

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

- Topics:
 - From single-level to multilevel empty means models
 - Intraclass correlation (ICC) and design effects
 - Cluster mean reliability and random intercept CIs
 - Comparing models for the variance using LRTs
 - Fixed slopes of level-2 predictors
 - Denominator DF and variance explained

Two Sides of a General Linear Model (GLM)

$$\boxed{p = \text{person}} \quad y_p = \boxed{\beta_0 + \beta_1(x1_p) + \beta_2(x2_p) + \dots} + \boxed{e_p}$$

Our focus

• Model for the Means (→ Predicted Values):

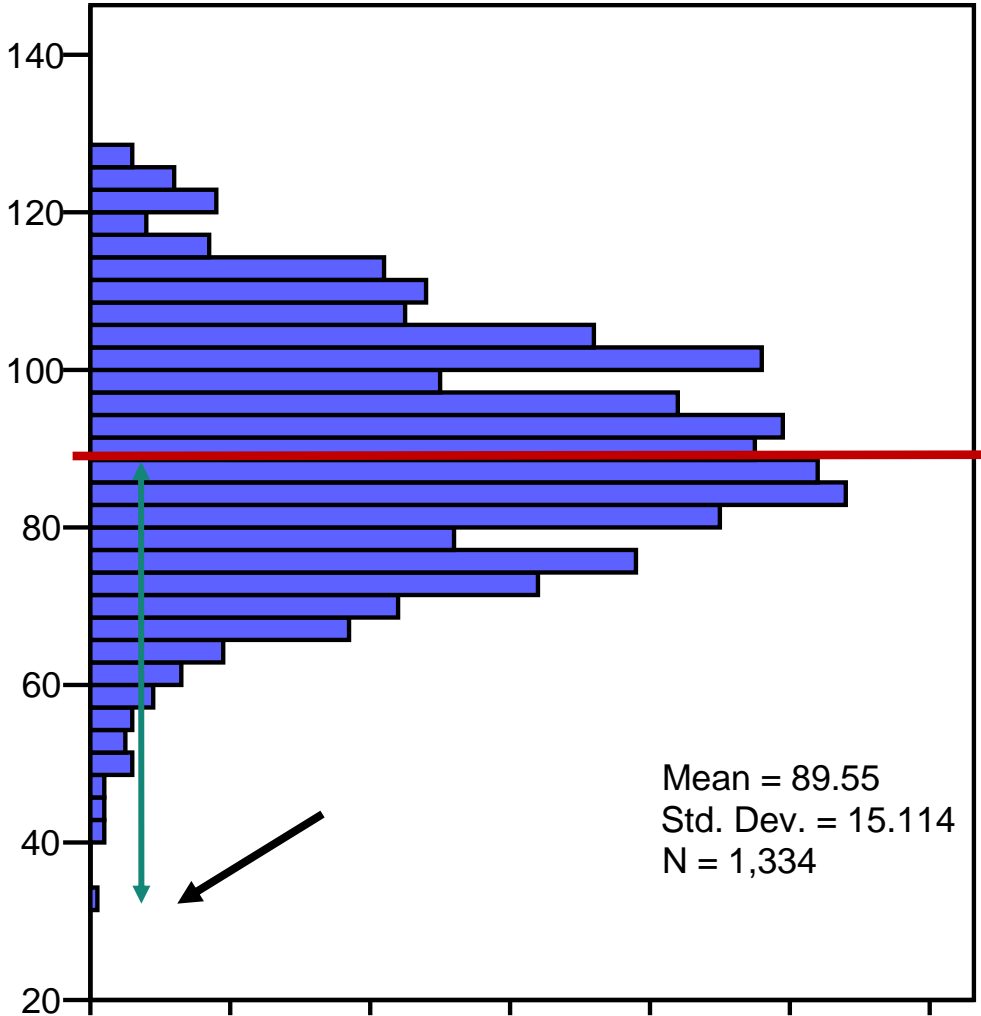
- Each person's expected (predicted) outcome is a weighted linear function of his/her values on $x1_p$ and $x2_p$ (and any other predictors); each variable is measured once per person (given by the p subscript)
- **Estimated β constants** are called **fixed effects** (intercept or slopes)

• Model for the Variance (→ "Piles" of Variance):

- $e_p \sim N(\mathbf{0}, \sigma_e^2) \rightarrow$ 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

