Introduction to Multilevel Models for Clustered Data

- Topics:
  - What do multilevel models do?
  - From single-level to multilevel empty means models
  - Intraclass correlation (ICC) and design effects
  - Fixed effects of level-2 predictors
  - Effect size for level-2 predictors

- By Lesa Hoffman, Professor of Educational Measurement and Statistics in the University of Iowa College of Education
  - Presented March 14, 2023, as part of the APA Free Science Trainings Series
  - Btw, my full course at Univ of Iowa on Clustered Multilevel Models is in Fall 2023!

Multilevel Models (MLMs) for Clustered* Data

- **Clustering = Nesting = Grouping = Hierarchies*
  - Key idea: Outcomes with >1 dimension of sampling simultaneously ("**micro**" units are nested in one or more types of "**macro**" units)
  - Each sampling dimension is considered its own "**level**" → **MLM**
  - MLMs can be used to predict outcomes from two-level (or more-level) sampling designs that result in nested and/or crossed observations

- The term "Multilevel Model" (MLM) has many synonyms:
  - **General Linear Mixed-Effects Models** (Fixed + Random = Mixed)
  - **Random Coefficients Models** (Random effects = latent variables)
  - **Hierarchical Linear Models** (HLM, but not = hierarchical regression)
  - Most MLM software is "univariate" → predict 1 outcome at a time
  - Multivariate MLMs can be estimated as "multilevel structural equation models" to predict 2+ outcomes at once (+ address missing predictors)

Examples of **Nested** Designs

- Examples of **two-level** sampling designs:
  - Students (level 1) nested in classes/teachers (level 2)
  - Patients (level 1) nested in doctors (level 2)
  - Citizens (level 1) nested in countries (level 2)

- Examples of **three-level** sampling designs:
  - Students (level 1) nested in classes/teachers (level 2) nested in schools (level 3)
  - Patients (level 1) nested in doctors (level 2) nested in hospitals (level 3)
  - Citizens (level 1) nested in survey years (level 2) nested in countries (level 3)

Examples of **Crossed** Designs

- Examples of **two-level cross-classified** sampling designs:
  - Two kinds of nesting: Students (level 1) nested in both schools (level-2) and neighborhoods (crossed at level 2)
  - Repeated measures: Responses (level 1) nested in both subjects (level 2) and items (crossed at level 2)
  - Reliability assessment: Ratings (level 1) nested in both raters (level 2) and targets (crossed at level 2)
  - Students who change classes over time: occasions (level 1) nested in both students (level 2) and classes (crossed at level 2)

- Example of **three-level cross-classified** sampling designs:
  - Ratings (level 1) nested in both children (level 2) and raters (crossed at level 2); raters are nested within sites (level 3)
  - Responses (level 1) nested in both students (level 2) and items (crossed at level 2); students are nested within schools (level 3)

Labels for Organizing Models

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

- **General Linear Models:** conditionally normal outcome distribution, **fixed effects** (identity link; only one dimension of sampling)

Note: Ordinary Least Squares is only for GLM

- **General*ized* Linear Models:** **any conditional outcome distribution**, **fixed** effects through **link functions**, no random effects (only one dimension)
- **General Linear Mixed Models:** conditionally normal outcome distribution, **fixed and random effects** (identity link, but **multiple dimensions** of sampling)
- **General*ized* Linear Mixed Models:** **any conditional outcome distribution**, **fixed and random effects** through **link functions** (**multiple dimensions**)
  - Same concepts as for general*ized* or mixed separately, but with more complexity in estimation

- "**Linear**" → fixed effects predict the *link-transformed* conditional mean of outcome in a **linear combination**: (effect*predictor) + (effect*predictor)…

Levels of Analysis in Two-Level Nested Data

- Between-Cluster (BC) Variation:
  - **Level-2** = "**INTER**-cluster differences" = cluster characteristics
- Within-Cluster (WC) Variation:
  - **Level-1** = "**INTRA**-cluster differences" = person characteristics
- **Any variable measured per person** could have both **L2 between and L1 within** variation!
  - BC = some clusters are higher/lower on average than other clusters
  - WC = some people are higher/lower than the rest of their cluster
  - Btw, univariate MLMs must address this differently for level-1 predictors vs. level-1 outcomes, but multivariate MLMs treat both the same way
    - *Stay tuned for APA Free Training 2 for level-1 person predictors!*
- **So how do MLMs "handle" multiple levels of sampling?**

The Two Sides of *Any* Model

- **Model for the Means:**
  - **Fixed Effects**, the "structural" part (= latent variables means)
  - What you are used to **caring about for testing hypotheses**
  - How the expected outcome for a given observation varies as a function of their values for the predictor variables

- **Model for the Variance:**
  - **Random Effects and Residuals**, the "stochastic" or "error" part
    - Btw, random effect variances = latent variable variances
  - What you are used to **making assumptions about** instead
  - How residuals are distributed and related across observations (persons, clusters, items, etc.) → these relationships are called "dependency" and ***this is the primary way that multilevel models differ from general linear models*** *(GLMs; "regression")*

