From Empty Models to Level-2 Predictors (in General Multilevel Models for Two-Level Nested Data)

- Topics:
  - Empty means model notation: from single-level to multilevel
  - Intraclass correlations, design effects, and effective sample size
  - Fixed slopes of level-2 predictors
    - Significance via Wald tests; effect sizes and explained variance
  - Random intercept variance
    - Significance via likelihood ratio tests; model comparison rules using REML or ML

Two Sides of a General Linear Model (GLM)

p = person

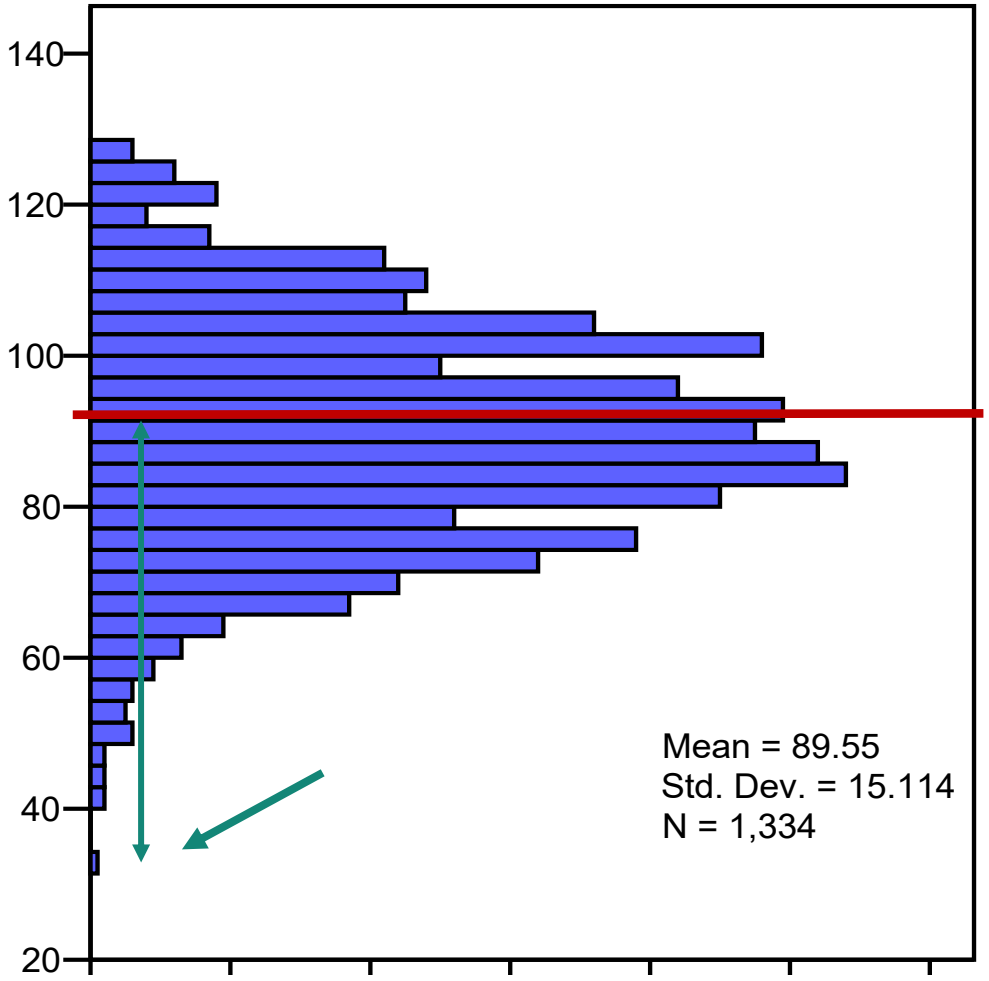
$$y_p = \beta_0 + \beta_1(x1_p) + \beta_2(x2_p) + \cdots + e_p$$

Our focus

- **Model for the Means (→ Predicted Values):**
  - Each person's expected (predicted) outcome is a weighted linear function of $x1_p$ and $x2_p$ (and any other predictors, …); each variable is measured once per person (as indicated by p subscript)
  - **Estimated β constants** are called **fixed effects** (two kinds: intercept or slopes)

- **Model for the Variance (→ "Piles" of Variance):**
  - $e_p \sim N(0, \sigma_e^2)$ → ONE (between-person) source of unexplained variation
  - In GLMs, e_p is supposed to be normally distributed with a mean of 0 with some estimated constant variance σ_e^2, unrelated to $x1_p$ and $x2_p$, and **independent** across all persons
  - **There is only ONE source of residual variance in the above GLM because it was designed to predict an outcome with only ONE dimension of sampling!**

An "Empty Means" General Linear Model → **Single-Level** Model *for the Variance*



$$y_p = \beta_0 + e_p$$

Filling in values:

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

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

Model for the Means

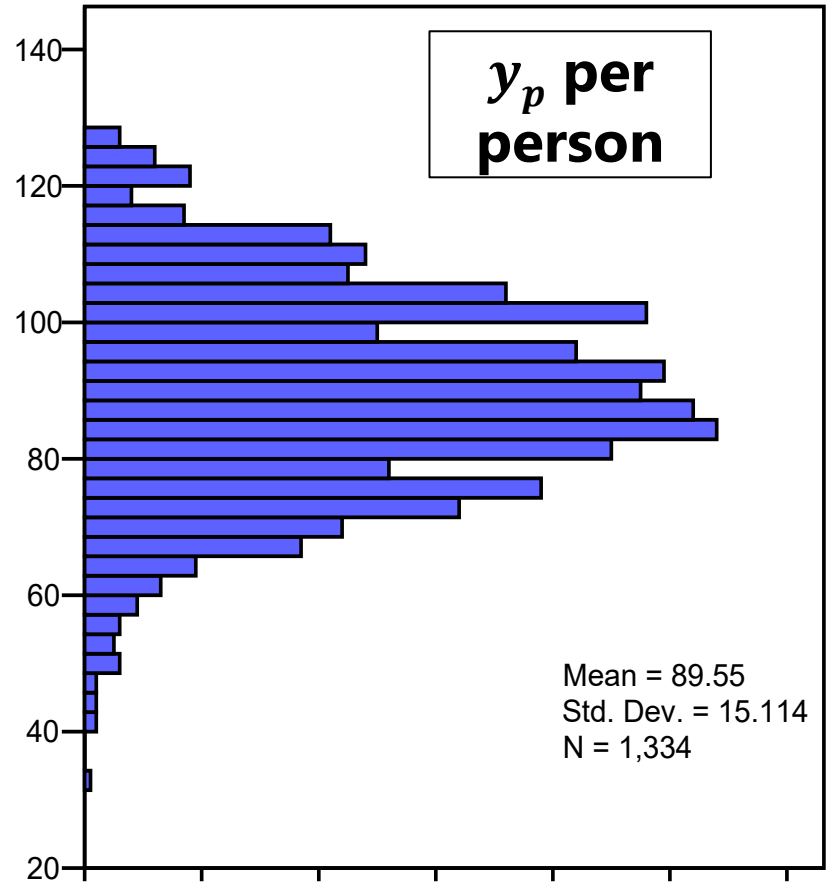
y_p residual variance:

