

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
 - Level-2 predictors
 - Level-1 predictors
 - Random slopes and cross-level interactions

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)

Labels for Organizing Models

- Outcome type: General (normal) vs. Generalized (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
- **Generalized 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)
- **Generalized Linear Mixed Models**: **any conditional outcome distribution**, **fixed and random effects** through **link functions** (**multiple dimensions**)
 - Same concepts as for generalized or mixed separately, but with more complexity in estimation
- “**Linear**” → fixed effects predict the *link-transformed conditional mean* of outcome in a **linear combination**: $(\text{effect} \times \text{predictor}) + (\text{effect} \times \text{predictor}) \dots$

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** between-cluster and within-cluster 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
- **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)

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

Our focus

- Model for the Means (→ Predicted Values):

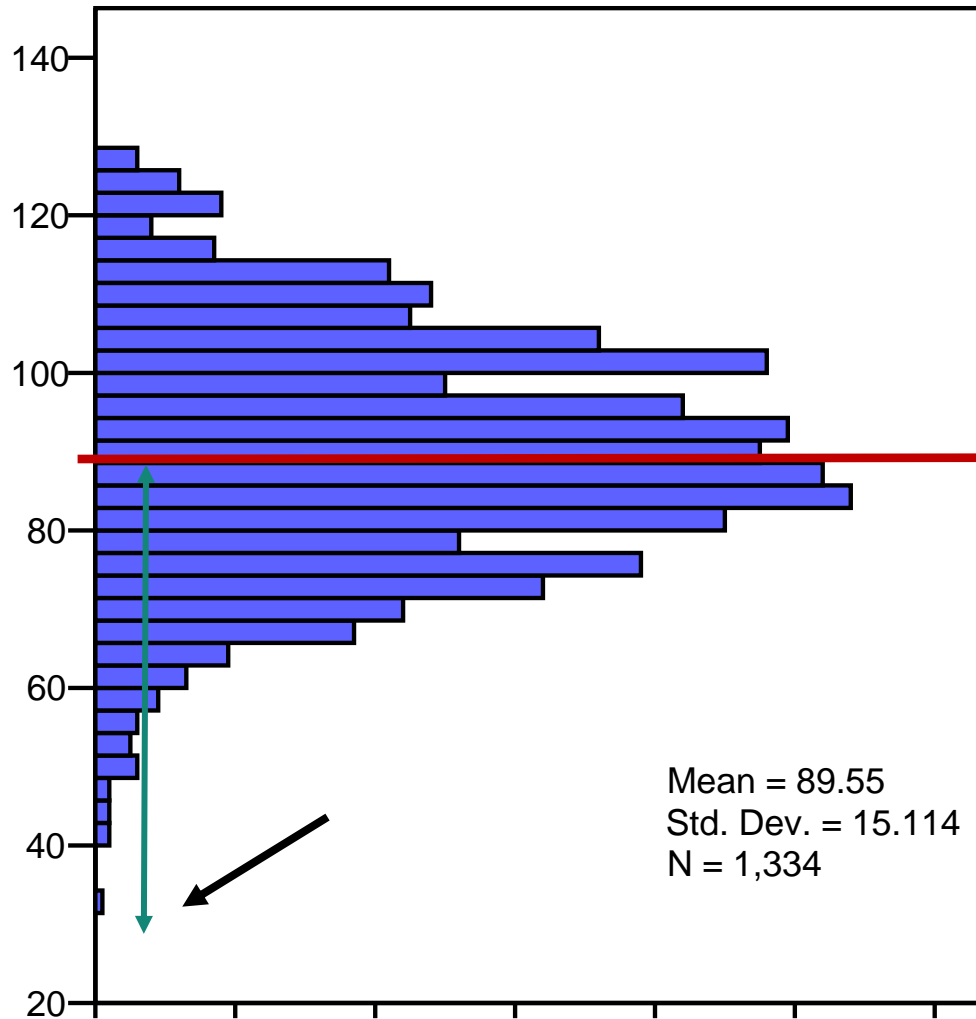
- Each person's expected (predicted) outcome is a weighted linear function of his/her values on $x1_p$ and $x2_p$ (and any other predictors); each variable is measured once per person (given by the p subscript)
- **Estimated β constants** are called **fixed effects** (intercept or slopes)