Two Sides of a General Linear Model (GLM)

p = person

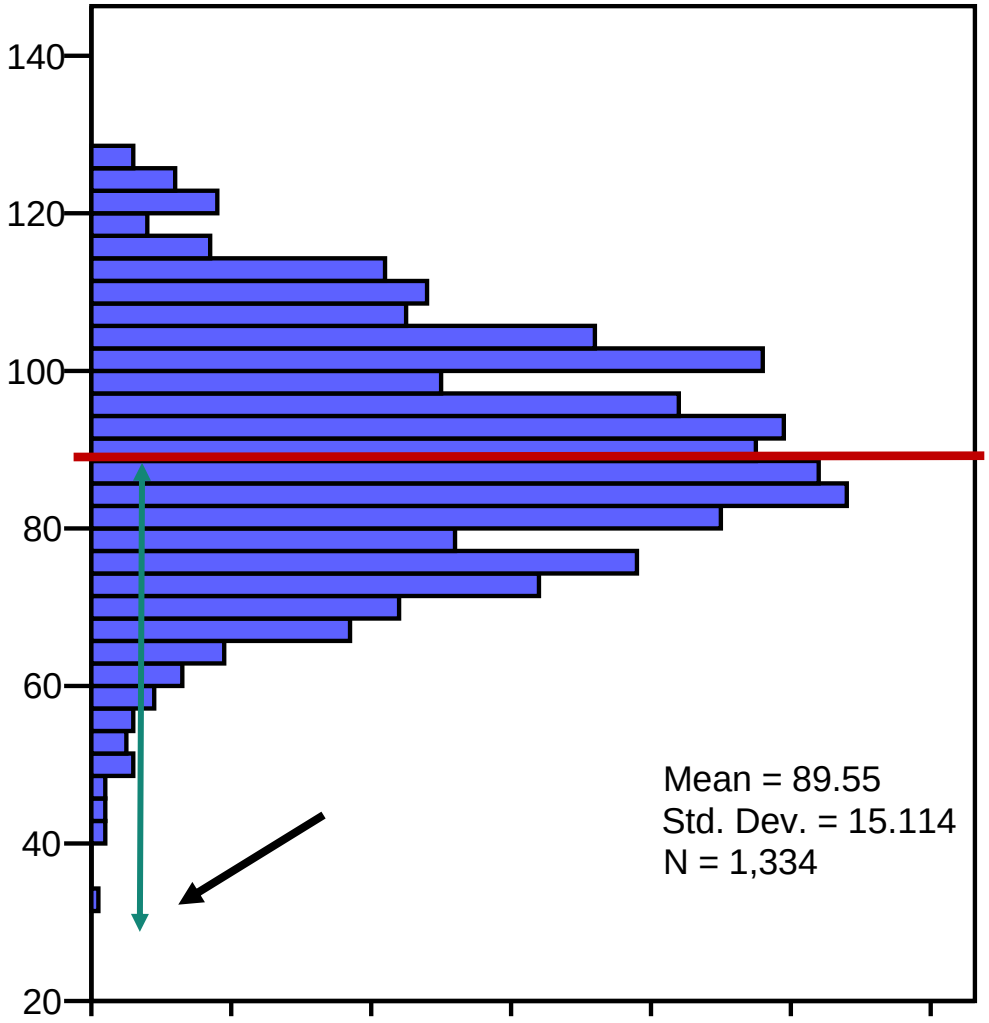
$$y_p = \beta_0 + \beta_1(x1_p) + \beta_2(x2_p) + \cdots + e_p$$

Our focus

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

- **Model for the Variance (→ "Piles" of Variance):**
  - $\boldsymbol{e_p \sim N(0, \sigma_e^2)}$ → ONE (between-person) source of unexplained variation
  - In GLMs, e_p has a mean of 0 with some estimated constant variance σ_e^2, is normally distributed, is unrelated to $x1_p$ and $x2_p$, and is **independent** across all observations (which is just one outcome per person here)
  - **There is only ONE source of residual variance in the above GLM because it was designed for 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

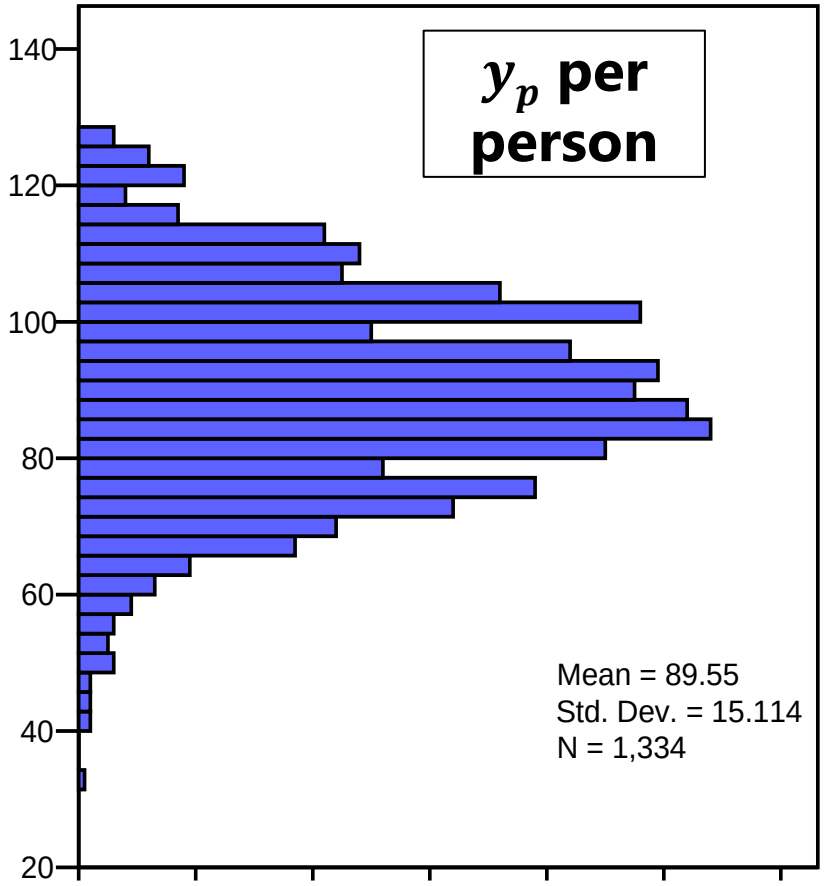
y_p residual variance:

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

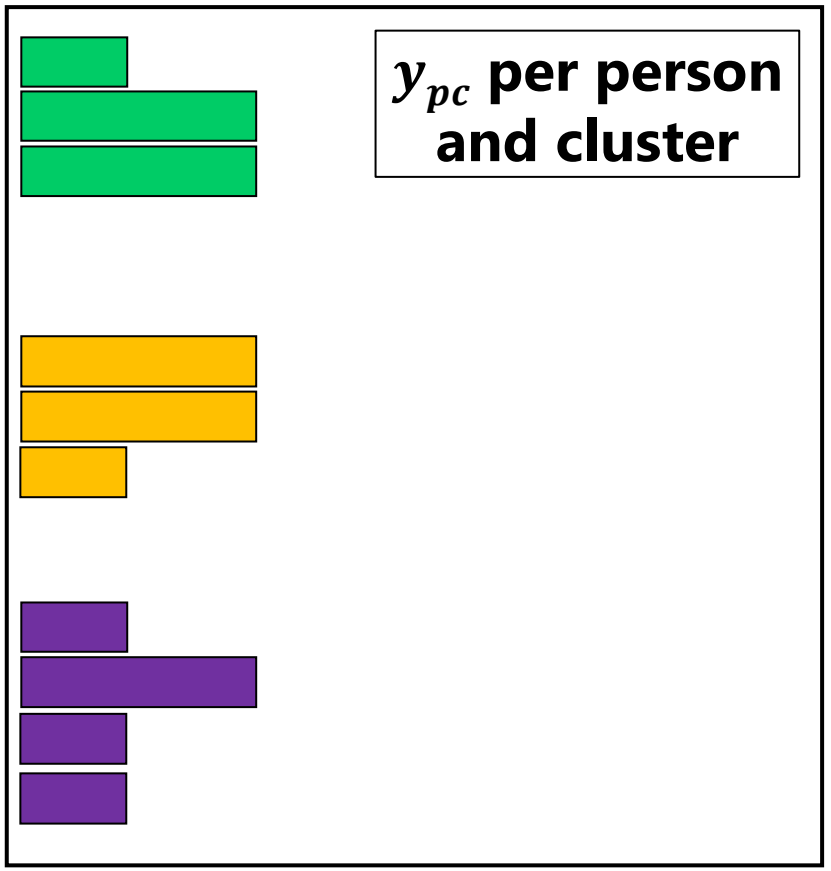
Without predictors, $\sigma_e^2 = \sigma_y^2$

Adding Multiple Persons per Cluster → **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 outcome's value:

= Grand Mean

→ **Fixed Intercept**

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

= Cluster Mean

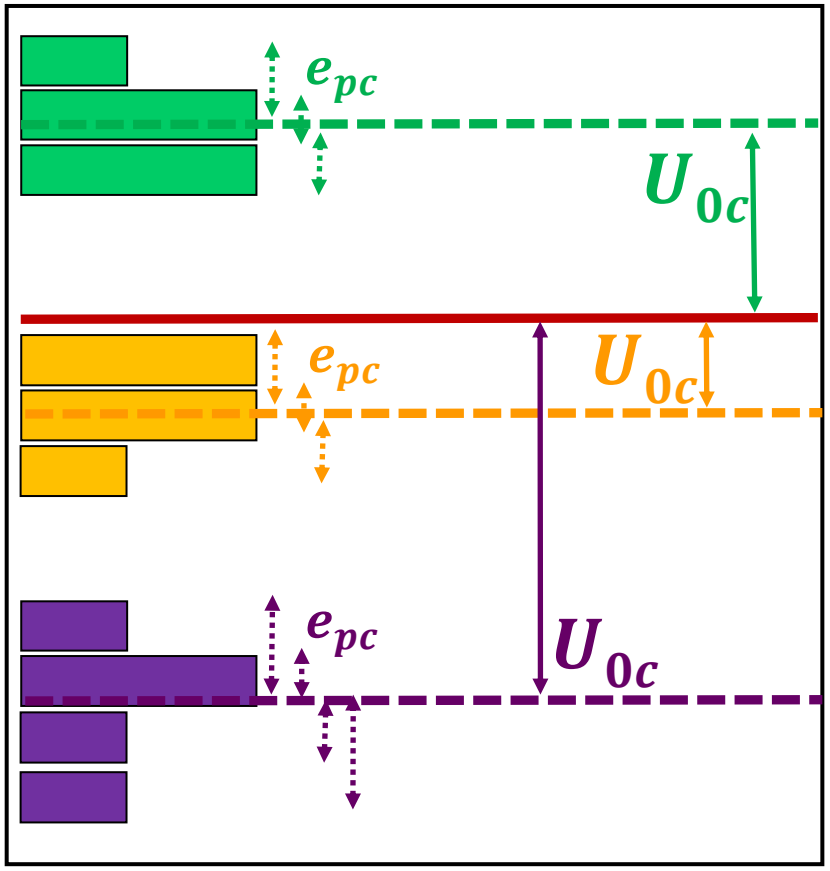
→ **Random Intercept**

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:

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

- Level 2:

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

3 total parameters:

Model for the Means (1):

- Fixed Intercept γ_{00}

Model for the Variance (2):

- Level-1 **WC** Variance of e_{ti} → σ_e^2
- Level-2 **BC** Variance of U_{0i} → 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)

$$\text{ICC} = \frac{\text{BC}}{\text{BC} + \text{WC}} = \frac{\text{L2 Intercept Var}}{\text{L2 Intercept Var} + \text{L1 Residual Var}}$$

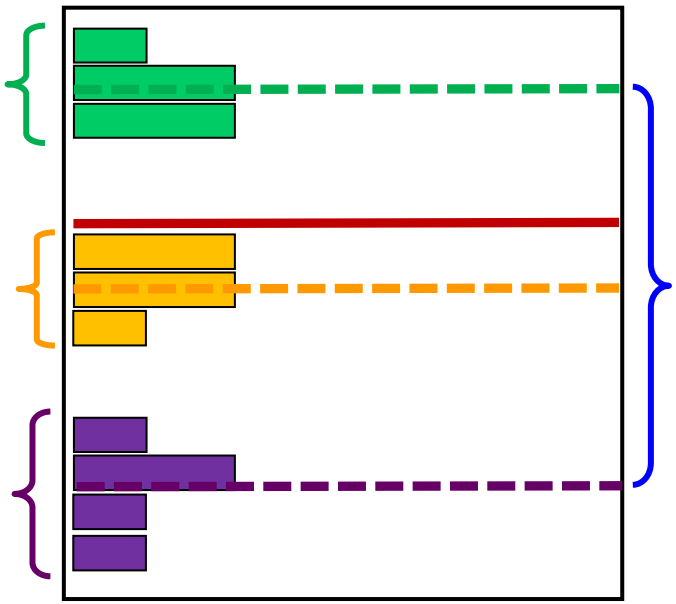
$$= \frac{\boldsymbol{\tau^2_{U_0}}}{\boldsymbol{\tau^2_{U_0} + \sigma^2_e}}$$

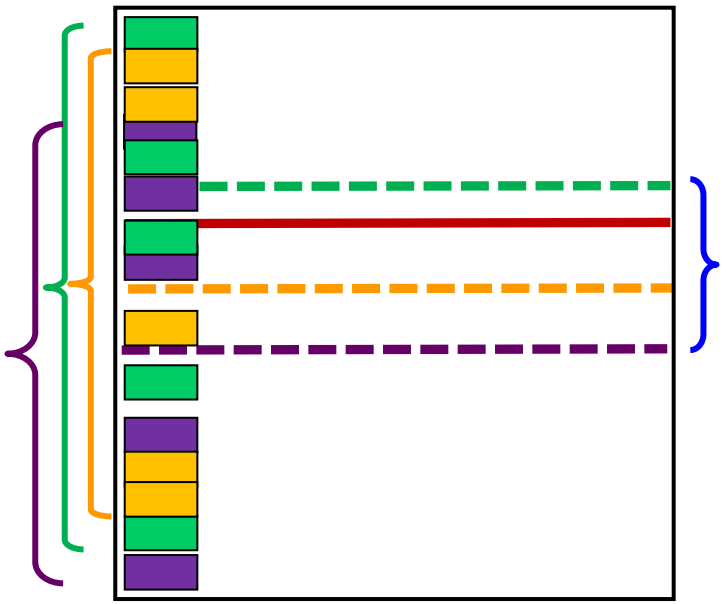
$\boldsymbol{\tau^2_{U_0}}$ → *Why don't all clusters have the same mean?*

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

- ICC = Proportion of total variance that is between clusters
- ICC = Average correlation of persons from same cluster
- ICC is a standardized way of expressing how much *dependency (correlation)* there is due to cluster mean differences
  → ICC is an effect size for *constant* cluster dependency
  - Dependency of other kinds can still be created by differences across clusters in the slopes of person predictors (stay tuned for part 2!)
- Btw, no variance has been "explained" yet (just 2 kinds of "error")