$$\sigma_e^2 = \frac{\sum(y_p - \hat{y}_p)^2}{N-1}$$

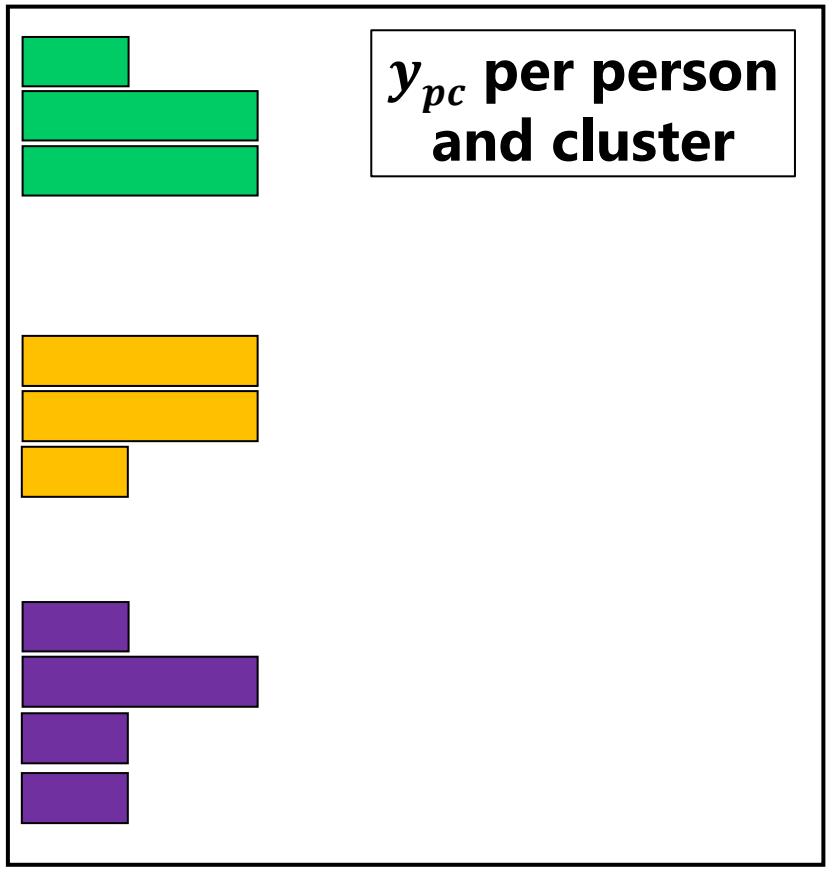
Without predictors, $\sigma_e^2 = \sigma_y^2$

Now with Multiple Persons in Multiple Clusters → **Two-Level** Model *for the Variance*

Full Sample Distribution



3 Clusters (c), 5 Persons (p)



Empty Means, Two-Level Model *for the Variance*

From a **one-level** to a **two-level model** for the variance:



Start off with the outcome's mean as a "best guess" for any value:

= Grand Mean

→**Fixed Intercept**

Can make *better* guess by taking advantage of cluster-common information:

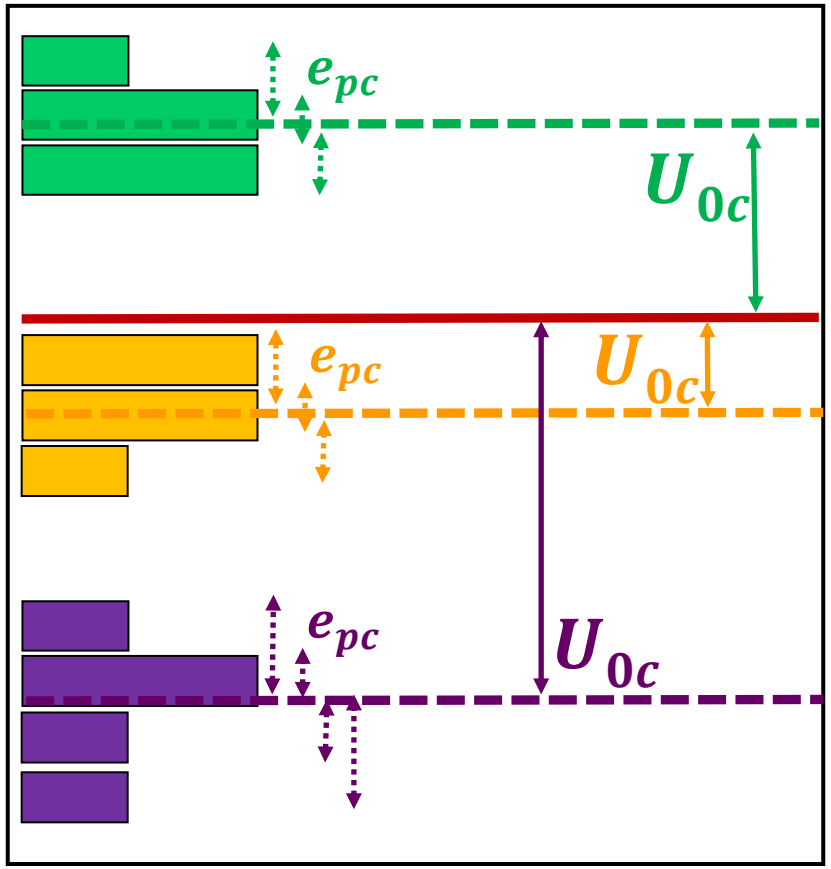
= Cluster Mean

→ **Random Intercept**

Empty Means, Two-Level Model *for the Variance*

β_0 → mean of cluster means

y_{pc} variance → 2 sources:



$$y_{pc} = \beta_0 + U_{0c} + e_{pc}$$

Level-2 random intercept U_{0c} (with variance labeled $\tau^2_{U_0}$):

- **Between**-cluster (BC) variance
- **INTER**-cluster differences to be explained by cluster predictors

Level-1 residual e_{pc} per person (with variance labeled σ^2_e):

- **Within**-cluster (WC) variance
- **INTRA**-cluster differences to be explained by person predictors

Two-Level Model Using Multilevel Notation: Empty Means, Random Intercept Model

GLM Empty Model:

$$y_p = \beta_0 + e_p$$

MLM Empty Model:

- Level 1 (L1):

$$y_{pc} = \beta_{0c} + e_{pc}$$

- Level 2 (L2):

$$\beta_{0c} = \gamma_{00} + U_{0c}$$

3 estimated parameters:

Model for the Means (1):

- Fixed Intercept γ_{00}

Model for the Variance (2):

- Level-1 **Within-Cluster** Variance of e_{pc} → σ_e^2
- Level-2 **Between-Cluster** Variance of U_{0c} → τ_{U0}^2

