changes not detected within 60 sec were "inaccurate"
- Only scenes with > 90% accuracy were included, but…

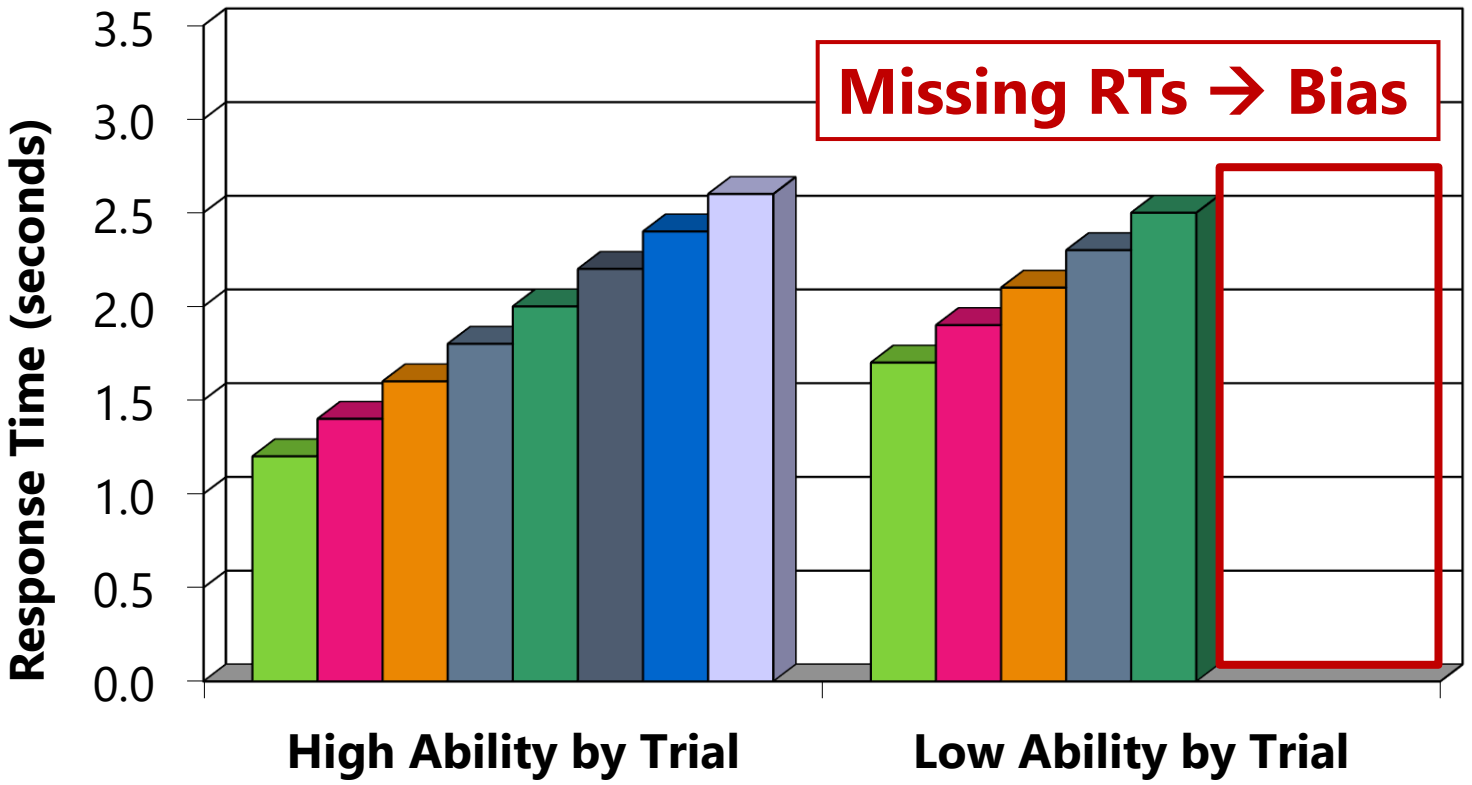
- RTs are more likely to be missing for difficult scenes
  - Downwardly biased condition mean RTs
  - Biased effects of predictor variables related to missingness
  - Loss of power due to listwise deletion