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





ICC = BTW / BTW + within

→ Large ICC

→ Large correlation within clusters

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

ICC = btw / btw + WITHIN

→ Small ICC

→ Small correlation within clusters

Effects of Clustering on Effective N

- **Design Effect** expresses how much sample size needs to be adjusted due to clustering → "**effective sample size**"
- **Design Effect** = ratio of the variance using a given sampling design to the variance using a simple random sample from the same population, given the same total sample size either way

- Design Effect = $1 + ([L1n - 1] * \text{ICC})$ $\boxed{L1n = \text{\# level-1 units}}$
- Effective sample size → $\text{Effective } N = \frac{\text{\# Total Observations}}{\text{Design Effect}}$

- As ICC and 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 = $1 + ([L1n - 1] * \text{ICC})$
- Effective sample size → $\text{Effective } N = \frac{\text{\# Total Observations}}{\text{Design Effect}}$

- $n = 5$ patients from each of 100 doctors, ICC = .30?
  - Patients Design Effect = 1 + ([5−1] * .30) = 2.20
  - **Effective N** = 500 / 2.20 = **227** (not 500)

- $n = 20$ students from each of 50 schools, ICC = .05?
  - Students Design Effect = 1 + ([20−1] * .05) = 1.95
  - **Effective N** = 1000 / 1.95 = **513** (not 1000)

Does a non-significant ICC mean you can ignore clustering and just do a regression?

- As ICC and cluster size go UP, effective N goes DOWN
  - So there is NO VALUE OF ICC that is "safe" to ignore, not even ~0!
  - An ICC=0 in an *empty means (unconditional)* model can become ICC>0 after adding person predictors because reducing the residual variance will then increase the random intercept variance (→ *conditional* ICC > 0)
    - 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 person predictor fixed slope SEs → **biased SEs**
    - Potential for **contextual effects** of person predictors (stay tuned!)
      - A "clustered-sampling correction" to the SEs **will not fix this problem**!

2 Options for Cluster Differences

Represent Cluster Differences via Fixed Effects

- Include (#clusters−1) binary predictors for cluster membership in the **model for the means** → **so cluster is NOT a model "level"**
  - Main effects control for cluster mean differences only; interactions with person predictors are also needed to control for cluster slope differences
- Useful if #clusters < 10ish or you care about specific clusters, but then you cannot include cluster predictors → saturated mean diffs

Represent Cluster Differences via Random Effects

- Include a random intercept variance across clusters in the **model for the variance** → **then cluster IS a model "level"**
  - A random intercept controls for cluster mean differences only; a random slope variance is needed for cluster differences in person predictor slopes
- Better if #clusters > 10ish or you want to **predict** cluster differences
- So let's see an example!

Empty Means, Random Intercept Model: (1b) Syntax by Univariate MLM Program

SAS:

```
PROC MIXED DATA=work.Example COVTEST NOCLPRINT IC METHOD=REML;
     CLASS schoolID;
     MODEL langpost = / SOLUTION DDFM=Satterthwaite;
     RANDOM INTERCEPT / VCORR TYPE=UN SUBJECT=schoolID; * VCORR gives ICC;
RUN;
```

R lmer from lme4 package—*using lmerTest package to get Satterthwaite denominator DF, and using performance package to get ICC from lmer:*

```
name = lmer(data=Example, REML=TRUE, formula=langpost~1+(1+|schoolID))
summary(name, ddf="Satterthwaite")
icc(name); ranova(name) # ICC and LRT for random intercept
```

STATA:

```
mixed langpost , || schoolID: , ///
      reml dfmethod(satterthwaite) dftable(pvalue) nolog
estat icc // Get ICC
```

SPSS:

```
MIXED langpost BY schoolID
      /METHOD   = REML
      /CRITERIA = DFMETHOD(SATTERTHWAITE)
      /PRINT    = SOLUTION TESTCOV
      /FIXED    =
      /RANDOM = INTERCEPT | COVTYPE(UN) SUBJECT(schoolID).
```

Example: Level-1 Students in Level-2 Schools

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

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

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

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

22% of total language variance is due to school mean differences (WC r = .22)

Results from SAS MIXED:

Without random intercept U_{0c}:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	41.4635	0.1476	3565	280.91	<.0001

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	V…	Pr	Z
Residual	77.6905	1.8402	42…	…001	

With random intercept U_{0c}:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	41.0791	0.3371	175	121.87	<.0001

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	V…	P…	Z
UN(1,1)	schoolID	17.8085	2.3063	7.72	0	1
Residual		62.2296	1.5179	4…	…001	

Adding Level-2 Cluster Predictors

- **Level-2 predictors** are constant over persons from the same cluster—they are cluster-level characteristics
  - Example: Level-1 (L1) students (p) nested in level-2 (L2) schools (c) that vary in homework amount and mixed grades (0=no, 1=yes)
- **Level-1:** $Lang_{pc} = \beta_{0c} + e_{pc}$

σ_e^2 → All possible L1 residual variance for within-school differences across students

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

$\tau_{U_0}^2$ → **All possible L2 random intercept variance** due to school mean differences

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

$\tau_{U_0}^2$ → L2 random intercept variance **leftover** after HW and mixed grade prediction

