# Multilevel Models for Subjects Crossed with Items: Motivation and Examples

- Topics:
  - The experimental psychologist's analytic toolbox
  - Examples of crossed random effects models:
    - 1: Psycholinguistic study (subjects by words)—see article & 945 Ex. 3a
    - 2: Visual search study (subjects by scenes)—chapter 12
    - 3: Eye tracking study (subjects by scenes)—see article
  - Example of nested model:
    - 4: Tracking and talking (speech within subjects)—see article

# Analytic Toolbox of the Experimental Psychologist

- Our friend, analysis of variance (ANOVA)
  - Between-group (*aka* between-subject, independent IV)
  - Within-group (*aka* within-subject, dependent, **repeated measures** IV)
  - Split-plot (*aka* mixed design of between- and within-group IVs)

- Expandable to include:
  - multiple IVs (factorial ANOVA)
  - main effects of continuous covariates (ANCOVA)
  - multiple outcomes (MANOVA/MANCOVA)

# RM 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)?

# 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 subjects, 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**?

# Example 1: Overview of Psycholinguistic Study Design

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

- Tests of effects of experimental treatment are typically conducted with the subject 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 | Item 001<br>Item 002<br>.........<br>Item 100 | Item 101<br>Item 102<br>.........<br>Item 200 |
| A2 | Item 201<br>Item 202<br>.........<br>Item 300 | Item 301<br>Item 302<br>.........<br>Item 400 |

**"F1" Within-Subjects ANOVA on *N* subjects:**

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

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

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

**Subject 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(Subject 1, Subject 2,... Subject *N*)<br>Item 002 = Mean(Subject 1, Subject 2,... Subject *N*)<br>......... Item 100 |
| A1, B2 | Item 101 = Mean(Subject 1, Subject 2,... Subject *N*)<br>Item 102 = Mean(Subject 1, Subject 2,... Subject *N*)<br>......... Item 200 |
| A2, B1 | Item 201 = Mean(Subject 1, Subject 2,... Subject *N*)<br>Item 202 = Mean(Subject 1, Subject 2,... Subject *N*)<br>......... Item 300 |
| A2, B2 | Item 301 = Mean(Subject 1, Subject 2,... Subject *N*)<br>Item 302 = Mean(Subject 1, Subject 2,... Subject *N*)<br>......... Item 400 |

# Choosing Amongst ANOVA Models

- **F1** Within-Subjects ANOVA on **subject** summary data:
  - Within-condition ***item*** variability is gone, so items assumed fixed
- **F2** Between-Items ANOVA on **item** summary data:
  - Within-item ***subject*** variability is gone, so subjects assumed fixed

- Historical proposed ANOVA-based resolutions:
  - **F′** → quasi-F test with random effects for both subjects and items (Clark, 1973), but requires complete data (**uses 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** × **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)…

# Sources of Variance (Clark, 1973)
## *t = #conditions, i = #items, s = #subjects*

| Label | | DF | Expected Mean Square |
|---|---|---|---|
| T | Treatments (t) | t−1 | $\sigma_e^2 + \sigma_{S\times I}^2 + i\sigma_{T\times S}^2 +$ ___ $+ s\sigma_I^2 + is\sigma_T^2$ |
| I w T | Items (i) within Treatments | t(i−1) | $\sigma_e^2 + \sigma_{S\times I}^2 +$ ___ $+$ ___ $+ s\sigma_I^2 +$ ___ |
| S | Subjects (s) | s−1 | $\sigma_e^2 + \sigma_{S\times I}^2 +$ ___ $+ t\sigma_S^2 +$ ___ $+$ ___ |
| T × S | Treatments by Subjects | (t−1)(s−1) | $\sigma_e^2 + \sigma_{S\times I}^2 + i\sigma_{T\times S}^2 +$ ___ $+$ ___ $+$ ___ |
| S × I w T | Subjects by Items within Treatments | t(i−1)(s−1) | $\sigma_e^2 + \sigma_{S\times I}^2 +$ ___ $+$ ___ $+$ ___ $+$ ___ |

# Effect of Treatment via $F_1$ ANOVA

*T numerator should differ from T×S denominator by 1 term*

| Label | | DF | Expected Mean Square |
|---|---|---|---|
| T | Treatments (t) | t−1 | $\sigma^2_e + \sigma^2_{S\times I} + i\sigma^2_{T\times S} +$ ___ $+ s\sigma^2_I + is\sigma^2_T$ |
| I w T | Items (i) within Treatments | t(i−1) | $\sigma^2_e + \sigma^2_{S\times I} +$ ___ + ___ $+ s\sigma^2_I +$ ___ |
| S | Subjects (s) | s−1 | $\sigma^2_e + \sigma^2_{S\times I} +$ ___ $+ t\sigma^2_S +$ ___ + ___ |
| T × S | Treatments by Subjects | (t−1)(s−1) | $\sigma^2_e + \sigma^2_{S\times I} + i\sigma^2_{T\times S} +$ ___ + ___ + ___ |
| S × I w T | Subjects by Items within Treatments | t(i−1)(s−1) | $\sigma^2_e + \sigma^2_{S\times I} +$ ___ + ___ + ___ + ___ |

# Effect of Treatment via $F_2$ ANOVA

*T numerator should differ from I×T denominator by 1 term*

| Label | | DF | Expected Mean Square |
|---|---|---|---|
| T | Treatments (t) | t−1 | $\sigma_e^2 + \sigma_{S\times I}^2 + i\sigma_{T\times S}^2 +$ ___ $+ s\sigma_I^2 + is\sigma_T^2$ |
| I w T | Items (i) within Treatments | t(i−1) | $\sigma_e^2 + \sigma_{S\times I}^2 +$ ___ + ___ $+ s\sigma_I^2 +$ ___ |
| S | Subjects (s) | s−1 | $\sigma_e^2 + \sigma_{S\times I}^2 +$ ___ $+ t\sigma_S^2 +$ ___ + ___ |
| T × S | Treatments by Subjects | (t−1)(s−1) | $\sigma_e^2 + \sigma_{S\times I}^2 + i\sigma_{T\times S}^2 +$ ___ + ___ + ___ |
| S × I w T | Subjects by Items within Treatments | t(i−1)(s−1) | $\sigma_e^2 + \sigma_{S\times I}^2 +$ ___ + ___ + ___ + ___ |

# Simultaneous Quasi-F Ratio (F′)

- F′ was proposed by Clark (1973) as a quasi-F test that treats both items and subjects as random factors

$$F'\left(df_{num}, df_{den}\right) = \frac{MS_T + MS_{SxI}}{MS_{TxS} + MS_I}$$

$$\text{where } df_{num} = \frac{\left(MS_T + MS_{SxI}\right)^2}{\frac{MS_T}{df_T} + \frac{MS_{SxI}}{df_{SxI}}} \text{ and } df_{den} = \frac{\left(MS_{TxS} + MS_I\right)^2}{\frac{MS_{TxS}}{df_{TxS}} + \frac{MS_I}{df_I}}$$

$$F'\left(df_{num}, df_{den}\right) = \frac{\left(2*\sigma_e^2\right) + \left(2*\sigma_{SxI}^2\right) + \left(\#I*\sigma_{TxS}^2\right) + \left(\#S*\sigma_I^2\right) + \left(\#I*\#S*\sigma_T^2\right)}{\left(2*\sigma_e^2\right) + \left(2*\sigma_{SxI}^2\right) + \left(\#I*\sigma_{TxS}^2\right) + \left(\#S*\sigma_I^2\right)}$$

- Numerator then exceeds the denominator by exactly the treatment variance as desired… except it requires complete data given that it relies on ordinary least squares
  - Not feasible in most real-world experiments