- ANOVA assumes RTs are missing completely at random, but an assumption of missing at random is more tenable
  - Missing at Random → probability of missingness is unrelated to unobserved outcome *after* predictors and observed responses are included in the model… possible by switching to likelihood estimation

Original RTs Across Trials by Ability



Biased Condition Mean RT



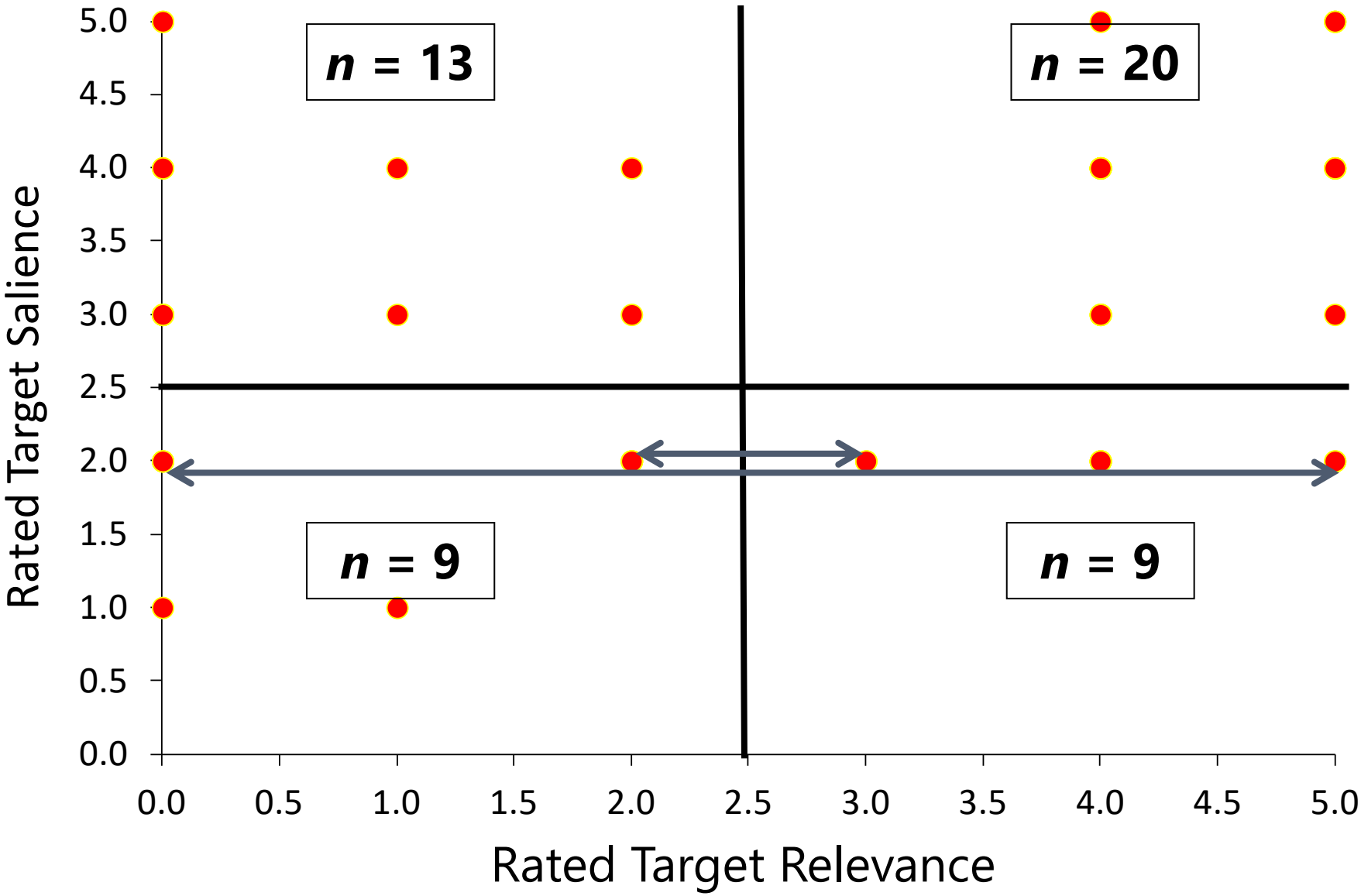
Analysis Plan, Reconsidered
Issue #3: Effects of Item Predictors