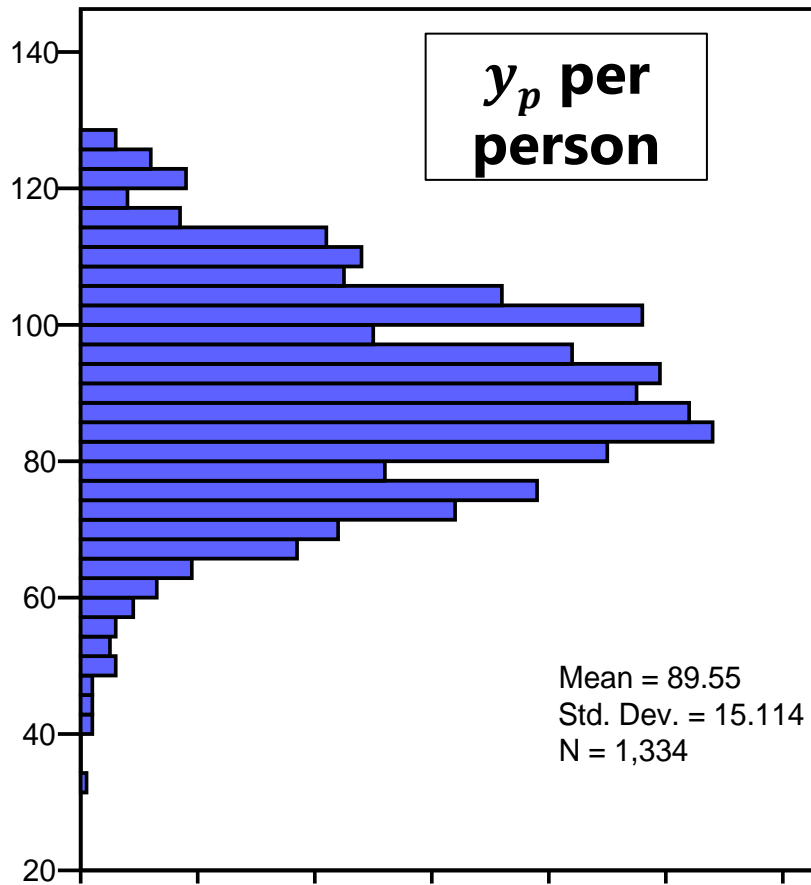
y_p residual variance:

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

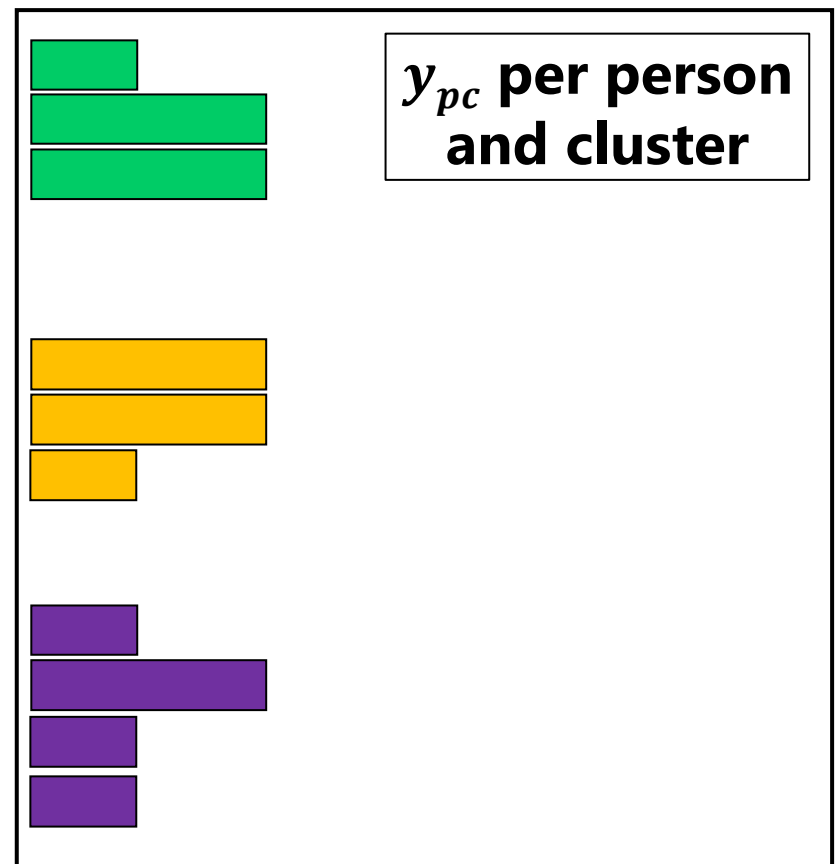
Without predictors, $\sigma_e^2 = \sigma_y^2$