L1 Residual = person-specific deviation from cluster-predicted outcome

Fixed Intercept = mean of cluster means (because no predictors yet)

L2 Random Intercept = cluster-specific deviation from predicted intercept

"Composite" equation:

$$y_{pc} = \gamma_{00} + U_{0c} + e_{pc}$$

Intraclass Correlation ("ICC" = ICC_1 specifically)

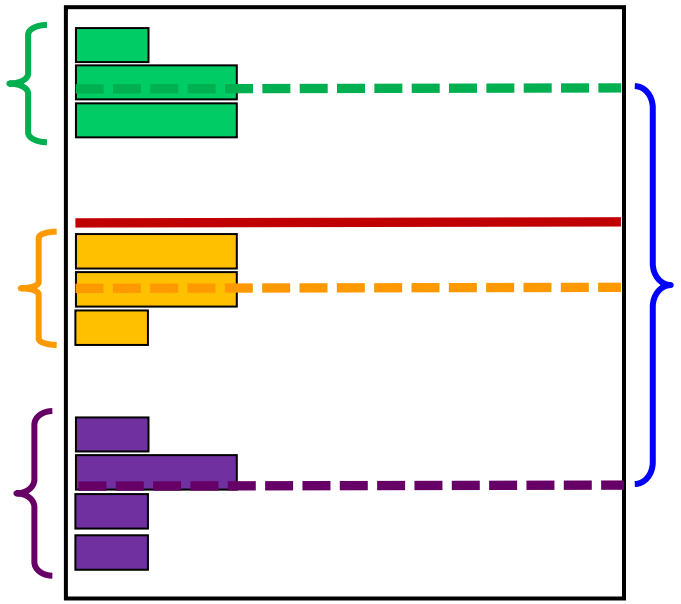
$$ICC_1 = \frac{\text{Between}}{\text{Betweeen} + \text{Within}} = \frac{\text{L2 Intercept Var}}{\text{L2 Intercept Var} + \text{L1 Residual Var}} = \frac{\boldsymbol{\tau^2_{U_0}}}{\boldsymbol{\tau^2_{U_0}} + \boldsymbol{\sigma^2_e}}$$

$\boldsymbol{\tau^2_{U_0}}$ → *Why don't all clusters have the same mean outcome?*

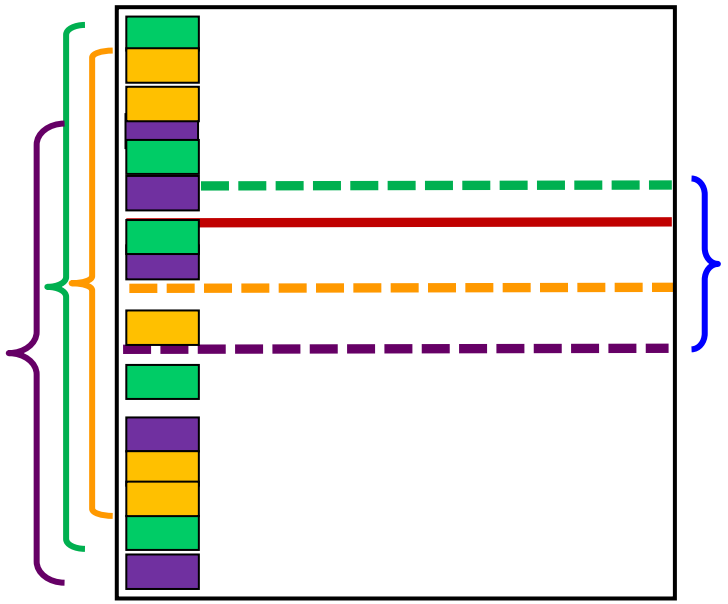
$\boldsymbol{\sigma^2_e}$ → *Why don't all people from the same cluster have the same outcome?*

- ICC_1 = **Proportion** of total variance that is **between clusters**
- ICC_1 = Average **correlation** of persons from same cluster
- ICC_1 is a standardized expression of how much *correlation* there is due to cluster mean differences → $\boldsymbol{ICC_1}$ **is an effect size for *constant* cluster dependency**
  - Non-constant cluster dependency can still be created by differences across clusters in the slopes of level-1 person predictors (i.e., random slopes—stay tuned!)
- Btw, **no variance has been "explained"** in an empty model → just 2 kinds of "error" *(side note: others have disagreed, but this is a hill I will die on)*

Even though **between-cluster variance** is the numerator, ICC_1 = mean **within-cluster** correlation!



$$ICC_1 = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \sigma^2_e}$$



ICC_1 = **BTW / BTW + within**

→ Large ICC_1

→ Large correlation within clusters

ICC_1 = **btw / btw + WITHIN**

→ Small ICC_1

→ Small correlation within clusters

Example: Level-1 Students in Level-2 Schools

Example from Snijders & Bosker (2012):
Predict language ($M = 41.46, VAR = 77.69$)
for 3,566 students (p) from 191 schools (c)

Level 1: $\boldsymbol{Lang_{pc} = \beta_{0c} + e_{pc}}$

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

$$\boldsymbol{ICC_1 = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \sigma^2_e} = \frac{17.809}{17.809 + 62.230} = .223}$$

22.3% of total language variance is due to school mean differences (WC $r = .223$)

Without random intercept $\boldsymbol{U_{0c}}$:

```
print("Single-Level Empty Means, No Random Intercept")
Model1 = lm(data=Lecture2, formula=langpost~1)
summary(Model1)
print("Get total variance"); summary(Model1)$sigma^2
```

```
[1] "Get total variance"
[1] 77.691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.464      0.148     281   <2e-16
```

With random intercept $\boldsymbol{U_{0c}}$:

```
print("Two-Level Empty Means, Random Intercept")
Model2 = lmer(data=Lecture2, REML=TRUE,
              formula=langpost~1+(1|schoolID))
summary(Model2)
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 schoolID (Intercept) 17.8     4.22
 Residual             62.2     7.89
Number of obs: 3566, groups:  schoolID, 191

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   41.079      0.337 174.914     122   <2e-16
```

Effects of Clustering on Effective N

- **Design Effect** expresses how much (planned) sample size needs to be adjusted due to clustering → "**effective sample size**"
- **Design Effect** = ratio of the variance using a clustered sampling design to the variance using a simple random sample from the same population, given the same total sample size either way
- Design Effect = $1 + ([L1n - 1] * ICC_1)$

$L1n$ = mean cluster size (# persons per cluster)

- Effective sample size → Effective $N = \frac{\text{\# Total Observations}}{\text{Design Effect}}$
- As ICC_1 and $L1n$ (cluster size) go UP, effective N goes DOWN
  - See Snijders & Bosker (2012) for more info and for a modified formula that takes unequal group sizes into account

Demonstrating Two-Level Design Effects

- Design Effect = $1 + ([L1n - 1] * ICC_1)$
- Effective $N = \frac{\text{\# Total Observations}}{\text{Design Effect}}$

Effective N ranges from total N (best case, $ICC_1 = 0$) to the # of clusters (worst case, $ICC_1 = 1$)

- $N = 500$: $n = 5$ patients from each of 100 doctors, $ICC_1 = .30$?
  - Patients Design Effect = $1 + ([5 - 1] * .30) = 2.20$
  - **Effective N** $= 500 \,/\, 2.20 \approx \mathbf{227}$ (not 500!)