# Minimum of Quasi-F Ratio (Min F′)

- Min F′ was developed to be used from $F_1$ and $F_2$ results:

$$\min F'\left(df_{num}, df_{den}\right) = \frac{MS_T}{MS_{TxS} + MS_I} = \frac{F_1 * F_2}{F_1 + F_2}$$

- But given that Min F′ is overly conservative, having to show significance by both models is often required instead:
  - the $F_1$ by $F_2$ criterion… but two wrongs don't make a right

- Wouldn't it be nice if we had some way to treat subjects and items as the random effects they actually are???
  - And to assess the extent to which items are actually exchangeable?
  - And that all the extraneous item variables were adequately controlled?
  - **Multilevel models to the rescue! … maybe?**

# Multilevel Model (MLM) Word Salad

- MLM is the same as other terms you have heard of:
  - **Linear Mixed-Effects Model** (fixed + random effects, of which intercepts and slopes are specific kinds of effects)
  - **Random Coefficients Model** (because coefficients also = effects)
  - **Hierarchical Linear Model** (not same as hierarchical regression)
- Special cases of MLM:
  - Random Effects ANOVA or Repeated Measures ANOVA
  - (Latent) Growth Curve Model (where "Latent" implies SEM software)
    - Btw, most MLMs can be equivalently estimated as single-level SEMS
  - Within-Person Fluctuation Model (e.g., for EMA or daily diary data)
    - See also "dynamic" SEM or multilevel SEM (even without measurement models!)
  - Clustered/Nested Observations Model (e.g., for kids in schools)
    - If followed over time in same group, is "clustered longitudinal model"
  - Cross-Classified Models (e.g., teacher "value-added" models)
  - Psychometric Models (e.g., factor analysis, item response theory, SEM)

# Multilevel Models to the Rescue?

**Original Data per Subject**

| | B1 | B2 |
|---|---|---|
| A1 | Item 001<br>Item 002<br>.........<br>Item 100 | Item 101<br>Item 102<br>.........<br>Item 200 |
| A2 | Item 201<br>Item 202<br>.........<br>Item 300 | Item 301<br>Item 302<br>.........<br>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: $y_{is} = \beta_{0s} + \beta_{1s}A_{is} + \beta_{2s}B_{is} + \beta_{3s}A_{is}B_{is} + e_{is}$

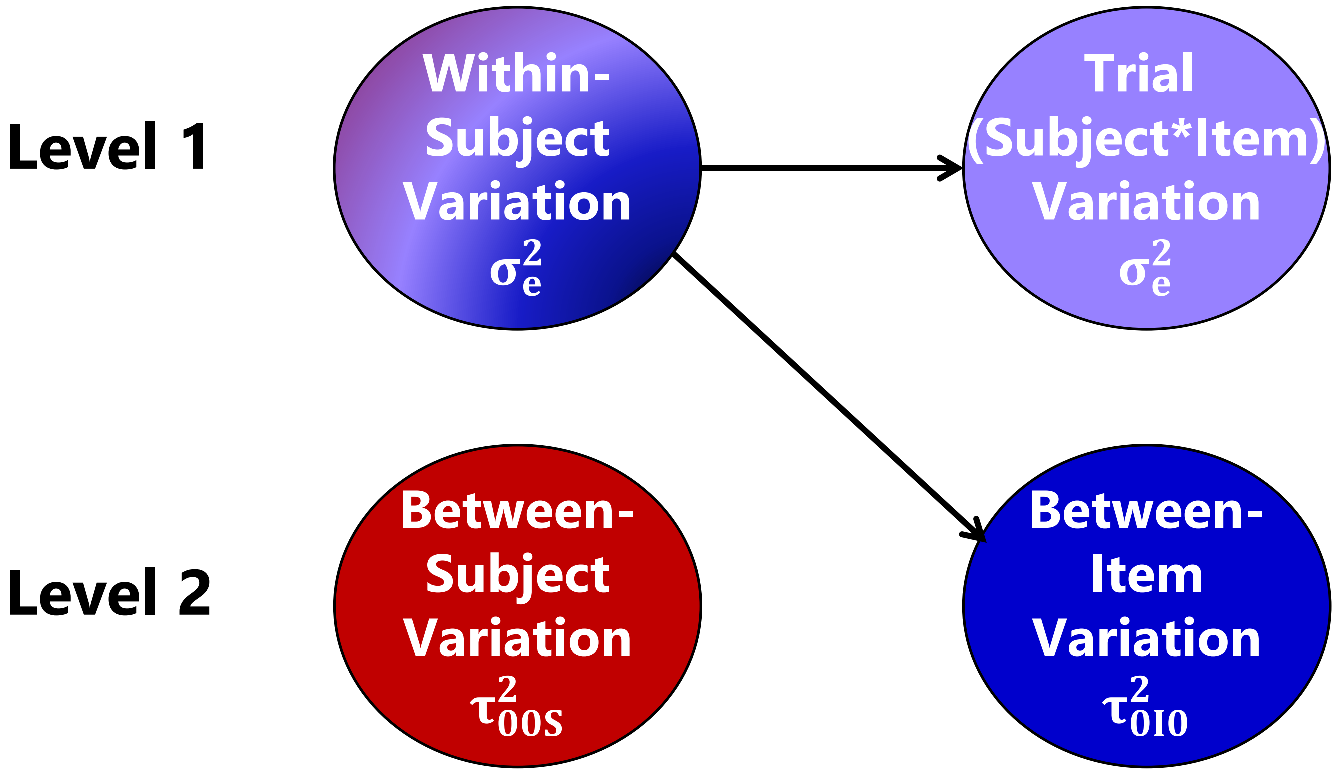
Level 2:

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

Level 1 = Within-Subject Variation (Across Items)

Level 2 = Between-Subject Variation

# Multilevel Models to the Rescue?



# Empty Means, Crossed Random Effects Models

- **Residual-only model:**
  - $\mathbf{RT_{tis}} = \gamma_{000} + \mathbf{e_{tis}}$
  - Assumes no dependency (correlation) of trials from the same subjects or the same items

- **Random person (or "subject") intercept:**
  - $\mathbf{RT_{tis}} = \gamma_{000} + \mathbf{U_{00s}} + \mathbf{e_{tis}}$
  - Includes systematic mean differences **between subjects** (which allows a correlation of trials from the same subject)

- **Random person and item intercepts:**
  - $\mathbf{RT_{tis}} = \gamma_{000} + \mathbf{U_{00s}} + \mathbf{U_{0i0}} + \mathbf{e_{tis}}$
  - Also includes systematic mean differences **between items** (which allows a correlation of trials from the same item, too)

# A Better Way of (Multilevel) Life

**Between-Subject Variation L2 $\tau^2_{00S}$**

**Between-Item Variation L2 $\tau^2_{0I0}$**

**Trial (Subject*Item) Variation $\sigma^2_e$**

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

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

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

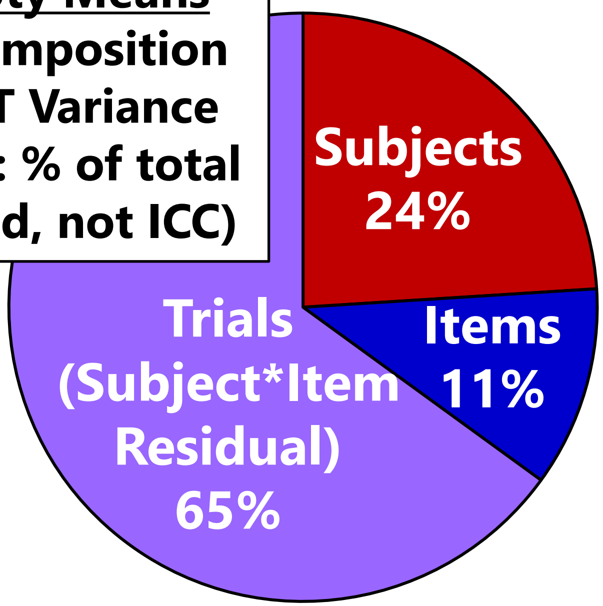
*t* trial
*i* item
*s* subject

- Both **subjects** and **items** as random effects:
  - Subject predictors explain between-subject mean variation: $\tau^2_{00S}$
  - Item predictors explain between-item mean variation: $\tau^2_{0I0}$
  - Trial predictors explain trial-specific residual variation: $\sigma^2_e$