- 51 scenes (items) varied in change relevance and salience

- Relevance and salience were separately rated for each scene on a continuous scale of 0-5
  - Relevance and salience r = .22
  - Median splits formed categories of "low" & "high"
  - Uneven number of scenes per "condition" by design (and because of timed-out inaccurate trials)

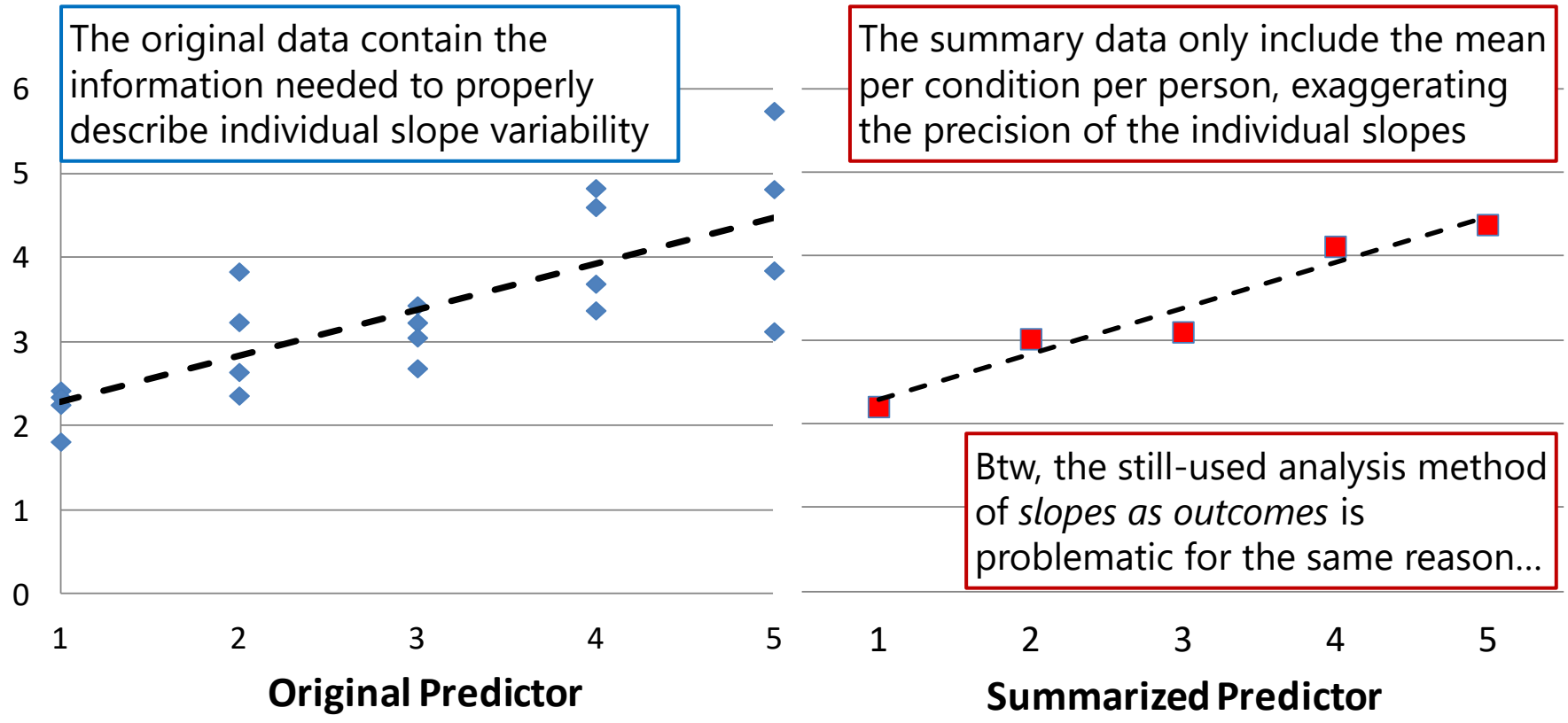
- Predictors of relevance and salience should be treated as continuous, which is not possible with RM ANOVA (that requires discrete conditions as outcomes)