- $N = 1000$: $n = 20$ students from each of 50 schools, $ICC_1 = .05$?
  - Students Design Effect = $1 + ([20 - 1] * .05) = 1.95$
  - **Effective N** $= 1000 \,/\, 1.95 \approx \mathbf{513}$ (not 1000!)

Example: Level-1 Students in Level-2 Schools

Example from Snijders & Bosker (2012):
Predict language for 3,566 students (p) from 191 schools (c), **with $L1n \approx 19$ per school**

Level 1: $\boldsymbol{Lang_{pc} = \beta_{0c} + e_{pc}}$

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

$$\boldsymbol{ICC_1 = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \sigma^2_e} = \frac{17.809}{17.809 + 62.230} = .223}$$

22.3% of total language variance is due to school mean differences (WC $r = .223$)

Design Effect $= 1 + ([L1n - 1] * ICC_1)$; Effective $N = \frac{\text{\# Total Observations}}{\text{Design Effect}}$

- $\bar{n} = 19$ students from 191 schools, $ICC_1 = .223$?
  - Design Effect $= 1 + ([19 - 1] * .223) = 5.01$
  - **Effective $N = 3{,}566 / 5.01 \approx 711$ "effective" students**

Does a non-significant ICC_1 mean you can ignore clustering and just do a standard general linear model?

- As ICC_1 and $L1n$ (cluster size) go UP, effective N goes DOWN
  - So there is NO VALUE OF ICC_1 that is "safe" to ignore, not even ~0!
  - Why? An $ICC_1 = 0$ in an *empty means (unconditional)* model can become $ICC_1 > 0$ after adding person predictors, because reducing the level-1 person residual variance $\boldsymbol{\sigma_e^2}$ can increase the level-2 cluster random intercept variance $\boldsymbol{\tau_{U0}^2}$ (→ *"conditional"* $ICC_1 > 0$)
    - **So design effects can increase** after including good person predictors!

- So just plan to do a multilevel analysis anyway…
  - Even if "that's not your question"… because people are in clusters, we still need to address **cluster dependency** (= **correlation**) because of:
    - Effect on SEs for predictor fixed slopes → **biased SEs**
    - Potential for **contextual effects** of person predictors (stay tuned!)
      - A "clustered-sampling correction" to the SEs **will not fix this problem**!!!!

Expected Increases in Random Intercept Variance

- Level-2 cluster random intercept variance $\tau^2_{U_0}$ will often increase after reducing the level-1 person residual variance σ^2_e via person predictors
- **Observed level-2 $\tau^2_{U_0}$** is NOT just between-cluster variance
  - Also contains a small part of within-cluster variance (**level-1 σ^2_e**), or:
    Observed level-2 $\tau^2_{U_0}$ = True level-2 $\tau^2_{U_0}$ + (σ^2_e / $L1n$)
    - With increasing $L1n$ persons per cluster, bias in true level-2 $\tau^2_{U_0}$ due to level-1 σ^2_e is minimized
  - Model estimates of "**True**" **level-2 $\tau^2_{U_0}$** use **(σ^2_e / $L1n$)** as correction factor:
    True level-2 $\tau^2_{U_0}$ = Observed level-2 $\tau^2_{U_0}$ – (σ^2_e / $L1n$)

- e.g., **Observed level-2 $\tau^2_{U_0}$ = 4.65**, **level-1 σ^2_e = 7.06**, **$L1n$ = 4**
  - **True level-2 $\tau^2_{U_0}$** = **4.65** – (**7.06** / **4**) = **2.88** in empty means model
  - Add level-1 within-cluster predictor → reduce **level-1 σ^2_e** from **7.06** to **2.17**
  - But now **True level-2 $\tau^2_{U_0}$** = **4.65** – (**2.17** / **4**) = **4.10** → more dependency than before!

"Effect Size" via Random Effect Confidence Intervals

- Practical magnitude can also be conveyed with a **random effect confidence interval** (CI)
- This is NOT a typical CI, which conveys a fixed effect *estimate's* uncertainty over repeated samples
- Instead, a random effect CI describes the effect's **predicted range over 95% of level-2 units** (still) using the fixed effect estimate as the center

Typical 95% CI for the fixed intercept:

$$CI = Est \pm t_{crit} * SE$$
$$CI = 41.08 \pm 1.96 * 0.34$$
$$CI = 40.42 \text{ to } 41.74$$

Random intercept 95% CI for expected range of language school means:

$$CI = Est \pm t_{crit}\sqrt{Var(U_{0c})}$$
$$CI = 41.08 \pm 1.96\sqrt{17.81}$$
$$CI = 32.81 \text{ to } 49.35$$

```
Random effects:
 Groups   Name        Variance Std.Dev.
 schoolID (Intercept) 17.8     4.22
 Residual             62.2     7.89
Number of obs: 3566, groups:  schoolID, 191

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   41.079      0.337 174.914     122   <2e-16
```

ICC_2: Reliability of the Cluster Mean

Example from Snijders & Bosker (2012):
Predict language for 3,566 students (p) from 191 schools (c), with $L1n = 19$ per school

Level-1: $\boldsymbol{Lang_{pc} = \beta_{0c} + e_{pc}}$

Level-2: $\boldsymbol{\beta_{0c} = \gamma_{00} + U_{0c}}$

$$\boldsymbol{ICC_1 = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \sigma^2_e} = \frac{17.809}{17.809 + 62.230} = .223}$$

22.3% of total language variance is due to school mean differences (WC $r = .223$)

ICC_2 = (Spearman-Brown) Reliability of the Cluster Mean

$$\boldsymbol{ICC_2 = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + (\sigma^2_e / L1n)} = \frac{17.809}{17.809 + (62.230/19)} = .845}$$

Reliability of the school language mean is ≈ .845

Cluster mean reliability (= ICC_2) increases with **cluster size** and/or mean differences across clusters (= $\boldsymbol{ICC_1}$)

Btw, you can also think of this as the empty means, random intercept model as a parallel items measurement model *(in which student responses are the items being predicted by a single latent factor, with all factor loadings fixed to 1 and all item residual variances held equal across students),* although it's never estimated this way

Side Note: Observed vs. Predicted Cluster Means

- Given an empty means, two-level model: $\boldsymbol{y_{pc} = \gamma_{00} + U_{0c} + e_{pc}}$
  - Each cluster's mean is *supposed* to be given by $\boldsymbol{\gamma_{00} + U_{0c}}$ but it may not be!