# Example 1: Psycholinguistic Study
(*Locker, Hoffman, & Bovaird, 2007*)

- Crossed design: 38 subjects by 39 items (words or nonwords)
- Lexical decision task: response time (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**

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

**Pseudo-$R^2$:**

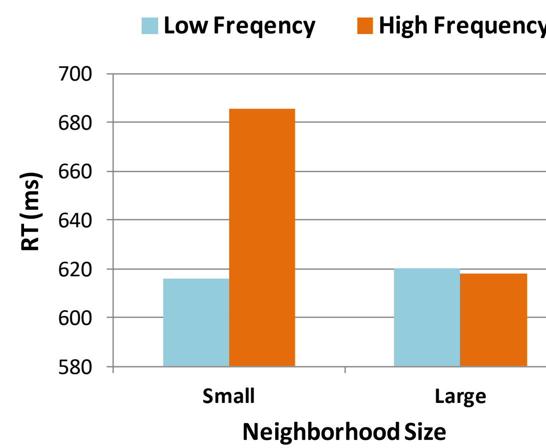
**Residual ≈ 0%**

**Subjects ≈ 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 |
|---|---|---|---|
| **$F_1$ Subjects ANOVA** | $F(1,37) = 16.1$<br>$p = .0003$ | $F(1,37) = 14.9$<br>$p = .0004$ | $F(1,37) = 38.2$<br>$p < .0001$ |
| **$F_2$ Words ANOVA** | $F(1,35) = 5.3$<br>$p = .0278$ | $F(1,35) = 4.5$<br>$p = .0415$ | $F(1,35) = 5.7$<br>$p = .0225$ |
| **F′ min (via ANOVA)** | $F(1,56) = 4.0$<br>$p = .0530$ | $F(1,55) = 3.5$<br>$p = .0710$ | $F(1,45) = 5.0$<br>$p = .0310$ |
| **Crossed MLM (via REML)** | $F(1,32) = 5.4$<br>$p = .0272$ | $F(1,32) = 4.6$<br>$p = .0393$ | $F(1,32) = 6.0$<br>$p = .0199$ |

# 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** | $F(1,37) = 16.1$<br>$p = .0003$ | $F(1,37) = 14.9$<br>$p = .0004$ | $F(1,37) = 38.2$<br>$p < .0001$ |
| **$F_2$ Words ANOVA** | $F(1,35) = 5.3$<br>$p = .0278$ | $F(1,35) = 4.5$<br>$p = .0415$ | $F(1,35) = 5.7$<br>$p = .0225$ |
| **F′ min (via ANOVA)** | $F(1,56) = 4.0$<br>$p = .0530$ | $F(1,55) = 3.5$<br>$p = .0710$ | $F(1,45) = 5.0$<br>$p = .0310$ |
| **Crossed MLM (via REML)** | $F(1,32) = 5.4$<br>$p = .0272$ | $F(1,32) = 4.6$<br>$p = .0393$ | $F(1,32) = 6.0$<br>$p = .0199$ |

# 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** | $F\ (1,37) = 16.1$<br>$p = .0003$ | $F\ (1,37) = 14.9$<br>$p = .0004$ | $F\ (1,37) = 38.2$<br>$p < .0001$ |
| **$F_2$ Words ANOVA** | $F\ (1,35) = 5.3$<br>$p = .0278$ | $F\ (1,35) = 4.5$<br>$p = .0415$ | $F\ (1,35) = 5.7$<br>$p = .0225$ |
| **F′ min (via ANOVA)** | $F\ (1,56) = 4.0$<br>$p = .0530$ | $F\ (1,55) = 3.5$<br>$p = .0710$ | $F\ (1,45) = 5.0$<br>$p = .0310$ |
| **Crossed MLM (via REML)** | $F\ (1,32) = 5.4$<br>$p = .0272$ | $F\ (1,32) = 4.6$<br>$p = .0393$ | $F\ (1,32) = 6.0$<br>$p = .0199$ |

# Ch. 12 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** |

# Example 1: Summary

- Although the $F_1 \times F_2$ criterion approach remains the current standard, its shortcomings are well known
  - $F_1$ ignores systematic variation across items
  - $F_2$ ignores systematic variation across subjects
  - Neither provides an accurate test of the effects of interest while considering **all** the relevant variation in response time

- Crossed random effects models may provide a tenable alternative with additional analytic flexibility…

…as illustrated by the next example…

# Example 2: Visual Search for Change (Hoffman & Rovine, 2007; Hoffman ch. 12)

- Outcome (DV)
  - Natural Log of RT to detect a change (up to 60 seconds)
  - 51 out of 80 natural scenes with > 90% accuracy
- Between-Subjects IV
  - Age: Younger (n = 96) vs. Older (n = 57) Adults
- Within-Subjects IVs
  - Change Meaningfulness to Driving (Low vs. High)
  - Change Salience (Low vs. High)
- Original Analysis Plan
  - $2 \times 2 \times 2$ mixed effects ANOVA on response time

# Analysis Plan, Reconsidered
# Issue #1: Systematic Item Differences

Can you find the change?



- Collapsing across scenes (as items) into condition means ignores systematic differences between scenes
- Treats items as fixed effects → $F_1$ ANOVA problem
  - Items will still vary in difficulty due to uncontrolled factors
  - Effect sizes may be inflated if that variability is not included
- ANOVA requires complete data to model variation across subjects and items simultaneously…