- Model for the Variance (→ “Piles” of Variance):

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

\hat{y}_p = “y-hat” model-predicted outcome

Model for the Means

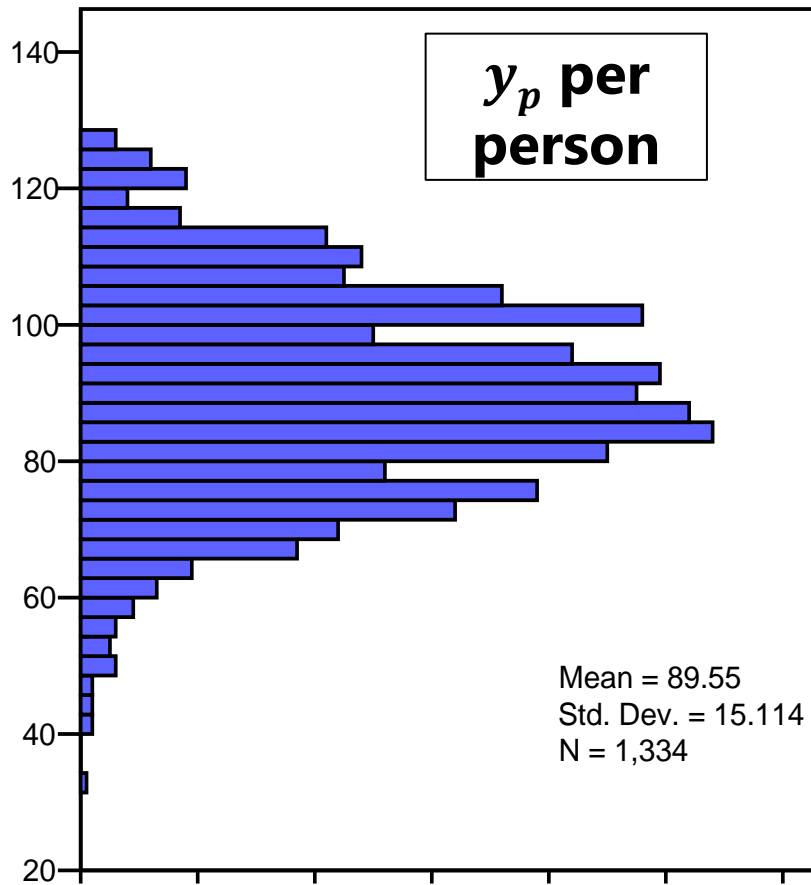
y_p residual variance:

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

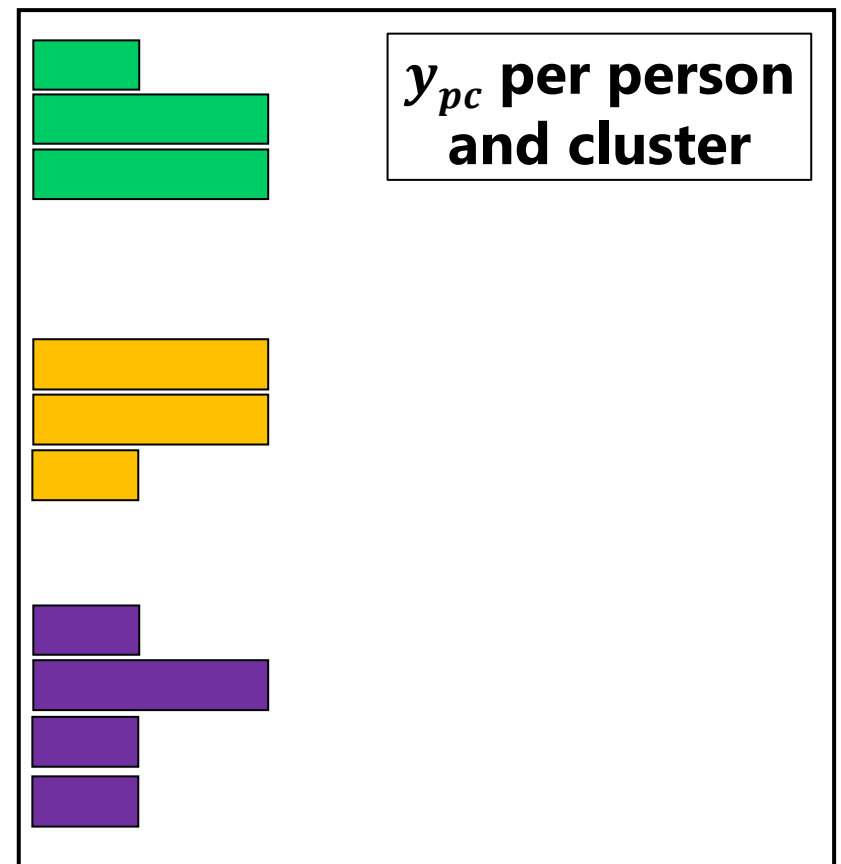
Without predictors, $\sigma_e^2 = \sigma_y^2$

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:

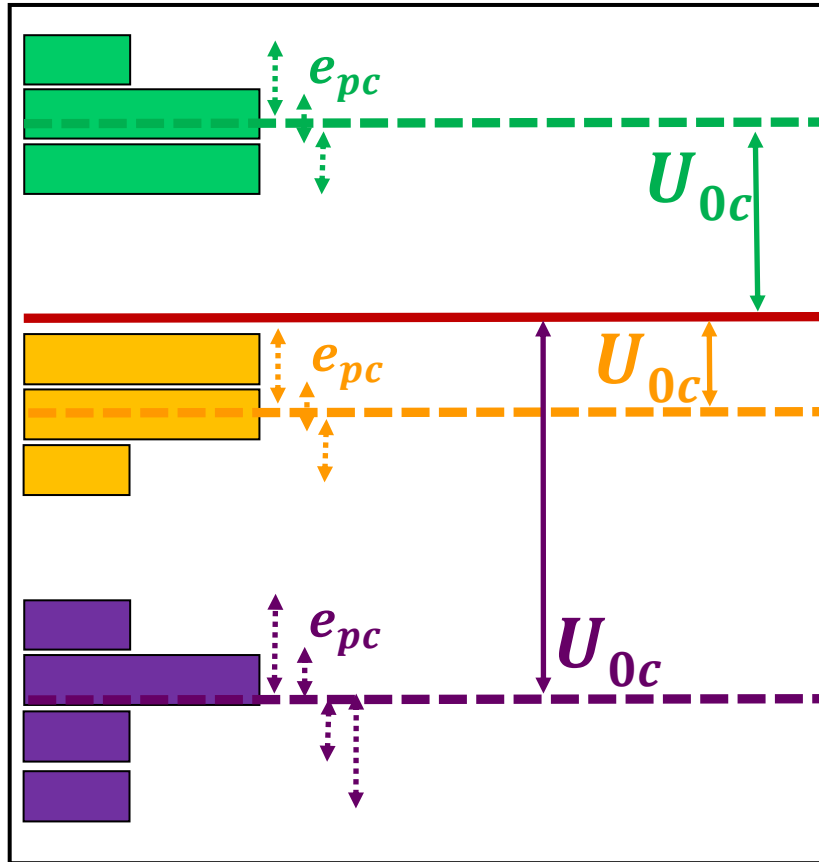
= Cluster Mean

→ **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:

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

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

L2 Random Intercept
= cluster-specific
deviation from
predicted intercept

3 total parameters:

Model for the Means (1):

- Fixed Intercept γ_{00}

Model for the Variance (2):

- Level-1 **WC** Variance of $e_{ti} \rightarrow \sigma_e^2$
- Level-2 **BC** Variance of $U_{0i} \rightarrow \tau_{U_0}^2$

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

Composite equation:

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

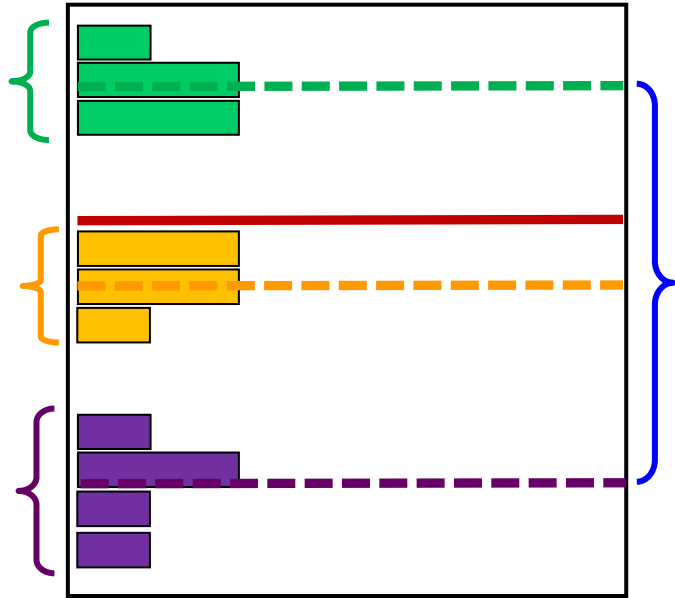
Intraclass Correlation (ICC)

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

$\tau_{U_0}^2 \rightarrow$ 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**
 - 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!**

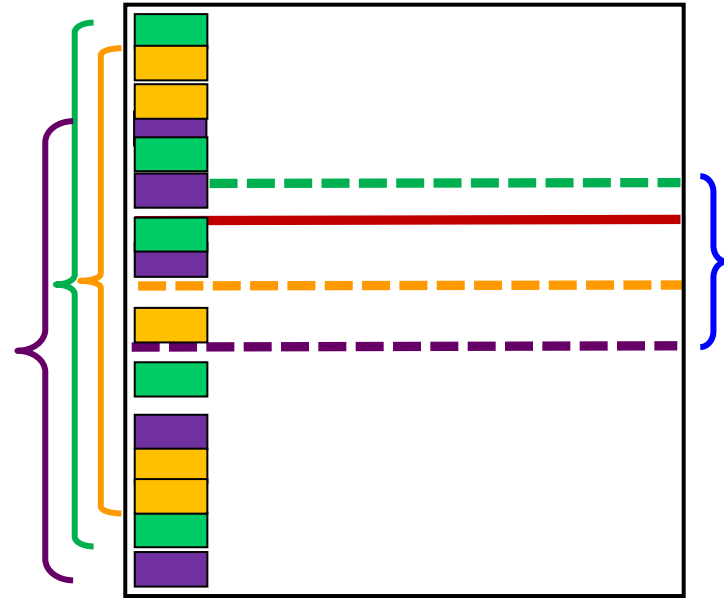


$$\text{ICC} = \text{BTW} / \text{BTW} + \text{within}$$

→ Large ICC

→ Large correlation within clusters

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



$$\text{ICC} = \text{btw} / \text{btw} + \text{WITHIN}$$

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

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 person predictor fixed slope SEs \rightarrow **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 $\#clusters < 10$ ish 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 $\#clusters > 10$ ish or you want to **predict** cluster differences
- Our examples will take this approach