- 3 indices of variance in cluster means, ordered from largest to smallest:
  1. Compute each cluster's observed mean; compute their variance
  2. Estimate the cluster $\boldsymbol{U_{0c}}$ variance as $\boldsymbol{\tau^2_{U_0}}$ directly (what we do in MLM)
  3. Save predicted $\boldsymbol{U_{0c}}$ terms as BLUPs, compute their variance (~2-stage MLM)

- Why? **Shrinkage**: the $\boldsymbol{U_{0c}}$ values for clusters with fewer people or more extreme means get "pushed" towards the center ($\boldsymbol{\gamma_{00}}$)
  - **This is why you should not work with predicted $\boldsymbol{U_{0c}}$ terms as observed variables—we should predict them directly in a MLM (or M-SEM) instead**

Intermediate Summary: A New Model for Variance

- Two-level models differ from single-level models by partitioning the total outcome variance into level-2 between and level-1 within
  - Level-2 between = random intercept variance $\boldsymbol{\tau_{U_0}^2}$ (mean differences across clusters here)
  - Level-1 within = residual variance $\boldsymbol{\sigma_e^2}$ (deviations across people within clusters here)
  - This partitioning does NOT explain variance—it just distinguishes two kinds: here, cluster predictors explain level-2 between; people predictors explain level-1 within
  - The idea of level 2 = between and level 1 = within holds across all sampling designs!

- A random intercept variance induces a constant correlation across level-1 persons from same level-2 cluster (which in turn lowers the effective sample size)
  - Expected correlation is given by between-cluster variance / total variance: $\boldsymbol{ICC_1 = \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + \sigma_e^2}}$
  - Reliability of cluster means is given by ICC_2, but is limited by ICC_1 and cluster size
  - Correlation can increase by decreasing level-1 residual variance via person predictors
  - Random intercepts should be predicted directly in a model, not saved as observed variables

Adding Fixed Slopes for Level-2 Cluster Predictors

- **Level-2 predictors** are cluster-level characteristics (constant over their persons)
  - Example: Level-1 (L1) students (p) nested in level-2 (L2) schools (c) that vary in homework amount ($HW_c - 2$) and mixed grades ($MixGrd_c$; 0=no, 1=yes)

- **Level 1:** $\boldsymbol{Lang_{pc} = \beta_{0c} + e_{pc}}$

$\boldsymbol{\sigma_e^2}$ → All possible level-1 residual variance for within-school differences across students

- "**Unconditional**" **Level 2** (before cluster predictors):
  - $\boldsymbol{\beta_{0c} = \gamma_{00} + U_{0c}}$

$\boldsymbol{\tau_{U_0}^2}$ → **All possible level-2 random intercept variance** due to school mean differences

- "**Conditional**" **Level 2** (after cluster predictors):
  - $\boldsymbol{\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + U_{0c}}$
  - Two new fixed slopes: $\boldsymbol{\gamma_{01}}$ and $\boldsymbol{\gamma_{02}}$
  - First subscript = which beta in level-1 model
    Second subscript = order of predictor in level-2 model for that beta

$\boldsymbol{\tau_{U_0}^2}$ → level-2 random intercept variance **leftover** after HW and mixed grade prediction

Review: Wald Significance Tests for Fixed Slopes

- Each **fixed slope** has 6 relevant characteristics (*essential to report):
  - ***Estimate** = $\boldsymbol{Est}$ = best guess for the fixed slope from our sample data
  - ***Standard Error** = $\boldsymbol{SE}$ = average distance of sample slope from population slope → expected *inconsistency* of slope across samples
  - "$\boldsymbol{t}$**-value**" = (Estimate − H_0) / SE = **test-statistic** for fixed slope against H_0 (usually $H_0 = 0$)
    - This is called a "**univariate Wald test**" (labeled as a "z-value" when not using denominator DF)
  - $\boldsymbol{p}$**-value** = (two-tailed) probability of fixed slope estimate *as or more extreme* if H_0 is true → how unexpected our result is on a t-distribution with 0=H_0, SD=SE given denominator DF
    - If p < alpha, then slope is "significant"; if p ≥ alpha, slope is "nonsignificant" (never "insignificant")
  - **(95%) Confidence Interval** = $\boldsymbol{CI} = Estimate \pm t_{critical} * SE$ = range in which true (population) value of estimate is expected to fall across (95%) of samples

- Same idea when testing multiple slopes, just in squared metric → "**Multivariate Wald test**"
  - Yields $\boldsymbol{F}$**-value** (with # slopes =numerator DF)

Uses Denominator Degrees of Freedom?	Test 1 slope	Test 2+ slopes
No: implies infinite N	z	$\chi^2 (= z^2$ if 1)
Yes: adjusts based on N	t	$\boldsymbol{F} (= t^2$ if 1)

Example: Adding Level-2 Cluster Predictors

Level-1: $Lang_{pc} = \beta_{0c} + e_{pc}$

Level-2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + U_{0c}$

Before adding level-2 predictors:

```
print("Two-Level Empty Means, Random Intercept")
Model2 = lmer(data=Lecture2, REML=TRUE,
              formula=langpost~1+(1|schoolID))
summary(Model2)

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolID (Intercept) 17.8     4.22
 Residual             62.2     7.89
Number of obs: 3566, groups:  schoolID, 191

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   41.079      0.337 174.914     122   <2e-16
```

After adding level-2 predictors:

```
print("Add 2 School-Level Predictors")
Model3 = lmer(data=Lecture2, REML=TRUE,
              formula=langpost~1+hw2+mixgrd+(1|schoolID))
summary(Model3)

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolID (Intercept) 17.2     4.14
 Residual             62.2     7.89
Number of obs: 3566, groups:  schoolID, 191

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   41.548      0.505 167.500   82.26   <2e-16
hw2            0.507      0.635 172.279    0.80   0.4261
mixgrd        -1.993      0.708 188.544   -2.81   0.0054
```

Example: Adding Level-2 Cluster Predictors

Level 1: $Lang_{pc} = \beta_{0c} + e_{pc}$

Level2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + U_{0c}$

Model for the Means:

- $\gamma_{00} = 41.55$ → fixed **intercept**: expected language for students in a school with homework=2 (~mean) and mixgrd=0 (unmixed)
- $\gamma_{01} = 0.51$ → fixed slope of **HW−2**: difference in school mean language for each unit more homework the school tends to assign
- $\gamma_{02} = -1.99*$ → fixed slope of **mixgrd**: difference in school mean language in schools with mixed grades instead of un-mixed grades

Model for the Variance:

- U_{0c} = level-2 random intercept → remaining deviation from fixed-effect-predicted school mean language for school c (with variance $\tau^2_{U_0} = 17.16$)
- e_{pc} = level-1 residual → remaining deviation of the observed outcome for student p from their outcome predicted by β_{0c} (with variance $\sigma^2_e = 62.20$)