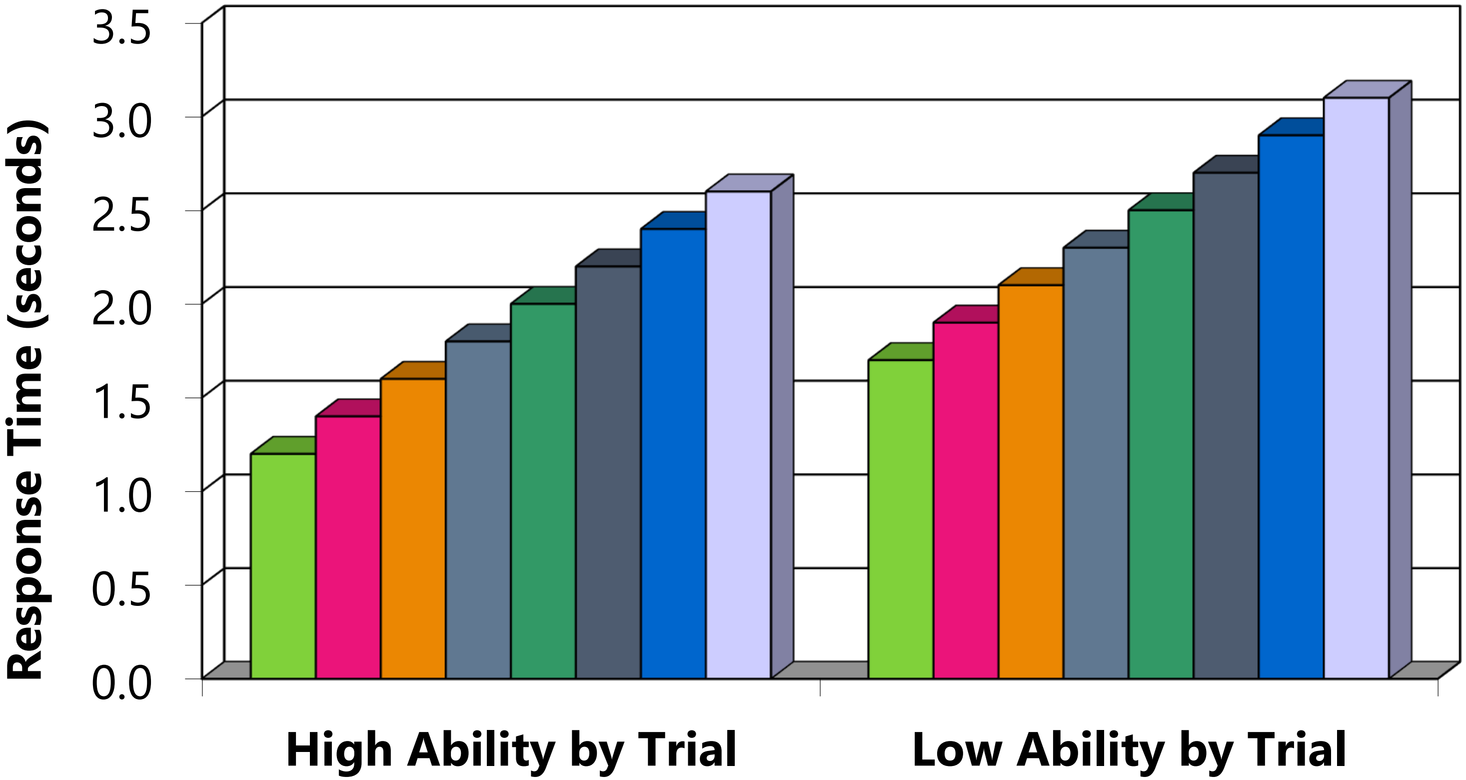
# Analysis Plan, Reconsidered
# Issue #2: Missing RTs for Incorrect Trials

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

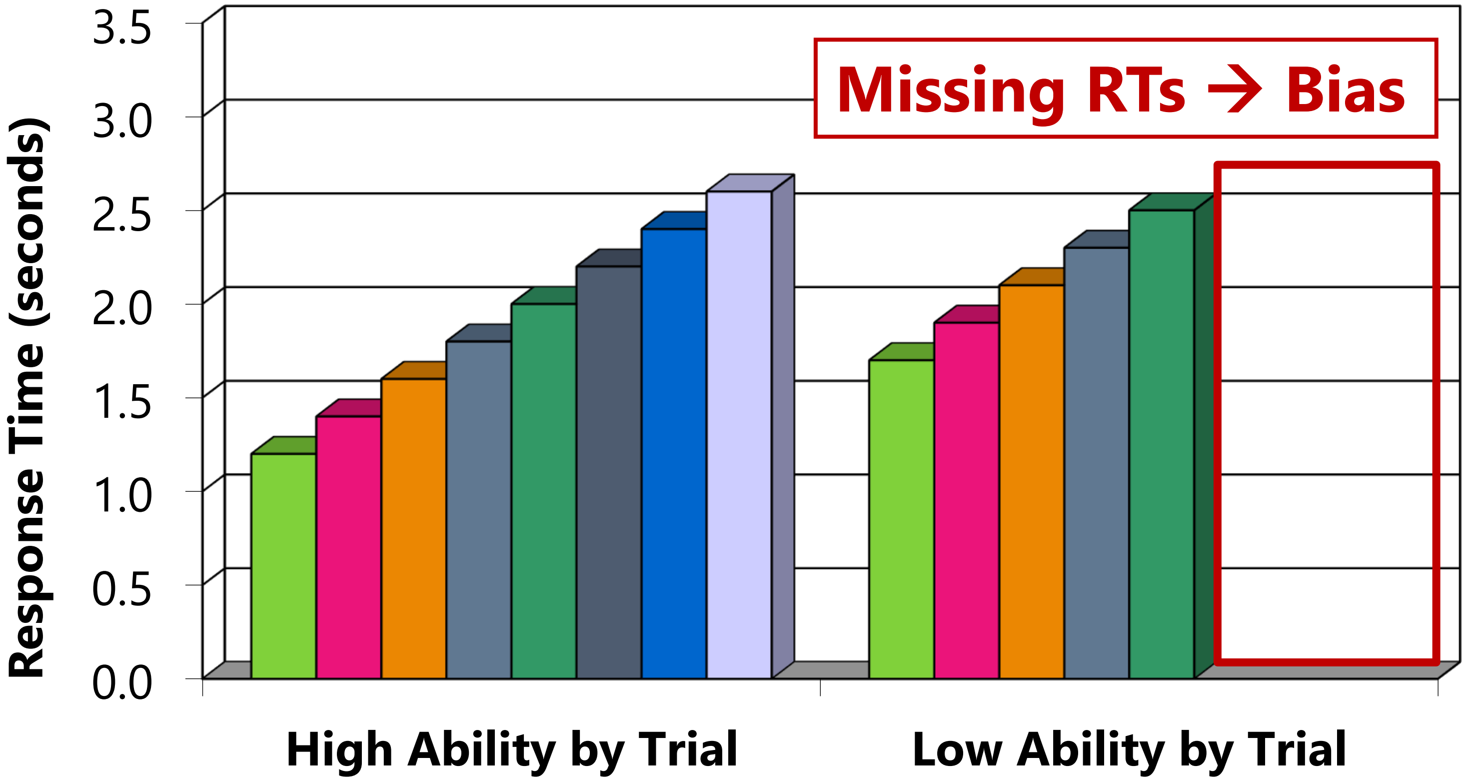
- RTs are more likely to be missing for difficult items
  - Downwardly biased condition mean RTs
  - Biased effects of predictor variables related to missingness
  - Loss of power due to listwise deletion