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:

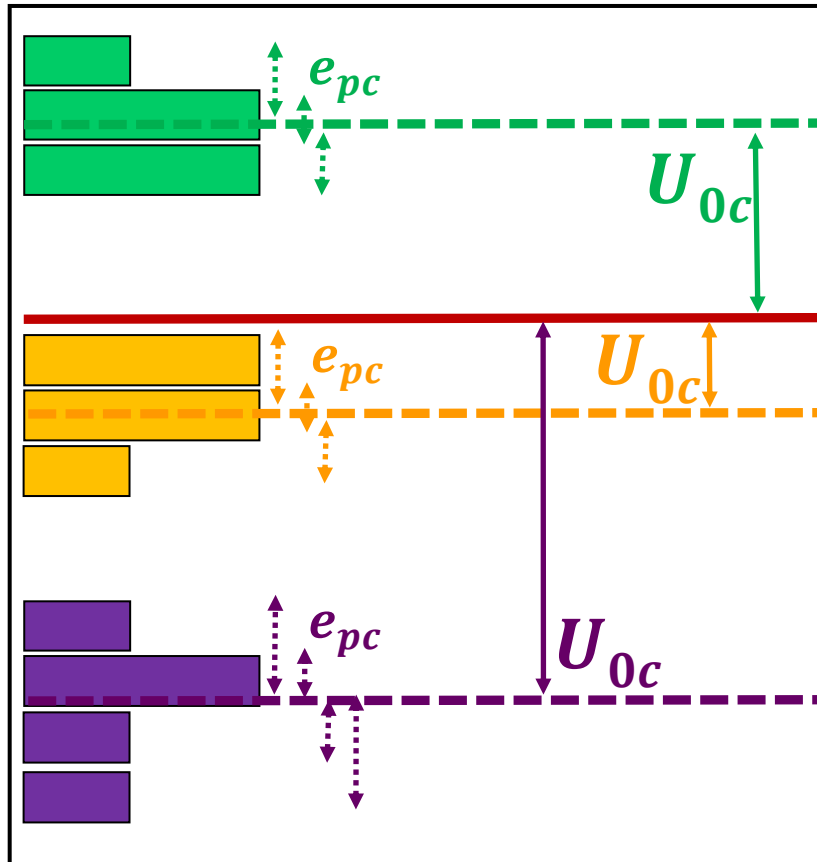
= Cluster Mean

→ **Random Intercept**

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

$\beta_0 \rightarrow$ mean of cluster means
 y_{pc} variance \rightarrow 2 sources:

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



Level-2 Random Intercept U_{0c}
(with variance labeled $\tau_{U_0}^2$):

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

Level-1 Residual e_{pc} per person
(with variance labeled σ_e^2):

- **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 $e_{pc} \rightarrow \sigma_e^2$
- Level-2 **BC** Variance of $U_{0c} \rightarrow \tau_{U_0}^2$

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:

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

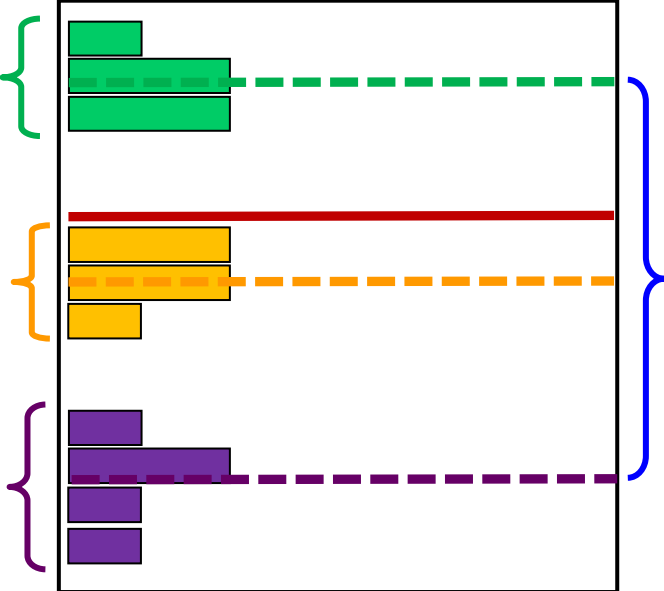
Intraclass Correlation (“ICC”=ICC1)

$$\text{ICC} = \frac{\text{BC}}{\text{BC} + \text{WC}} = \frac{\text{L2 Intercept Var}}{\text{L2 Intercept Var} + \text{L1 Residual Var}}$$
$$= \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + \sigma_e^2}$$