Effect Sizes for Fixed Slopes of Level-2 Cluster Predictors

- Direct: convert t-statistic for fixed effect into Cohen's d or partial r
  - Cohen's $d = \frac{2t}{\sqrt{DF_{den}}}$, $partial\ r = \frac{t}{\sqrt{t^2 + DF_{den}}}$

Note: These formulas can be used with any model (multilevel or not)

- Indirect: explained variance of two complementary kinds
  - **Pseudo-R^2**: amount of variance explained *per variance component (per pile)*
    - Pseudo-$R^2 = \frac{variance_{fewer} - variance_{more}}{variance_{fewer}} = 1 - \frac{variance_{more}}{variance_{fewer}}$

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

    - Pseudo-R^2 can go negative if adding useless predictors or if the level-1 model is mis-specified (stay tuned!); these problems can be remedied by calculating it by using model-implied total variance instead (see Rights & Sterba, 2019; 2020, in R packages for certain types of models)
    - Only pseudo-R^2 for the L2 random intercept variance is relevant for L2 predictors
  - **Total-R^2**: amount of total variance explained (combined across piles)
    - Generate model-predicted $\hat{y}_{pc}$ values from fixed effects ONLY and correlate them with observed outcomes; square that correlation to get total-R^2 (same as in single-level models)

Example: Effect Sizes for Level-2 Fixed Slopes

- Results from example predicting student language:
  - Empty: Level-1 $\boldsymbol{\sigma_e^2 = 62.230}$ and Level-2 $\boldsymbol{\tau_{U_0}^2 = 17.809}$, so $\boldsymbol{ICC_1 = .223}$
  - Conditional: Level-1 $\boldsymbol{\sigma_e^2 = 62.201}$ (≈ because no person predictors yet),
    Level-2 $\boldsymbol{\tau_{U_0}^2 = 17.164}$ after controlling for mixed grade and HW,
    creating a **conditional** (after predictors) $\boldsymbol{ICC_1 = .216}$ (smaller correlation)

- **Variance explained** by the two level-2 cluster fixed slopes:
  - **Pseudo-$R_{U_0}^2$** $= \frac{17.809 - 17.164}{17.809} = .036$ → **3.6%** of the level-2 random intercept variance (due to school mean differences) in language
  - **Total-R^2 = $r(\hat{y}_{pc}, y_{pc})^2 = .0792 = .006$** → 0.6% of the total variance in language (22.3% of which was originally due to school mean diffs)
  - Because these R^2 values mean very different things (and there are many other kinds), it is essential to clearly describe how you calculated them and what they then mean!

What if we had used single-level GLM instead?

Level 1: $Lang_{pc} = \beta_{0c} + e_{pc}$

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + U_{0c}$

So which result should be reported and believed???

With random intercept U_{0c}:

```
print("Add 2 School-Level Predictors")
Model3 = lmer(data=Lecture2, REML=TRUE,
              formula=langpost~1+hw2+mixgrd+(1|schoolID))
summary(Model3)

Random effects:
 Groups   Name        Variance Std.Dev.
 schoolID (Intercept) 17.2     4.14
 Residual             62.2     7.89
Number of obs: 3566, groups:  schoolID, 191

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)   41.548      0.505 167.500   82.26   <2e-16
hw2            0.507      0.635 172.279    0.80   0.4261
mixgrd        -1.993      0.708 188.544   -2.81   0.0054
```

Without random intercept U_{0c}:

```
print("Add 2 School-Level Predictors
      WITHOUT Random Intercept")
Model4 = lm(data=Lecture2,
            formula=langpost~1+hw2+mixgrd)
summary(Model4)
print("Residual variance"); summary(Model4)$sigma^2

          [1] "Residual variance"
          [1] 77.251

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept)   41.637      0.216  192.50  < 2e-16
    hw2            0.411      0.280    1.47     0.14
    mixgrd        -1.474      0.347   -4.25 0.000022
```

Introducing Relative Model Fit Comparisons

- Nested models—in which one model is a subset of the other—in two-level models can now differ from each other in two distinct ways

- **Model for the Means** → which fixed effects (for which predictors)
  - **Does not** require assessment of relative model fit because we can still use univariate (t) or multivariate (F) Wald tests for this, same as in single-level models

- **Model for the Variance** → which random effects (for the pattern of variance and covariance of residuals from the same sampling unit)
  - **DOES require assessment of relative model fit, now using likelihoods (up next)**
  - Btw, variances should not use Wald test *p*-values—they use a two-sided sampling distribution for what the variance could be, but variances cannot be negative, so those *p*-values will not be valid, more so for variances close to 0

Introducing Likelihood Estimation

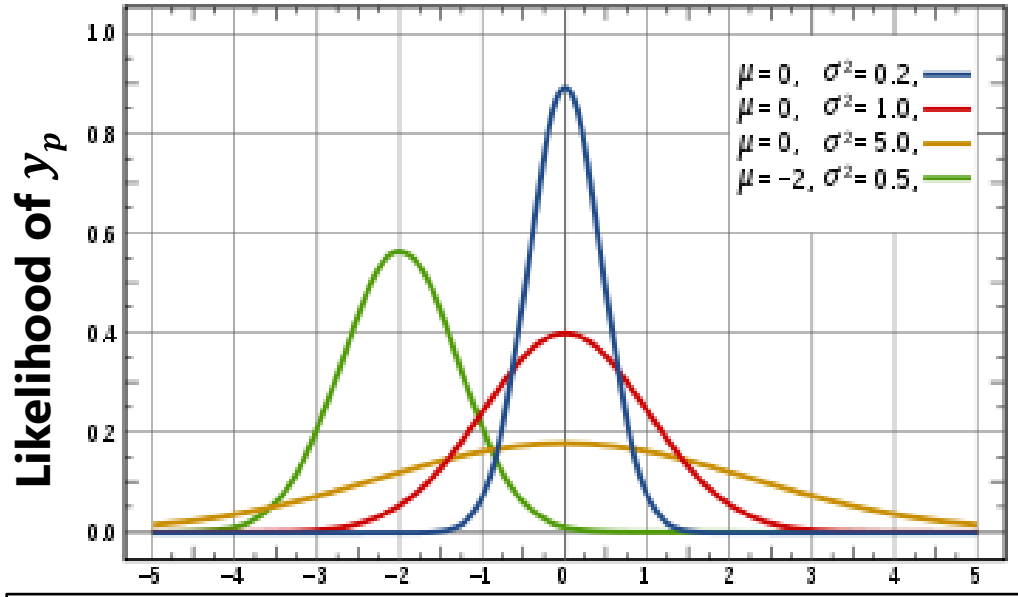
- Single-level general linear models use Ordinary Least Squares (OLS) estimation
  - "Best" fixed effects are those that **minimize the sum of squared residuals** (e_p^2)
  - How? Compute sums of squares → mean squares → F-ratios…