- 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

# Original RTs Across Trials by Ability



# Biased Condition Mean RT



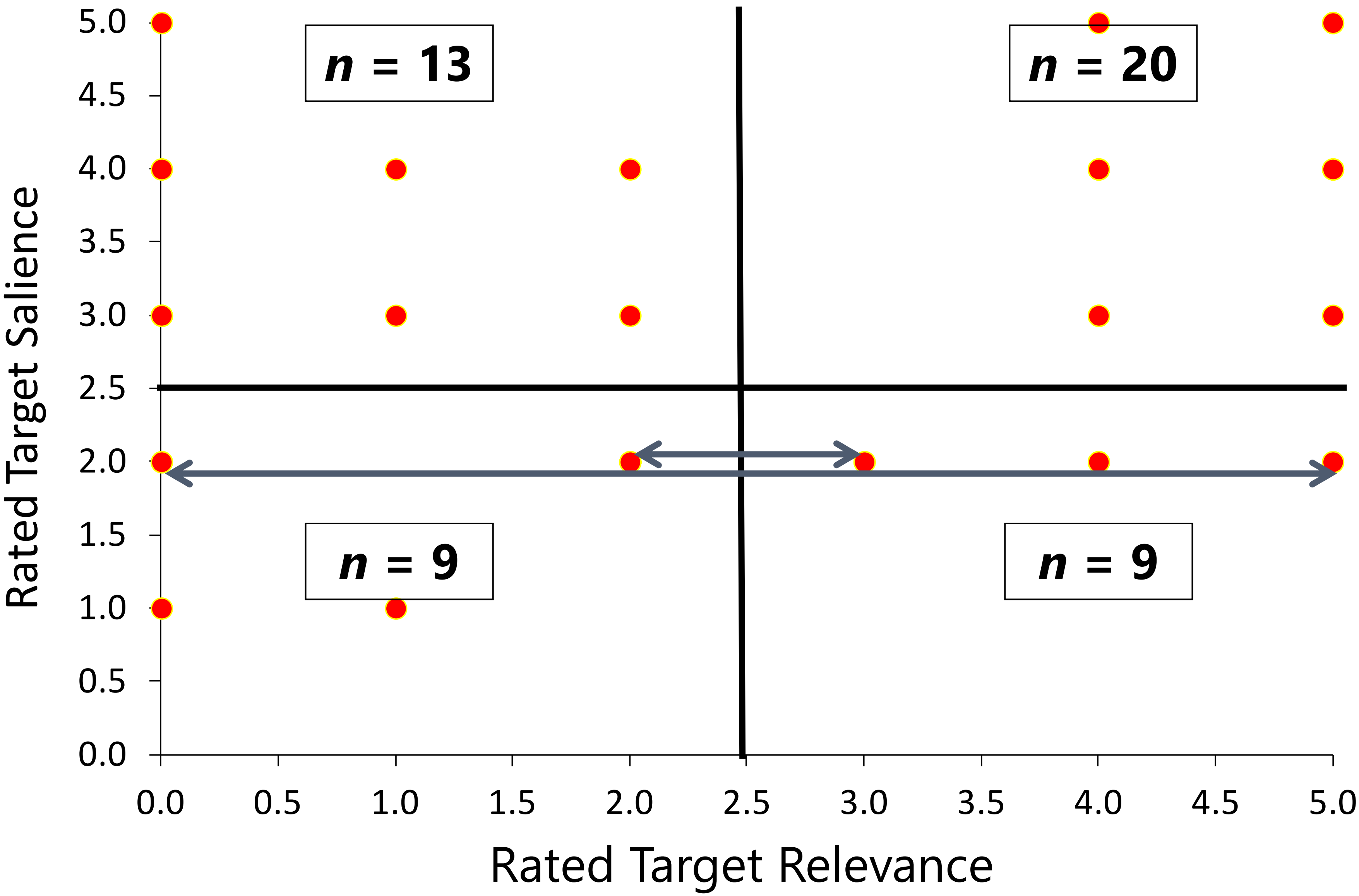
# Analysis Plan, Reconsidered
# Issue #3: Effects of Item Predictors

- 51 scenes 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 trials)

- Predictors of meaning and salience should be treated as continuous, which is problematic with an ANOVA