Creating "Conditions" (r = .22 → r ≈ 0)



Individual Differences in Predictor Effects?

- RM ANOVA allows individual differences in means (intercepts) only
- …What about variability in the ***effects of item manipulations***?

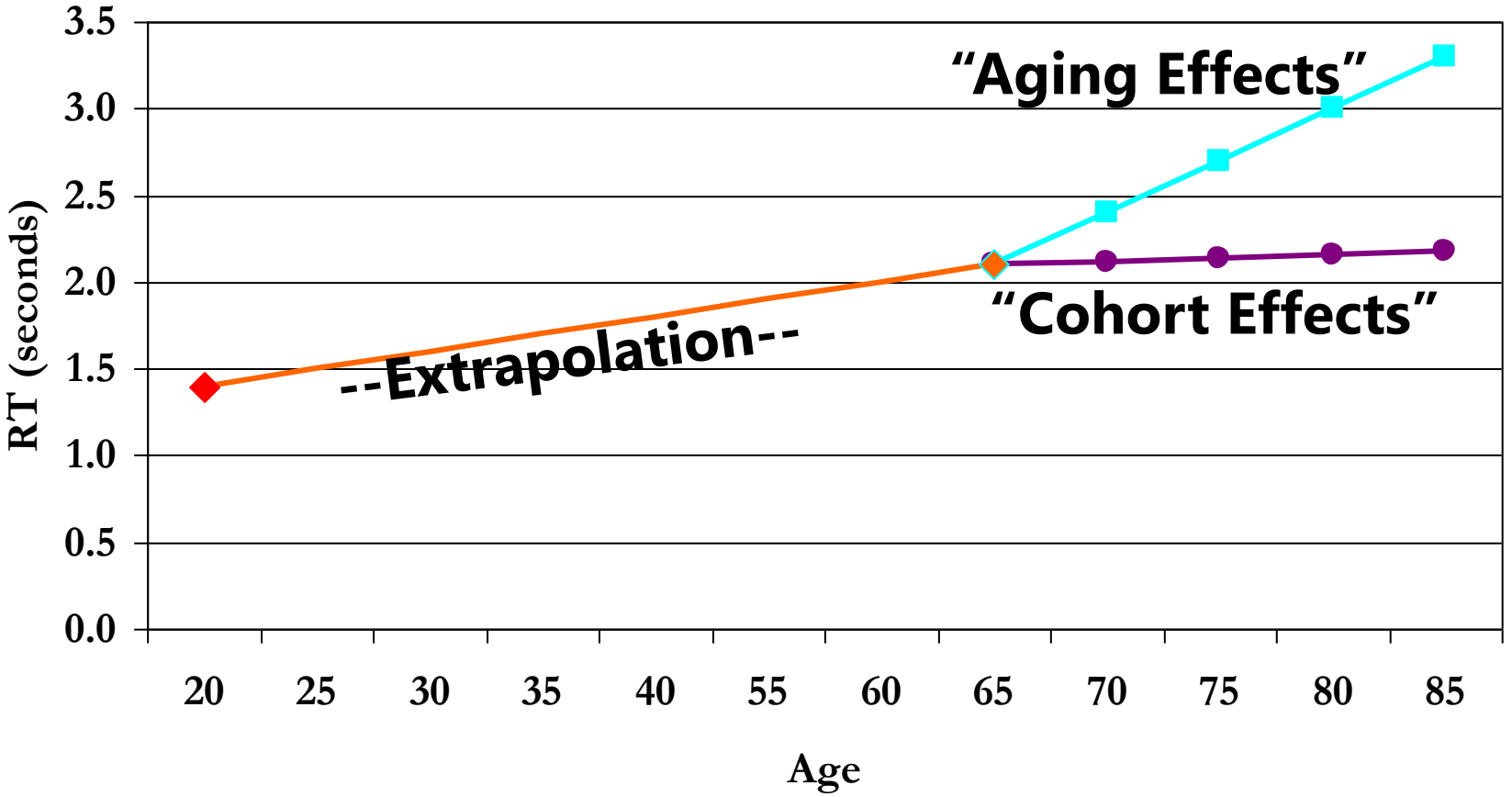


Analysis Plan, Reconsidered
Issue #4: Age Differences in Means

- "Younger" and "Older" adults were sampled, but…
  - Much more variability in age in the older group
    - 18-32 years (mostly 18-21) vs. 65-86 years
  - Age in this design is not a strict dichotomy:
    - Including a single mean age group difference is not adequate
    - Separating "young-old" from "old-old" doesn't really help, either

- Two fixed effects of age are needed:
  - "Age Group" → difference between young and old
  - "Years over 65" → slope of age in the older group
  - This can be created using "piecewise" linear slopes
  - (Could also be done using RM AN**C**OVA)

Piecewise (Semi-Continuous) Effects of Age on RT



Analysis Plan, Reconsidered
Issue #5: Age Differences in Variance

- In addition to modeling differences in RT means by age, the **RT variances** are likely to differ by age as well:
  - Older adults are likely to be more different *from each other* than are younger adults
  - Older adults are likely to be more inconsistent *across trials* than are younger adults

- The model needs to accommodate heterogeneity of variance across age groups at multiple levels of analysis
  - "**Location–scale**" variants of mixed-effects models can do this! (see work by Don Hedeker at the University of Chicago)

Summary: Mixed-Effects Models

- **ANOVAs on summary data can be problematic:**
  - Ignoring non-randomly missing responses; discretizing item predictors
  - Significance and effect sizes of item-specific predictors will be distorted if items are not exchangeable but they are modeled that way