$\tau_{U_0}^2 \rightarrow$ Why don't all clusters have the same mean?
 $\sigma_e^2 \rightarrow$ 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**
 - Non-constant cluster dependency can still be created by differences across clusters in the slopes of level-1 person predictors
- Btw, no variance has been “explained” yet (just 2 kinds of “error”)

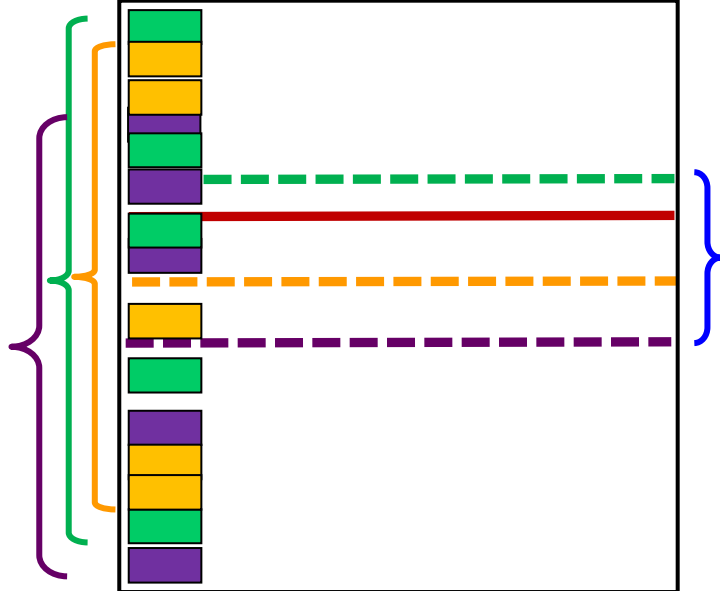
Even though **between-cluster variance** is the numerator, **ICC = within-cluster correlation!**