# Creating "Conditions" ($r = .22 \rightarrow r \approx 0$)

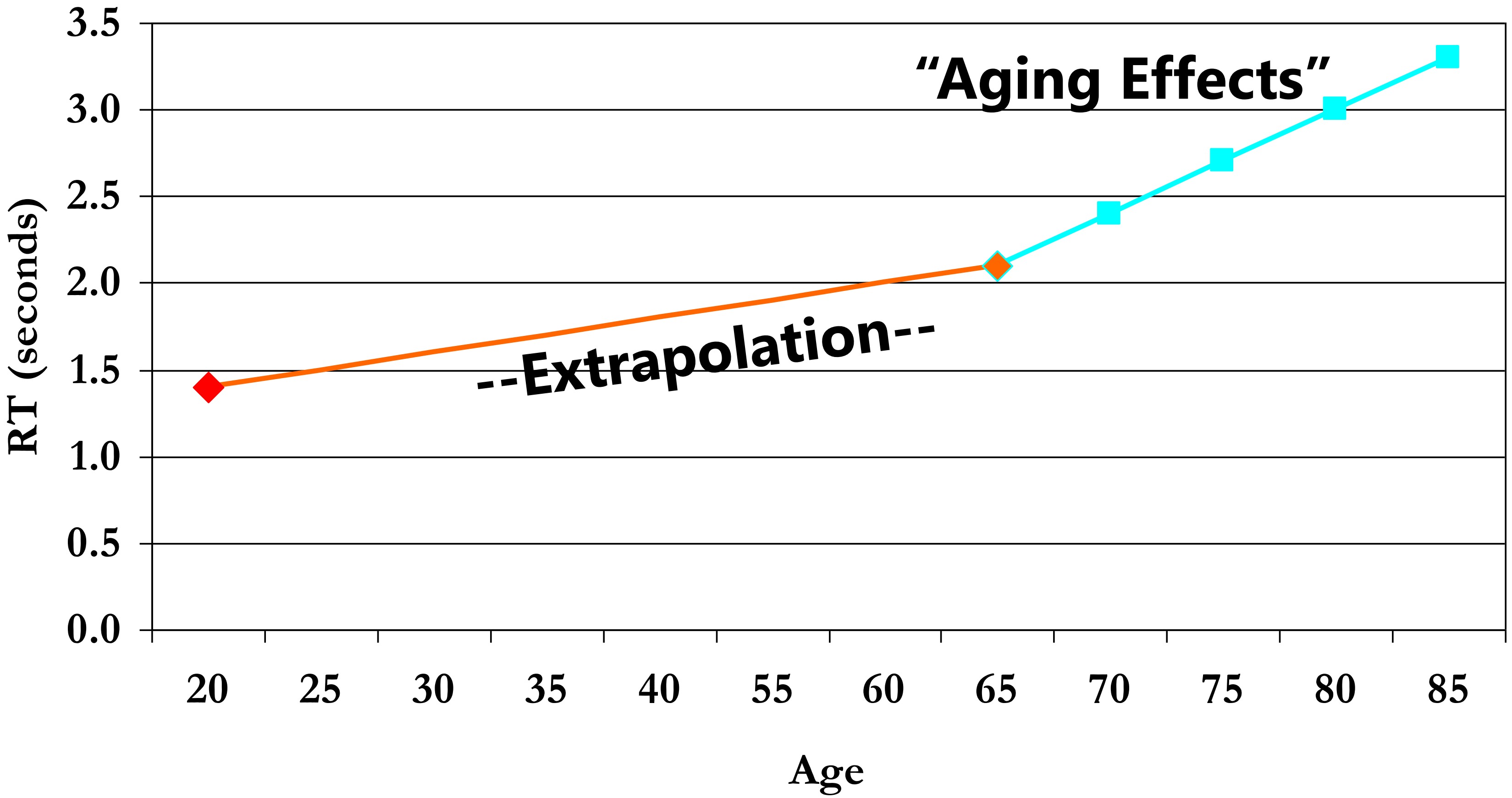


# Analysis Plan, Reconsidered
# Issue #4: Age Differences in Means

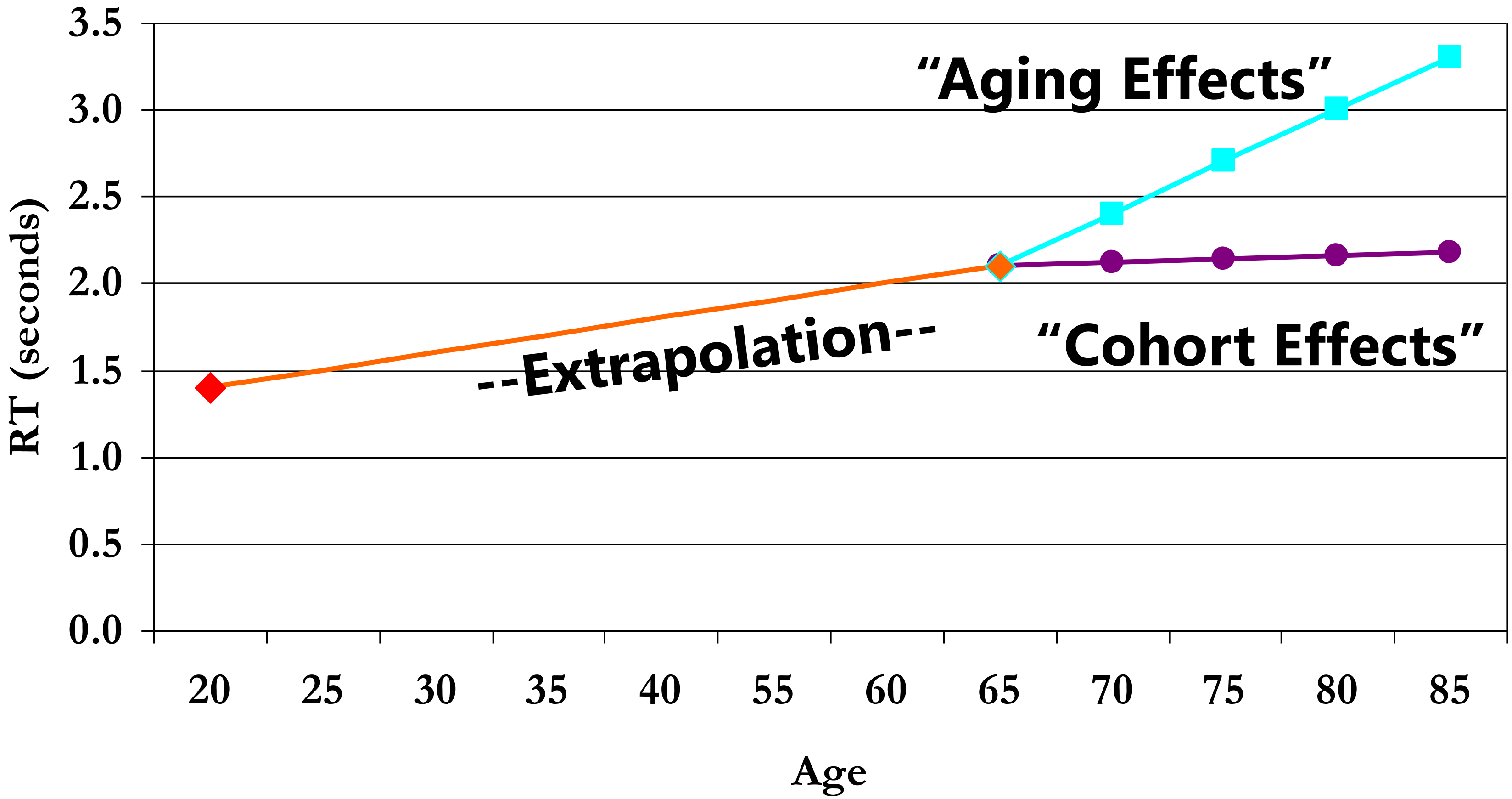
- "Younger" and "Older" adults were sampled, but…
  - Much more variability in age in the older group
    - 18–32 years (mostly 18–21) vs. 65–86 years
  - Age is not a strict dichotomy:
    - Including a single mean age group difference is not adequate
    - Separating "young-old" from "old-old" doesn't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

- Two effects of age are needed:
  - "Age Group" → difference between young and old
  - "Years over 65" → slope of age in the older group
  - This is a piecewise (spline) model of age!

# Piecewise (Semi-Continuous) Effects of Age on RT



# Piecewise (Semi-Continuous) Effects of Age on RT



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

- In addition to modeling differences in the means by age, the variances are likely to differ by age as well:
  - Older adults are likely to be more different *from each other* than are younger adults
    - Greater between-person variation in older group
  - Older adults are likely to be more variable *across trials* than are younger adults
    - Greater within-person variation in older group

- The model needs to accommodate heterogeneity of variance across age groups at multiple analysis levels

# Analysis Model, Reconsidered

- Scene predictors of relevance and salience should be modeled as continuous; the effect of age should be semi-continuous.
  - MLM allows categorical or continuous predictors at any level.

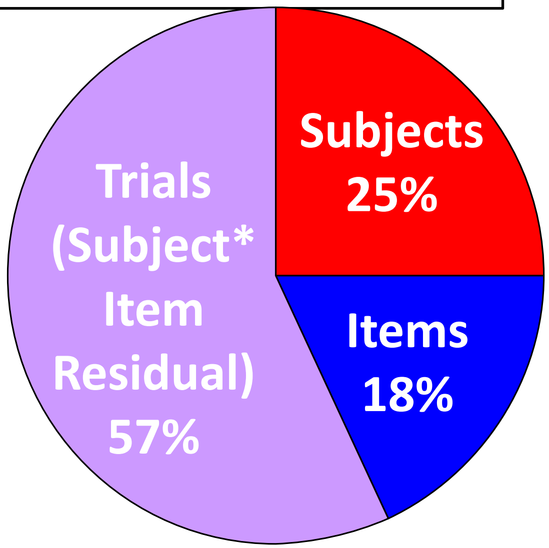
- RTs are not missing completely at random.
  - MLM only assumes missing at random.

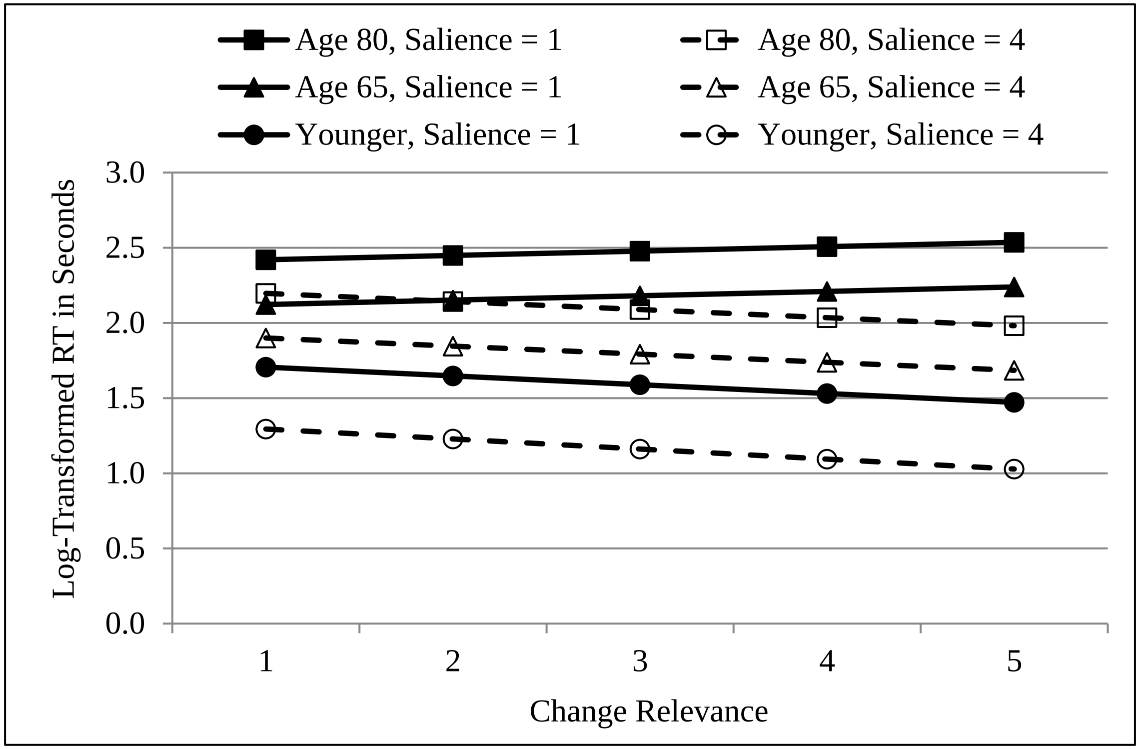
- Systematic differences between scenes should be included as a component of overall variance in RT.
  - MLM allows crossed random effects of subjects and items.

- Magnitude of variation between persons and within-persons (between trials) should be allowed to differ by age group.
  - MLM allows for heterogeneous variances by group at any level.