Adding Level-2 Cluster Predictors: (1c) Syntax by Univariate MLM Program

SAS:

```
PROC MIXED DATA=work.Example COVTEST NOCLPRINT IC METHOD=REML;
     CLASS schoolID;
     MODEL langpost = hw2 mixgrd / SOLUTION DDFM=Satterthwaite;
     RANDOM INTERCEPT / TYPE=UN SUBJECT=schoolID;
RUN;
```

R lmer from lme4 package—*using lmerTest package to get Satterthwaite denominator DF:*

```
name = lmer(data=Example, REML=TRUE, formula=langpost~1+hw2+mixgrd+(1+|schoolID))
summary(name, ddf="Satterthwaite")
```

STATA:

```
mixed langpost c.hw2 c.mixgrd, || schoolID: , ///
      reml dfmethod(satterthwaite) dftable(pvalue) nolog
```

SPSS:

```
MIXED langpost BY schoolID WITH hw2 mixgrd
      /METHOD   = REML
      /CRITERIA = DFMETHOD(SATTERTHWAITE)
      /PRINT    = SOLUTION TESTCOV
      /FIXED    = hw2 mixgrd
      /RANDOM   = INTERCEPT | COVTYPE(UN) SUBJECT(schoolID).
```

Example: Adding Level-2 Cluster Predictors

Level-1 (L1): $Lang_{pc} = \beta_{0c} + e_{pc}$

Level-2 (L2): $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + U_{0c}$

Before adding L2 predictors:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	41.0791	0.3371	175	121.87	<.0001

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z
UN(1,1)	schoolID	17.8085	2.3063	
Residual		62.2296	1.5179	

After adding L2 predictors:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	41.5479	0.5051	168	82.26	<.0001
hw2	0.5068	0.6352	172	0.80	0.4261
mixgrd	-1.9931	0.7083	189	-2.81	0.0054

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z
UN(1,1)	schoolID	17.1640	2.2341	
Residual		62.2013	1.5165	

Example: Adding Level-2 Cluster Predictors

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

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

Model for the Means:

- γ_{00} **= 41.55** = fixed **intercept**: expected language for students in a school with homework=2 (~mean) and mixgrd=0 (=not mixed)
- γ_{01} **= 0.51** = fixed slope of **HW−2**: difference in school mean language for each unit more homework the school tends to assign
- γ_{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 = deviation between actual and predicted school mean language for school c (with variance $\tau^2_{U_0}$ **= 17.16**)
- e_{pc} = level-1 residual = deviation of the actual outcome for student p from their outcome predicted by β_{0c} (with variance σ^2_e **= 62.20**)

What if we used single-level GLM instead?

Level-1 (L1): $Lang_{pc} = \beta_{0c} + e_{pc}$

Level-2 (L2): $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + U_{0c}$

Without random intercept U_{0c}:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	41.6370	0.2163	3563	192.50	<.0001
hw2	0.4110	0.2796	3563	1.47	0.1416
mixgrd	-1.4745	0.3472	3563	-4.25	<.0001

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	V…	P… Z
Residual	77.2515	1.8303	42…	…001

With random intercept U_{0c}:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	41.5479	0.5051	168	82.26	<.0001
hw2	0.5068	0.6352	172	0.80	0.4261
mixgrd	-1.9931	0.7083	189	-2.81	0.0054

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	V…	P… Z
UN(1,1)	schoolID	17.1640	2.2341	68	…
Residual		62.2013	1.5165	41…	…001

Effect Size for Level-2 Cluster Predictors

- Direct: convert t-statistic for fixed effect into d or partial r
  - $d = \frac{2t}{\sqrt{DF_{den}}}, \quad r = \frac{t}{\sqrt{t^2 + DF_{den}}}$

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*
    - Pseudo-$R^2 = \frac{variance_{fewer} - variance_{more}}{variance_{fewer}}$

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

    - It 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 with model-implied total variance instead (see Rights & Sterba, 2019; 2020)
    - Only pseudo-R^2 for the L2 random intercept var is relevant for L2 predictors
  - **Total-R^2**: amount of total variance explained (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

Example Conditional MLM Effect Size

- 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 **ICC = .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

- **Variance explained** by the two level-2 cluster fixed slopes:
  - **Pseudo-R^2_{U0}** $= \frac{17.809-17.164}{17.809}$ **= .036 → 3.6%** of the level-2 random intercept variance (due to school mean differences) in language
  - **Total-R^2 = correlation($\hat{y}_{pc}$, y_{pc})² = .079² = .006** → 0.6% of the total variance in language (22.3% of which was due to school mean diffs)
    - Total-R^2 approximation when there is only a random intercept:
      Total-R^2 = (Pseudo-$R^2_{U_0}$* ICC) + (Pseudo-R^2_e* [1−ICC]) = .008 (see excel)
  - Because these R^2 values mean very different things, it is essential to clearly describe how you calculated them and what they then mean!