- The good news: **Maximum Likelihood (ML) estimation can be used to find the "best" answers** with way more flexibility in **many more kinds of data**
  - Minimizing sum of squared residuals ≈ maximizing likelihood of the data
  - OLS is actually a computational shortcut to "residual" ML (**REML**; what we will use)

- **The even better news**: If you understand **this**, then you understand the **basics of ML or REML**
  - We can use excel functions to describe the basics, starting with a single-level outcome for simplicity



Univariate Normal Probability Distribution Function



Univariate Normal PDF:

$$f(y_p) = \frac{1}{\sqrt{2\pi\sigma_e^2}} * \exp\left[-\frac{1}{2} * \frac{\left(y_p - y_p\right)^2}{\sigma_e^2} \right]$$

Sum over persons of log of $f(y_p)$ = Model Log-Likelihood (LL) → Fit Index

- This PDF tells us how **likely** (i.e., **tall**) any value of $\boldsymbol{y_p}$ is given two things:
  - Predicted value $\hat{\boldsymbol{y}}_{\boldsymbol{p}}$ → $\boldsymbol{\mu}$ on the left
  - Residual variance $\boldsymbol{\sigma_e^2}$ → $\boldsymbol{\sigma^2}$ on the left
- We can see how this works using the NORMDIST function in excel!
  - Easiest for **empty** single-level model:

$$\boldsymbol{y_p = \beta_0 + e_p}$$

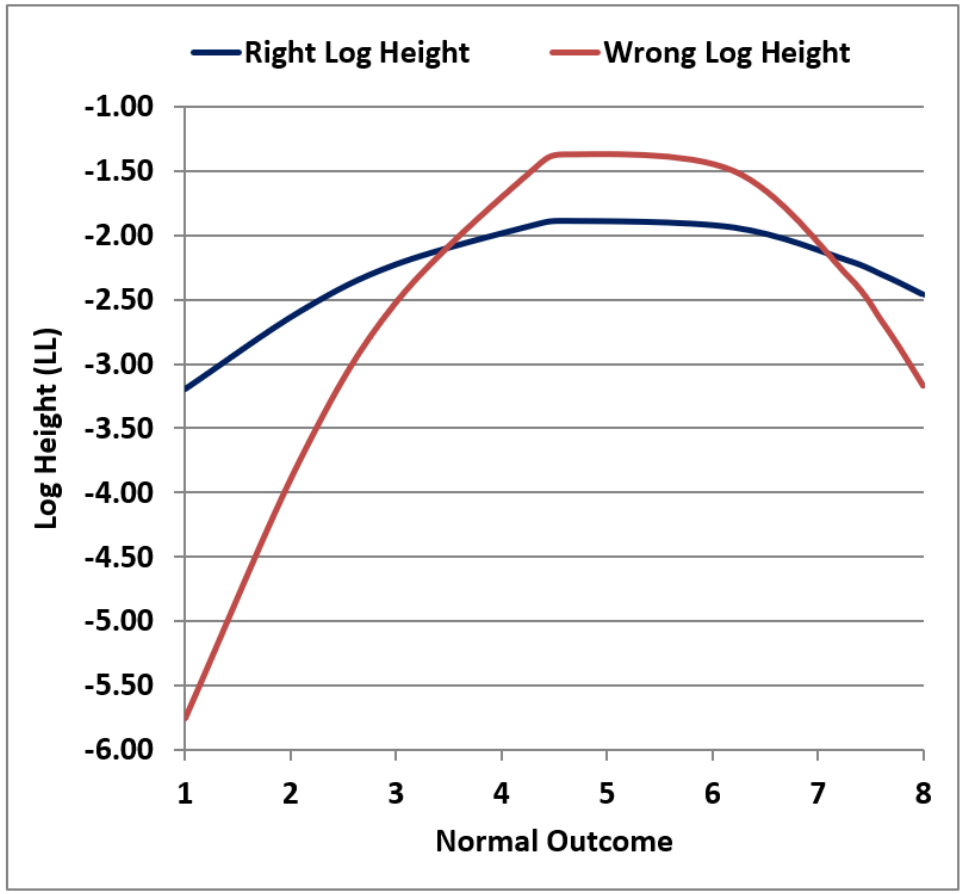
  - Two-level models use a multivariate normal PDF instead (that builds in a correlation of persons from the same cluster), which works the same way but is harder to visualize

ML via Excel NORMDIST: Example

Key idea: Normal Distribution formula → data height
Right answers = tallest possible function for everyone!

	Right Values	Wrong Values
Mean	5.19	5.24
Variance	6.56	2.00
ML Variance	*5.90*	*1.80*

Normal Outcome	Right Height	Right Log Height	Wrong Height	Wrong Log Height
1.0	0.04	-3.20	0.00	-5.76
2.1	0.08	-2.59	0.02	-3.73
3.0	0.11	-2.22	0.08	-2.52
4.3	0.15	-1.92	0.23	-1.49
4.6	0.15	-1.89	0.25	-1.37
6.2	0.14	-1.94	0.22	-1.50
7.3	0.11	-2.20	0.10	-2.33
7.6	0.10	-2.30	0.07	-2.66
7.8	0.09	-2.38	0.05	-2.90
8.0	0.09	-2.46	0.04	-3.17
SUM = ML LL = taller is better		-23.09		-27.42
ML: -2LL = smaller is better		46.19		54.84



Comparing Nested Models for the Variance

- Unlike **fixed effects** (which can always use Wald-type tests), testing a **random effect variance** requires assessing **relative model fit**: How well does the model fit relative to a nested model without that random effect variance?
  - When using **REML** estimation, the fixed effects must be kept the same (stay tuned)!

- Each model's fit is indexed by overall model **log-likelihood** ($\boldsymbol{LL}$):
  - Multivariate log-height for each cluster's outcomes given model parameters
  - Sum of log-heights across all (independent) clusters = **model $\boldsymbol{LL}$**

- We use a "**likelihood ratio test**" (LRT) to compare LL values across models, in which $fewer$ is the simpler model and $more$ is the augmented model:
  - Using likelihoods = L_{fewer}/L_{more}; using log-likelihoods = $LL_{fewer} - LL_{more}$
  - Either way, the result multiplied by -2 creates a χ^2 ("chi-square") test statistic

Comparing Nested Models for the Variance

Likelihood ratio test (LRT) = −2ΔLL test (aka, "χ^2 test" in SEM; "deviance 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 **−2ΔLL:**
   if given −2LL, use $-2\Delta LL = (-2LL_{fewer}) - (-2LL_{more})$
   if given LL, use $-2\Delta LL = -2 * (LL_{fewer} - LL_{more})$
2. Calculate **ΔDF =** $(\#\ Parms_{more}) - (\#\ Parms_{fewer})$
3. **Compare −2ΔLL to χ^2 distribution with numerator DF = ΔDF, get p-value**

Because −2 is needed, $-2LL$ may be given (as "deviance") instead of or in addition to LL