ICC = BTW / BTW + within

- Large ICC
- Large correlation within clusters

$$ICC = \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + \sigma_e^2}$$



ICC = btw / btw + WITHIN

- Small ICC
- Small correlation within clusters

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).
```

Electronic materials for this example from my 2023 APA training sessions are [here](#)

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

$$ICC = \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + \sigma_e^2} = \frac{17.809}{17.809 + 62.230} = .223$$

22.3% 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	Value	Pr > Z	Z
Residual	77.6905	1.8402	42	<.0001	

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	Value	Pr > Z	Z
UN(1,1)	schoolID	17.8085	2.3063	7.72	<.0001	
Residual		62.2296	1.5179	41	<.0001	

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 = $1 + ([L1n - 1] * ICC)$ $L1n = \# \text{ level-1 units}$
- Effective sample size → 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 = $1 + ([L1n - 1] * ICC)$
- Effective sample size \rightarrow 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)

Example: Level-1 Students in Level-2 Schools

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

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

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

$$ICC = \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + \sigma_e^2} = \frac{17.809}{17.809 + 62.230} = .223$$

22.3% of total language variance is due to school mean differences (WC $r = .22$)

$$\text{Design Effect} = 1 + ([L1n - 1] * ICC)$$

$$\text{Effective } N = \frac{\# \text{ Total Observations}}{\text{Design Effect}}$$

- $\bar{n} = 19$ students from 191 schools, $ICC = .30$?
 - Design Effect = $1 + ([19 - 1] * .223) = 5.005$
 - **Effective $N = 3,566 / 5.005 = 712$ “effective” students**

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 (\rightarrow *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 SEs for person predictor fixed slopes \rightarrow **biased SEs**
 - Potential for **contextual effects** of person predictors (stay tuned!)
 - A “clustered-sampling correction” to the SEs **will not fix this problem!**

Increases in Random Intercept Variance

- Level-2 random intercept variance $\tau_{U_0}^2$ will often increase as a consequence of reducing level-1 residual variance σ_e^2
- **Observed level-2** $\tau_{U_0}^2$ is NOT just between-cluster variance
 - Also has a small part of within-cluster variance (**level-1** σ_e^2), or:
Observed $\tau_{U_0}^2 = \text{True } \tau_{U_0}^2 + (\sigma_e^2/L1n)$
 - With increasing $L1n$ persons, bias in true $\tau_{U_0}^2$ due to level-1 σ_e^2 is minimized
 - Model estimates of “**True**” $\tau_{U_0}^2$ use $(\sigma_e^2/L1n)$ as correction factor:
True $\tau_{U_0}^2 = \text{Observed } \tau_{U_0}^2 - (\sigma_e^2/L1n)$
- e.g., **Observed level-2** $\tau_{U_0}^2 = 4.65$, **level-1** $\sigma_e^2 = 7.06$, $L1n = 4$
 - **True** $\tau_{U_0}^2 = 4.65 - (7.06/4) = 2.88$ in empty means model
 - Add L1 within-cluster predictor → reduce σ_e^2 from **7.06** to **2.17**
 - But now **True** $\tau_{U_0}^2 = 4.65 - (2.17/4) = 4.10$ → more dependency!

ICC2: Reliability of the Cluster Mean

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

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

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

$$ICC1 = \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + \sigma_e^2} = \frac{17.809}{17.809 + 62.230} = .223$$

22.3% of the total language variance is due to school mean differences (WC $r = .22$)

ICC2 = (Spearman-Brown) Reliability of the Cluster Mean

$$ICC2 = \frac{\tau_{U_0}^2}{\tau_{U_0}^2 + (\sigma_e^2/L1n)} = \frac{17.809}{17.809 + (62.230/19)} = .845$$

The reliability of the school language mean is approximately .845 (*different ways of addressing differences in cluster size across clusters will yield ~different answers*)

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

Random Intercept Confidence Interval

- Another way to convey magnitude of random intercept variance in practical terms is a **random intercept confidence interval** (CI)
- HOWEVER—this is NOT the typical CI you are used to, which conveys an estimate's uncertainty across repeated samples
- Instead, a random intercept CI describes the **predicted range of 95% of the cluster means** (assuming normality)

Typical 95% CI for fixed intercept:

$$CI = \mathbf{Est} \pm z_{crit} * SE$$

$$CI = \mathbf{41.08} \pm 1.96 * 0.34$$

$$CI = 40.42 \text{ to } 41.74$$

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

$$CI = \mathbf{Est} \pm z_{crit} \sqrt{Var(U_{0c})}$$

$$CI = \mathbf{41.08} \pm 1.96 \sqrt{17.81}$$

$$CI = 32.81 \text{ to } 49.35$$

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	DF	Pr > ChiSq
UN(1,1)	schoolID	17.8085	2.3063	72	.0001
Residual		62.2296	1.5179	41	<.0001

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)

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 $-2 \times \log$ 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**”:
 - **$-2\Delta LL$ 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 **$-2\Delta LL$** : if given $-2LL$, use $-2\Delta LL = (-2LL_{\text{fewer}}) - (-2LL_{\text{more}})$
if given LL , use $-2\Delta LL = -2 * (LL_{\text{fewer}} - LL_{\text{more}})$
 2. Calculate **ΔDF** = (# Params_{more}) – (# Params_{fewer})
 3. **Compare $-2\Delta LL$ 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 $-2\Delta LL$ test means:
 - If you **ADD** parameters, then your model can get **better** (if $-2\Delta LL$ test is significant) or **not better** (not significant)
 - If you **REMOVE** parameters, then your model can get **worse** (if $-2\Delta LL$ 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

LRT ($-2\Delta LL$) Example: Single-Level vs. Two-Level Model for the Variance

Without random intercept U_{0c} :

Fit Statistics	
-2 Res Log Likelihood	25642.7
AIC (Smaller is Better)	25644.7
AICC (Smaller is Better)	25644.7
BIC (Smaller is Better)	25650.9

With random intercept U_{0c} :

Fit Statistics	
-2 Res Log Likelihood	25191.0
AIC (Smaller is Better)	25195.0
AICC (Smaller is Better)	25195.0
BIC (Smaller is Better)	25201.5

$$25,642.7 - 25,191.0 = 451.75$$

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

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
1	451.75	<.0001

Conclusion? Two-level fits significantly better (or single-level model fits significantly worse), which means $ICC = .223$ is significantly > 0

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 <i>default</i> (-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)
<u>Nested?</u> YES, can do significance tests via...	Fixed effect p -values from ML or REML -- OR -- ML $-2\Delta LL$ only (NO REML $-2\Delta LL$)	NO p -values REML $-2\Delta LL$ (ML $-2\Delta LL$ is ok if big N)	ML $-2\Delta LL$ only (NO REML $-2\Delta LL$)
<u>Non-Nested?</u> 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

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)

$\sigma_e^2 \rightarrow$ All possible L1 residual variance for within-school differences across students

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

- **“Unconditional” Level-2** (before cluster predictors):

➤ $\beta_{0c} = \gamma_{00} + U_{0c}$

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

- **“Conditional” Level-2** (after cluster predictors):

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

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

- First subscript = which beta in level-1 model
Second subscript = order of predictor in level-2 model

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).
```