# Example #2: Final Model

**<u>Empty Means Model</u>**
**Decomposition of RT Variance (note: % of total is used, not ICC)**




Final model had random subject intercepts and salience slopes, with separate **G** and **R** matrices per age group

$$
\begin{aligned}
RT_{tis} = {} & \gamma_{000} + \gamma_{010}\left(Relevance_i - 3\right) + \gamma_{020}\left(Salience_i - 3\right) + \gamma_{030}\left(Relevance_i - 3\right)\left(Salience_i - 3\right) \\
& + \gamma_{001}\left(OlderGroup_s\right) + \gamma_{002}\left(YearsOver65_s\right) \\
& + \gamma_{011}\left(OlderGroup_s\right)\left(Relevance_i - 3\right) + \gamma_{021}\left(OlderGroup_s\right)\left(Salience_i - 3\right) \\
& + \gamma_{031}\left(OlderGroup_s\right)\left(Relevance_i - 3\right)\left(Salience_i - 3\right) + U_{00s} + U_{02s}\left(Salience_i - 3\right) + U_{0i0} + e_{tis}
\end{aligned}
$$

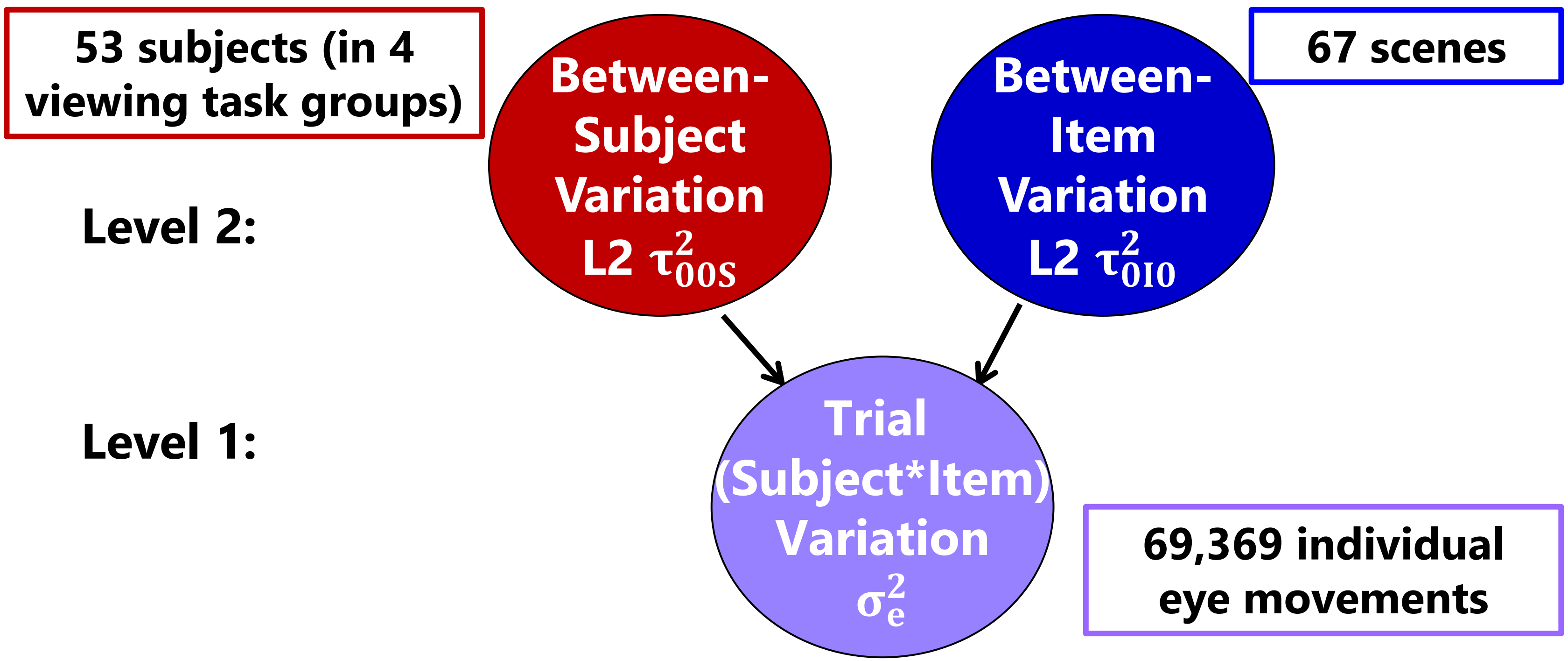
# Example #3: Eye Tracking (Mills et al., 2011)

- Does change over time in eye movements depend on the purpose of looking at a scene?
  - DVs: Fixation duration, saccadic amplitude
  - Each of the 53 subjects viewed the same 67 scenes for 6 sec
  - 4 between-subject viewing groups:
    - Free-view, Memorize, Rate Pleasantness, Search for n/z

- Original analysis: Mixed-effects ANOVA
  - Between-subjects task by chopped-up viewing time
    - Average over scenes; average within 20 “time” 500 msec conditions