- But when testing random effect variances (that can't be negative), a "**mixture**" χ^2 distribution should be used (otherwise the test is slightly conservative)
  - e.g., Add a random intercept variance? DF is mixture of 1 (when positive) and 0 (when it would have been negative), so you can just cut the obtained p-value in half

LRT (−2ΔLL) Example: Single-Level vs. Two-Level Model for the Variance

```
print("Show conditional ICC1 and its LRT")
icc(Model3); ranova(Model3)

# Intraclass Correlation Coefficient

    Adjusted ICC: 0.216
  Unadjusted ICC: 0.214
ANOVA-like table for random-effects: Single term deletions

Model:
langpost ~ 1 + hw2 + mixgrd + (1 | schoolID)
               npar logLik   AIC LRT Df Pr(>Chisq)
<none>            5 -12590 25190
(1 | schoolID)    4 -12811 25630 442  1     <2e-16
```

χ^2 Critical value = 3.84 for DF =1 (or 2.71 for a mixture of DF = 0,1)

Conclusion? Two-level (with the random intercept) fits significantly better (or single-level model fits significantly worse), which means the conditional (leftover) $ICC1 = .216$ is significantly > 0.

Therefore, we believe the two-level model results over the single-level model results.

More about Comparing Models via $-2\Delta LL$

- What your *p*-value for the −2ΔLL test means:
  - **ADD** parameters? Model can get **better** (significant −2ΔLL) or **not better** (nonsig)
  - **DROP** parameters? Model can get **worse** (significant −2ΔLL) or **not worse** (nonsig)
    - Better model has bigger LL (as "height") or smaller $-2LL$ (as "shortness")

- Nested or non-nested (not subset) models can also be compared by **Information Criteria** that also reflect model parsimony
  - No wide-spread significance tests or critical values, just "smaller is better"
  - **AIC** = Akaike IC = $-2LL$ + 2 *(#parameters)
  - **BIC** = Bayesian IC = $-2LL + \log(N)$*(#parameters)
  - How N and #parameters are computed varies across software, however, because…

More about Comparing Models via $-2\Delta LL$

- Maximum likelihood estimation—and thus model LL values—come in 2 flavors:

- **"Maximum likelihood" (ML; also called FIML*)**
  - Is available for any model, even with non-continuous outcomes and latent variables
  - Maximizes the height of the full data, so any nested model comparison is valid
  - In small level-2 samples, ML under-estimates random effect variances (analogous to population vs sample formulas for computing variance) → $\frac{\sum(y_{pc} - \hat{y}_{pc})^2}{N}$

- **"Restricted (or residual) maximum likelihood" = REML (≈ OLS)**
  - Only available for general linear models or general linear mixed models (i.e., multilevel models with normally distributed residuals at all levels of analysis)
  - Maximizes the height of the residuals, so cannot be used to compare model fixed effects!
  - Estimates variances like OLS by accounting for k = number of estimated fixed effects (more accurate in smaller level-2 samples) → $\frac{\sum(y_{pc} - \hat{y}_{pc})^2}{N - k}$

- **FI = Full information → it uses all original data, and actually they both do*

LRTs using ML vs. REML in a nutshell

All comparisons must use exact same sample to be valid!!!	**ML**	**REML**
In estimating random effect variances, it treats fixed effects as…	**Known** (DF for having to estimate fixed effects is **not** factored in)	**Unknown** (DF for having to also estimate fixed effects **is** factored in)
So, in small level-2 samples, random effect variances (and the standard errors for associated fixed effects) will be…	**Too small** (but less of a difference as level-2 sample size increases)	**Unbiased** (correct)
But because it indexes the fit of the…	**Entire model** (fixed effects in means model + random effects in variance model)	**Variance model only**
You can use $-2\Delta LL$ to compare models differing in…	**Fixed and/or random effects** (either/both)	**Random effects only** (hold fixed effects constant)

Summary of Rules for Comparing Models

All observations must be the same across models!

Compare Models Differing In:

Type of Comparison:	**Means Model (Fixed Effects) Only**	**Variance Model (Random Effects) Only**	**Both Means Model and Variance Model (Fixed and/or Random)**
Nested models? YES, can do significance tests via…	Fixed effect *p*-values from ML or REML -- *OR* -- ML −2ΔLL only (NO REML −2ΔLL)	NO *p*-values REML −2ΔLL (ML −2ΔLL is ok if big N)	ML −2ΔLL only (NO REML −2ΔLL)
Non-nested models? NO significance tests, instead look for smaller…	ML AIC, BIC (NO REML AIC, BIC)	REML AIC, BIC (ML ok if big N)	ML AIC, BIC only (NO REML AIC, BIC)

Nested = one model is a direct subset of the other

Non-nested = one model is not a direct subset of the other

Summary

- By modeling cluster differences using **random effects** (by including at least a level-2 random intercept variance $\boldsymbol{\tau^2_{U_0}}$) instead of fixed effects (treating cluster as a categorical predictor in a single-level model), we can test fixed slopes of other level-2 between-cluster predictors
  - Wald tests ($t = Est/SE$ or F) for significance as usual (and $t \rightarrow d$ or partial r effect sizes)
  - Level-2 predictors explain level-2 random intercept variance ($\boldsymbol{\tau^2_{U_0}}$ = cluster mean differences)
    - Reduction in **level-2 intercept variance** can be quantified by **pseudo-$\boldsymbol{R^2_{U_0}}$** (per pile index)
    - Reduction in **total variance** can be quantified by **total-$\boldsymbol{R^2}$** = $\boldsymbol{r(\hat{y}_{pc}, y_{pc})^2}$ (overall index)

- We test the significance of random effect variances using likelihood ratio tests: $\chi^2 = -2$ times the difference in LL values between nested models
  - Using (the more accurate) REML instead of ML? Use the same fixed effects across models!
- Use random effects CIs to convey practical effect size instead of any R^2

Formative Assessment Questions

1. If patients from a single hospital clinic (seen by the same physician) are randomly assigned to one of two groups (treatment or control), how many "levels" of a model would you need to predict patient outcomes?
2. What if the previous study were extended to 50 hospitals, in which each 25 hospitals were treatment sites and 25 hospitals were control sites -- then how many levels of a model would be needed to predict patient outcomes?
3. What is the common underlying feature of general linear models (e.g., regression, ANOVA, and ANCOVA) or generalized linear models (e.g., logistic, ordinal, count regression) that makes them insufficient for multilevel data?
4. How is ICC1 computed? What are three potential interpretations of ICC1?
5. A colleague computes an ICC1 = .50 and concludes that a random intercept has explained 50% of the variance in their outcome. Is this correct? Why or why not?
6. What is the role of level-2 predictors in two-level models of persons nested in clusters?
7. What does the model log-likelihood convey about the outcome data? What will we use it for in multilevel models?
8. If you had to pick a "theme song" or "walk up song" for yourself (i.e., a song in any language that you'd envision playing loudly in an arena for you as a professional sports player), what would it be???