Electronic materials for this example from my 2023 APA training sessions are [here](#)

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:

After 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

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	Chi-Square	Pr > ChiSq	Z
UN(1,1)	schoolID	17.8085	2.3063	57.72	<.0001	1
Residual		62.2296	1.5179	41.1	<.0001	

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Z
UN(1,1)	schoolID	17.1640	2.2341	56.8	<.0001	1
Residual		62.2013	1.5165	41.1	<.0001	

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:

- $\gamma_{00} = 41.55$ = fixed **intercept**: expected language for students in a school with homework=2 (~mean) and mixgrd=0 (=not mixed)
- $\gamma_{01} = 0.51$ = 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 = deviation of the original from predicted school mean language for school c (with variance $\tau_{U_0}^2 = 17.16$), where "original" is from the empty means, random intercept model
- e_{pc} = level-1 residual = deviation of the observed outcome for student p from their outcome predicted by β_{0c} (with variance $\sigma_e^2 = 62.20$)

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	Value	Pr > Z	
Residual	77.2515	1.8303	42	<.0001	

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	Value	Pr > Z
UN(1,1)	schoolID	17.1640	2.2341	68	<.0001
Residual		62.2013	1.5165	41	<.0001

Significance of Fixed Effects in MLM

	Denominator DF is infinite (Proper Wald test)	Denominator DF is estimated instead ("Modified" Wald test)
Numerator DF = 1 (<i>test one fixed effect</i>) is Univariate Wald Test	use z distribution (Mplus, STATA default)	use t distribution (SAS, SPSS, STATA with dfmethod option)
Numerator DF > 1 (<i>test 2+ fixed effects</i>) is Multivariate Wald Test	use χ^2 distribution (Mplus, STATA default)	use 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 L2n clusters and level-1 for L1n persons (like in RM ANOVA)
 - **Level-2 DDF** = $L2n - \text{\#level-2 fixed effects}$
 - **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 $L2n$ 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

Effect Size for Level-2 Cluster Predictors

- Direct: convert t -statistic for fixed effect into d or partial r

$$\triangleright d = \frac{2t}{\sqrt{DF_{den}}}, \quad r = \frac{t}{\sqrt{t^2 + DF_{den}}}$$

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*

- $\text{Pseudo-}R^2 = \frac{\text{variance}_{\text{fewer}} - \text{variance}_{\text{more}}}{\text{variance}_{\text{fewer}}}$

"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 $\sigma_e^2 = 62.230$ and Level-2 $\tau_{U_0}^2 = 17.809$, so **ICC = .223**
 - Conditional: Level-1 $\sigma_e^2 = 62.201$ (\approx because no person predictors yet), Level-2 $\tau_{U_0}^2 = 17.164$ after controlling for mixed grade and HW
- **Variance explained** by the two level-2 cluster fixed slopes:
 - **Pseudo- $R_{U_0}^2 = \frac{17.809 - 17.164}{17.809} = .036 \rightarrow 3.6\%$** of the level-2 random intercept variance (due to school mean differences) in language
 - **Total- $R^2 = \text{correlation}(\hat{y}_{pc}, y_{pc})^2 = .079^2 = .006 \rightarrow 0.6\%$** 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- $R^2 = (\text{Pseudo-}R_{U_0}^2 * \text{ICC}) + (\text{Pseudo-}R_e^2 * [1 - \text{ICC}]) = .008$ (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!

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 ($\tau_{U_0}^2$)
 - Level-1 within-cluster person differences → residual (σ_e^2)
 - Dependency effect size via Intraclass Correlation: $\text{ICC} = \tau_{U_0}^2 / (\tau_{U_0}^2 + \sigma_e^2)$
 - 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 $\tau_{U_0}^2$ 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- $R_{U_0}^2$**
 - Reduction in **total variance** can be quantified by **total- R^2** ($\approx \text{pseudo-}R_{U_0}^2 * \text{ICC}$)