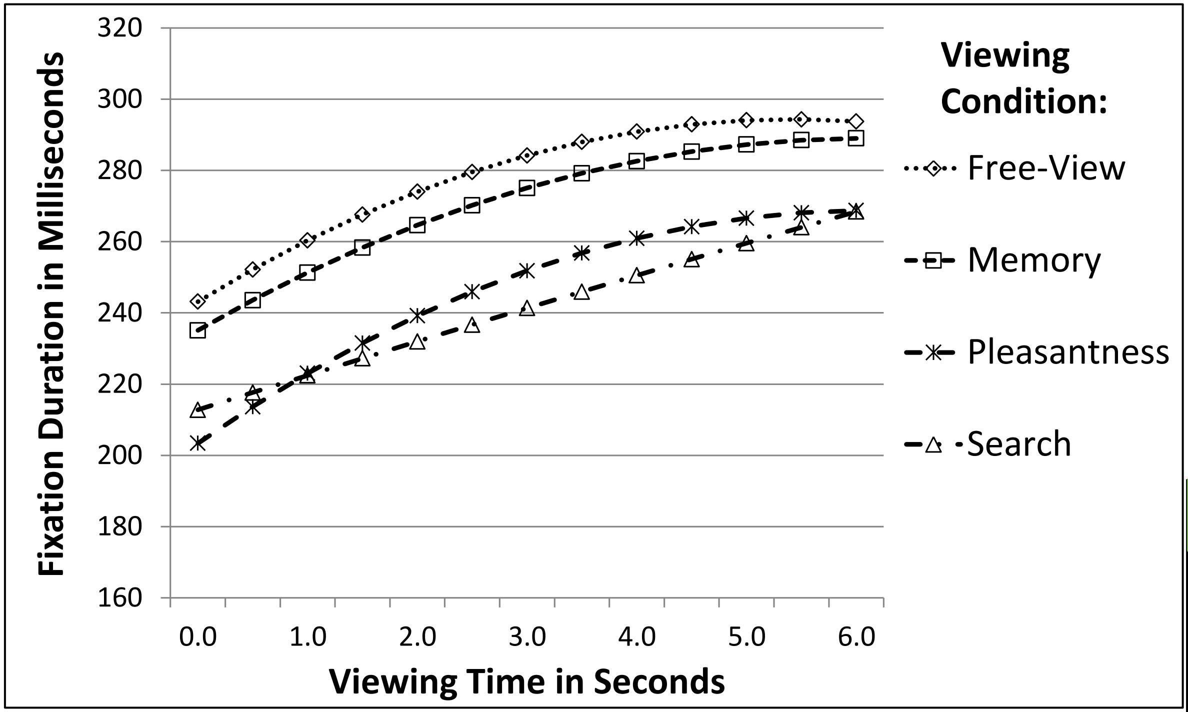
# Example #3: Eye Tracking

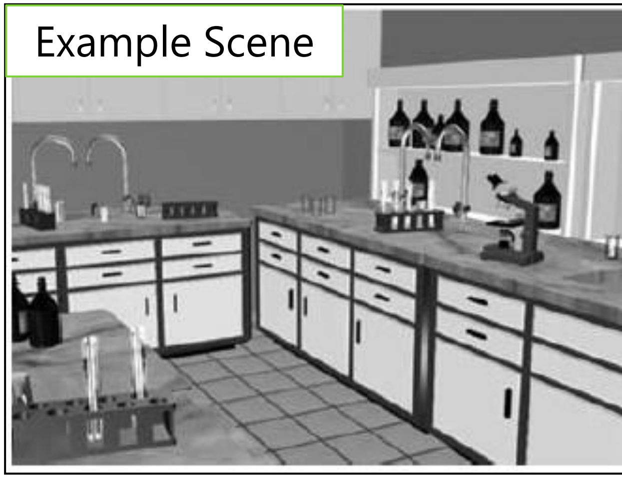
- New analysis: Growth curve modeling of eye movements!
  - Individual eye movements nested within scenes and within subjects
  - Scenes (items) and subjects are crossed random effects
  - Subject predictor = which viewing task they did, no scene predictors
  - Level-1 predictor = viewing time (with a random slope over subjects)



# Example #3: Eye Tracking

## Fixation duration changes *during* scene viewing based on goals



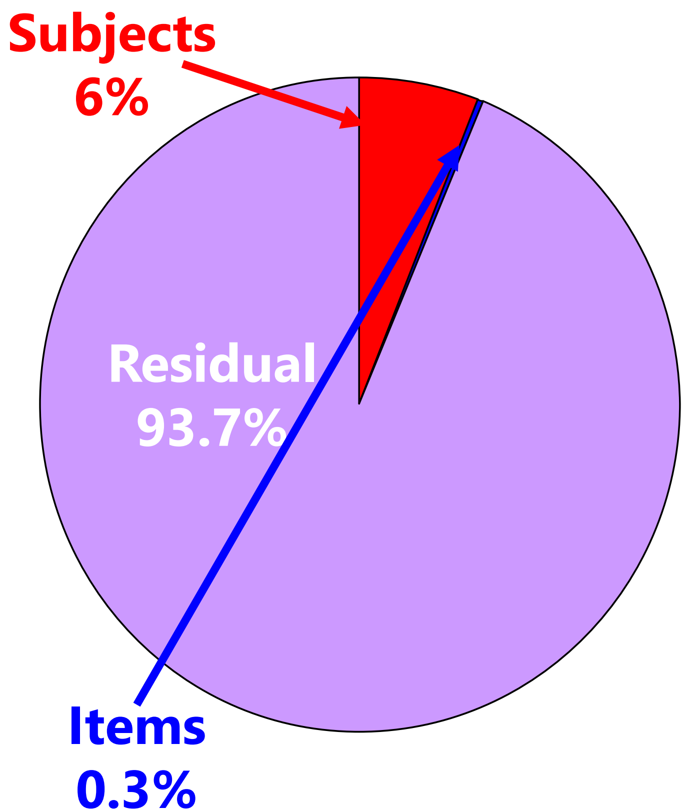


**UNL Psychology Program:** Visual Attention, Memory, and Perception Lab

Left: Mark Mills and Eye Tracker

# Example #3: Eye Tracking

**Empty Means Model**
**Decomposition of Fixation Duration Variance (note: % of total is used, not ICC)**

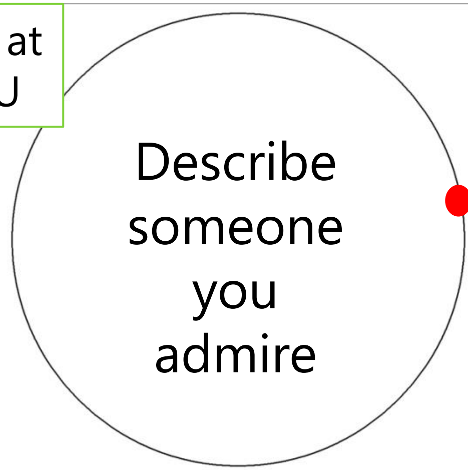


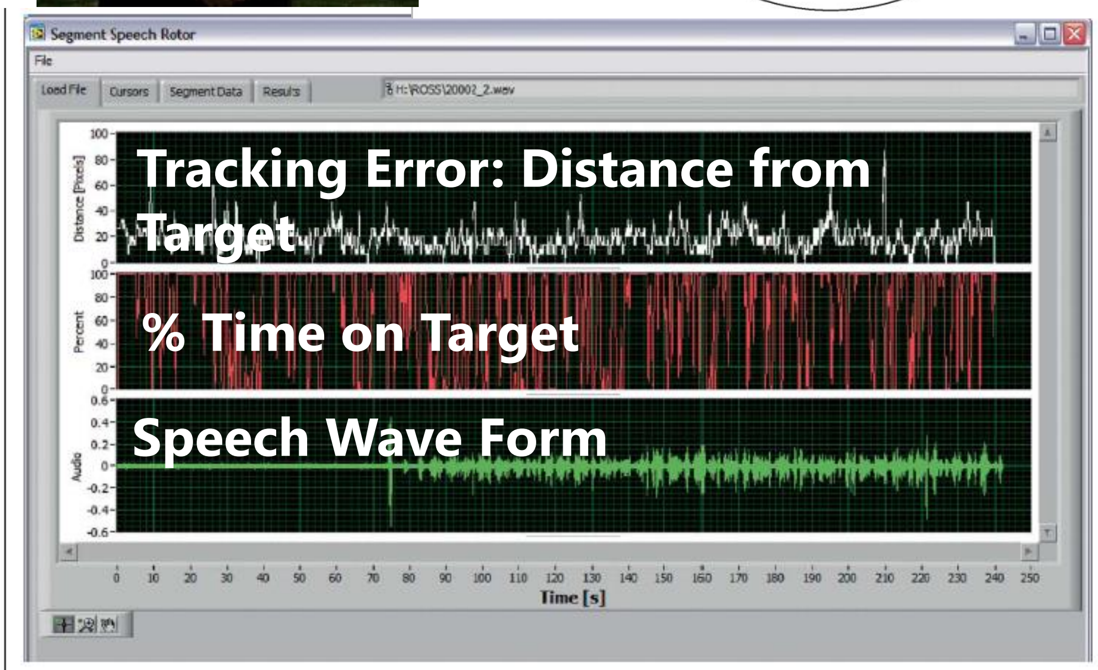

- Empty means models:
  - Residual variance only
  - + Subject, + item random intercepts
- Unconditional models:
  - + Linear and quadratic fixed time slopes
  - + Random linear time slope over subjects (could be random over items, too )
- Conditional models for task effects:
  - Main effect of viewing task → $R^2 \approx .32$ for subject intercept variance
  - Task * linear time → $R^2 \approx .03$ for subject linear time slope variance
  - Task * quadratic time → $R^2 \approx .00$ for residual variance (no random quadratic)

# Example #4: Tracking and Talking:

[Kemper, Hoffman, Schmalzried, Herman, & Kieweg (2011)]

Susan Kemper at Fraser Hall, KU




- **Model:** speech nested within subjects (no "items")
- **Dual task:** Track red ball with mouse while talking to examine costs of…
- **Speech planning:** current tracking suffers if *next* speech utterance is more complicated
- **Speech production:** current tracking suffers and becomes more variable while producing more complex speech and immediately after

# Conclusions

- An 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

- ANOVA is a special case of a more general family of 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

# References for Papers Mentioned


- Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

- 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.

- 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.