Part 1: Summary

- MLMs begin with an empty model to determine how much outcome variance is attributable to each dimension of sampling:
  - Level-2 between-cluster mean differences → random intercept ($\boldsymbol{\tau^2_{U_0}}$)
  - Level-1 within-cluster person differences → residual ($\boldsymbol{\sigma^2_e}$)
  - Dependency effect size via Intraclass Correlation: $\mathbf{ICC} = \boldsymbol{\tau^2_{U_0}}$ **/** $(\boldsymbol{\tau^2_{U_0}} + \boldsymbol{\sigma^2_e})$
    - ICC = proportion of total variance due to cluster mean differences
    - ICC = average correlation of persons from same cluster
    - Higher ICC and level-1 sample size → larger design effect → smaller effective N

- Modeling cluster differences using random effects (by including $\boldsymbol{\tau^2_{U_0}}$ at a minimum, possibly random slope variances, stay tuned!) allows us to test the effects of level-2 between-cluster predictors
  - Significance tests via Wald tests (usually with denominator DF) as usual
  - Adding fixed slopes for level-2 predictors (cluster characteristics) can explain level-2 random intercept variance (cluster mean differences)
    - Reduction in **level-2 intercept variance** can be quantified by **pseudo-**$\boldsymbol{R^2_{U_0}}$
    - Reduction in **total variance** can be quantified by **total-**$\boldsymbol{R^2}$ (≈ pseudo-$R^2_{U_0}$* ICC)

Foreshadowing Part 2

- **Part 2 on Thursday March 16:** Adding Level-1 Predictors to Multilevel Models for Clustered Data
  - Fixed slopes of level-1 person predictors
    - Cluster-mean-centering, constant-centering, and latent centering
  - Random slopes of level-1 person predictors
  - Cross-level interactions and systematically varying effects

- Thank you for your attention!

Begin Bonus Material

- Significance testing for each side of the model
- Likelihood ratio tests and information criteria
- Maximum likelihood (ML) vs.
  residual maximum likelihood (REML)
- Model comparisons in ML vs. REML
- Why explaining level-1 residual variance will increase level-2 random intercept variance (and the design effect)

Relative Model Fit by Model Side

- Nested models (i.e., in which one is a subset of the other) can now differ from each other in two distinct ways

- **Model for the Means** → which predictors and which fixed slopes for them are included in the model
  - **Does not** require assessment of relative model fit using −2LL because we can still use univariate or multivariate Wald tests for this (although we have more choices for denominator degrees of freedom)

- **Model for the Variance** → what the pattern of variance and covariance of residuals from the same sampling unit should be
  - **DOES** require assessment of relative model fit using −2LL
  - Cannot use the Wald test *p*-values (even if they show up on the output) for testing the significance of variances because those *p*-values use a two-sided sampling distribution for what the variance could be (but variances cannot be negative, so those *p*-values are not valid)

Significance of Fixed Effects in MLM

	Denominator DF is infinite (Proper Wald test)	**Denominator DF is estimated instead** ("Modified" Wald test)
Numerator DF = 1 *(test one fixed effect)* is **Univariate Wald Test**	use $\boldsymbol{z}$ distribution (Mplus, STATA default)	use $\boldsymbol{t}$ distribution (SAS, SPSS, STATA with dfmethod option)
Numerator DF > 1 *(test 2+ fixed effects)* is **Multivariate Wald Test**	use $\boldsymbol{\chi^2}$ distribution (Mplus, STATA default)	use $\boldsymbol{F}$ distribution (SAS, SPSS, STATA with dfmethod option)
Options for estimating Denominator DF (DDF)	not applicable	SAS, STATA: Kenward-Roger SAS, STATA, SPSS: **Satterthwaite**

In R, the default and optional DDF behavior vary across packages:

- Kenward-Roger and Satterthwaite are available through the lmerTest package (for use with the lmer function from the lme4 package)
- Satterthwaite DDF may not always work in nlme package (gls or lme functions)

Denominator DF (DDF) Methods

- **Between-Within** (DDFM=BW in SAS, REPEATED in STATA):
  - Total DDF comes from total number of observations, separated into level-2 for L2*n* clusters and level-1 for L1*n* persons (like in RM ANOVA)
    - **Level-2 DDF** = L2*n* – #level-2 fixed effects
    - **Level-1 DDF** = Total DDF – Level-2 DDF – #level-1 fixed effects
    - Level-1 effects with random slopes still get level-1 DDF

- **Satterthwaite** (DDFM=Satterthwaite in SAS and STATA, available in LME and LMER in R, default in SPSS):
  - More complicated, but analogous to two-group *t*-test given unequal residual variances and unequal group sizes
  - Incorporates contribution of variance components at each level
    - Level-2 DDF will resemble Level-2 DDF from BW method
    - Level-1 DDF will resemble Level-2 DDF from BW method if the level-1 effect also has a random slope, but it will resemble level-1 DDF otherwise

Denominator DF (DDF) Methods

- **Kenward-Roger** (DDFM=KR in SAS, KROGER in STATA, available in LME and LMER in R, available in SPSS 26+):
  - Adjusts the asymptotic covariance matrix of the fixed effects to reflect the uncertainty introduced by using large-sample techniques of maximum likelihood estimation in small L2*n* samples
  - This creates different (larger) SEs for the fixed effects
  - Then uses Satterthwaite DDF, new SEs, and *t* to get *p*-values

- Differences in inference not likely to matter often in practice unless sample sizes are very small
  - e.g., critical *t*-value at DDF=20 is 2.086, at infinite DDF is 1.960 instead

- When in doubt, use KR (is overkill at worst, becomes Satterthwaite)
  - I use Satterthwaite in my teaching for comparability across programs

Comparing Models for the Variance

- Unlike **fixed effects** (which can always use Wald-type tests), testing **random effects** requires assessment of **relative model fit**: how well does the model fit relative to other possible models?