    - Relevant whenever *all* item variation isn't accounted for by fixed effects
    - But "exchangeable enough" is always an empirical question!

- **Mixed-effects models** also provide a way to quantify and **predict individual differences** in cognitive processes, such as…
  - Changes in eye movements by cognitive task during scene viewing
    Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd (2011, *Journal of Vision*)
  - Executive function and semantic processing in verbal fluency
    McDowd, Hoffman, Rozek, Lyons, Pahwah, Burns, & Kemper (2011, *Neuropsychology*)
  - Dual-task costs for mouse-tracking during natural speech production
    Kemper, Hoffman, Schmalzried, Herman, & Kieweg (2011, Aging, *Neuropsychology, and Cognition*)
  - Btw, these fixed effects are known as "cross-level interactions"

Thank you! Here are some references:

- Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39*(1), 101-117. Electronic appendix available at Lesa's website.

- Kemper, S., Hoffman, L., Schmalzried, R., Herman, R., & Kieweg, D. (2011). Tracking talking: Dual task costs of planning and producing speech for young versus older adults. *Aging, Neuropsychology, and Cognition, 18*(3), 257-279.

- Locker Jr., L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling in the analysis of psycholinguistic data. *Behavior Research Methods, 39*(4), 723-730. Electronic appendix available at Lesa's website.

- McDowd, J. M., Hoffman, L., Rozek, E., Lyons, K., Pahwa, R., Burns., J., & Kemper, S. (2011). Understanding verbal fluency in healthy aging, Alzheimer's disease, and Parkinson's disease. *Neuropsychology, 25*(2), 210-225. PMID: 21381827.

- Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision, 11*(8), 1-15.

- Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York, NY: Routledge/Taylor & Francis Group.