- Model fit is indexed by overall model **log-likelihood** (**LL**):
  - Multivariate height for each cluster's outcomes given model parameters
  - Sum heights across all (independent) clusters = **model LL**
  - Two flavors in MLM: Maximum Likelihood (ML) or Restricted ML (REML)

- What you get for this on your output varies by software…

- Given as −2*log likelihood (−2LL) in SAS or SPSS MIXED, some R: **−2LL** gives BADNESS of fit, so **smaller** value = better model

- Given as just log-likelihood (LL) in STATA MIXED and Mplus, some R: **LL** gives GOODNESS of fit, so **bigger** value = better model

Comparing Models for the Variance

- Nested models are compared using a **"likelihood ratio test": −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. & 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**
4. Get *p*-value (from CHIDIST in excel, LRTEST in STATA, R/ANOVA in R)

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

Comparing Models for the Variance

- What your *p*-value for the −2ΔLL test means:
  - If you **ADD** parameters, then your model can get **better** (if −2ΔLL test is significant) or **not better** (not significant)
  - If you **REMOVE** parameters, then your model can get **worse** (if −2ΔLL test is significant) or **not worse** (not significant)

- Nested or non-nested models can also be compared by **Information Criteria** that also reflect model parsimony
  - No significance tests or critical values, just "smaller is better"
  - **AIC** = Akaike IC = −2LL + 2 *(#parameters)
  - **BIC** = Bayesian IC = −2LL + log(N)*(#parameters)
  - What "parameters" means depends on flavor (not in R or STATA!):
    - ML = ALL parameters; REML = variance model parameters only

Flavors of Maximum Likelihood

- For MLMs, maximum likelihood estimation comes in 2 flavors:
- **"Restricted (or residual) maximum likelihood"**
  - Only available for general linear models or general linear mixed models (key: based on normally distributed residuals at all levels of analysis)
  - **REML = OLS** given complete outcomes, but it doesn't require them
  - Estimates variances the same way as in OLS (accurate) → $\frac{\sum(y_{pc} - \hat{y}_{pc})^2}{N - k}$
- **"Maximum likelihood" (ML; also called FIML*)**
  - Is more general, is available for all of the above, as well as for non-normal outcomes and models with latent variables (CFA/SEM/IRT/DCM)
  - Is NOT equivalent to OLS: It under-estimates variances by not accounting for number of estimated fixed effects → $\frac{\sum(y_{pc} - \hat{y}_{pc})^2}{N}$
- **FI = Full information → it uses all original data (they both do)*

LRTs using ML vs. REML in a nutshell

All comparisons must use exact same sample to be valid!!!	**ML**	**REML**
To select, type…	METHOD=ML (-2 log likelihood)	METHOD=REML *default* (-2 res log likelihood)
In estimating variances, it treats fixed effects as…	**Known** (DF for having to also estimate fixed effects is **not** factored in)	**Unknown** (DF for having to also estimate fixed effects **is** factored in)
So, in small samples, L2 variances will be…	**Too small** (but less of a difference after Level-2 sample size = 100 or so)	**Unbiased** (correct)
But because it indexes the fit of the…	**Entire model** (means + variances)	**Variance model only**
You can compare models differing in…	**Fixed and/or random effects** (either/both)	**Random effects only** (same fixed effects)

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 and Variance Model (Fixed and Random)**
Nested? 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? NO signif. tests, instead see…	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

Increases in Random Intercept Variance

- Level-2 random intercept variance $\tau^2_{U_0}$ will often increase as a consequence of reducing level-1 residual variance σ^2_e
- **Observed level-2 $\tau^2_{U_0}$** is NOT just between-cluster variance
  - Also has a small part of within-cluster variance (**level-1 σ^2_e**), or:

    Observed $\tau^2_{U_0}$ = True $\tau^2_{U_0}$ + $(\sigma^2_e/L1n)$
    - With increasing $L1n$ perons, bias in true $\tau^2_{U_0}$ due to level-1 σ^2_e is minimized
  - Model estimates of "**True**" $\tau^2_{U_0}$ use **$(\sigma^2_e/L1n)$** as correction factor:

    True $\tau^2_{U_0}$ = Observed $\tau^2_{U_0}$ – $(\sigma^2_e/L1n)$

- e.g., **Observed level-2 $\tau^2_{U_0}$ = 4.65**, **level-1 σ^2_e = 7.06**, **$L1n$ = 4**
  - **True $\tau^2_{U_0}$** = **4.65** –(**7.06/4**) = **2.88** in empty means model
  - Add L1 within-cluster predictor → reduce σ^2_e from **7.06** to **2.17**
  - But now **True $\tau^2_{U_0}$** = **4.65** –(**2.17/4**) = **4.10** → more dependency!