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

$\sigma_e^2 \rightarrow$ All possible L1 residual variance for within-school differences across students
- **"Unconditional" Level-2** (before cluster predictors):
 - $\beta_{0c} = \gamma_{00} + U_{0c}$

$\tau_{U_0}^2 \rightarrow$ **All possible L2 random intercept variance** due to school mean differences
- **"Conditional" Level-2** (after cluster predictors):
 - $\beta_{0c} = \gamma_{00} + \gamma_{01}(L2x1_c) + \gamma_{02}(L2x2_c) + U_{0c}$

$\tau_{U_0}^2 \rightarrow$ L2 random intercept variance **leftover** now
 - First subscript = which beta in level-1 model
Second subscript = order of predictor in level-2 model

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

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 ($\tau_{U_0}^2$)
 - 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** ($\approx \text{pseudo-}R_{U_0}^2 * \text{ICC}$)

Level-1 Predictors: What **Not** to Do!

- Level-2 predictors ($L2x_c$ below) are **cluster** characteristics
- Level-1 predictors ($L1x_{pc}$ below) are **person** characteristics
 - *What if we added a L1 predictor directly (as we did before at L2)?*

Level-1: $y_{pc} = \beta_{0c} + \beta_{1c}(L1x_{pc}) + e_{pc}$

Level-2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(L2x_c) + U_{0c}$
 $\beta_{1c} = \gamma_{10}$

γ_{00}	= fixed intercept (at pred=0)
γ_{01}	= fixed slope of $L2x_c$
γ_{10}	= fixed slope of $L1x_{pc}$
U_{0c}	= level-2 random intercept
e_{pc}	= level-1 residual

- First subscript = which beta in level-1 model
Second subscript = order of predictor in level-2 model
- All good, right? Many researchers mistakenly think so, but this model is **VERY LIKELY to be mis-specified...**
 - ... For the **exact same reasons** we need MLM in the first place!

Level-1 (Person-Level) Predictors

- Modeling level-1 predictors is complicated (and often done incorrectly) because **each level-1 predictor is usually really 2 predictor variables** (each with their own slope), **not 1**
- Textbook example: Student Socioeconomic Status (SES)
 - Some **kids** have higher SES than others in their school:
 - **L1 WC variation in SES** (*represented directly as deviation from school mean*)
 - Some **schools** have more high-SES students than other schools:
 - **L2 BC variation in SES** (*represented as school mean or via external info*)
- Can quantify each source of variance with an empty model ICC
 - $ICC = (L2 \text{ between variance}) / (L2 \text{ between variance} + L1 \text{ within variance})$
 - **ICC < 1?** L1 predictor has **WC** variation (so it *could* have a **L1 WC** slope)
 - **ICC > 0?** L1 predictor has **BC** variation (so it *could* have a **L2 BC** slope)

Between- vs. Within-Cluster Effects

- Between- and within-cluster slopes in SAME direction
 - SES → Achievement in students
 - **WC: Kids with more money than other kids in their school may have greater achievement than other kids in their school (regardless of school mean SES)**
 - **BC: Schools with more money than other schools may have greater mean achievement than schools with less money**
- Between- and within-cluster slopes in OPPOSITE directions
 - Body mass → life expectancy in animals ([Curran and Bauer, 2011](#))
 - **WC: Within a species, relatively bigger animals have shorter life expectancy (e.g., over-weight ducks die sooner than healthy-weight ducks)**
 - **BC: Larger species tend to have longer life expectancies than smaller species (e.g., whales live longer than cows, cows live longer than ducks)**
- L1 within-cluster and L2 between-cluster slopes usually differ
 - Why? Because variables have different **meanings** at each level!
 - Why? Because variables have different **scales** at each level!

What **Not** to Do: Smushed Effects!

Level-1: $y_{pc} = \beta_{0c} + \beta_{1c}(L1x_{pc}) + e_{pc}$

Level-2: $\beta_{0c} = \gamma_{00} + U_{0c}$
 $\beta_{1c} = \gamma_{10}$

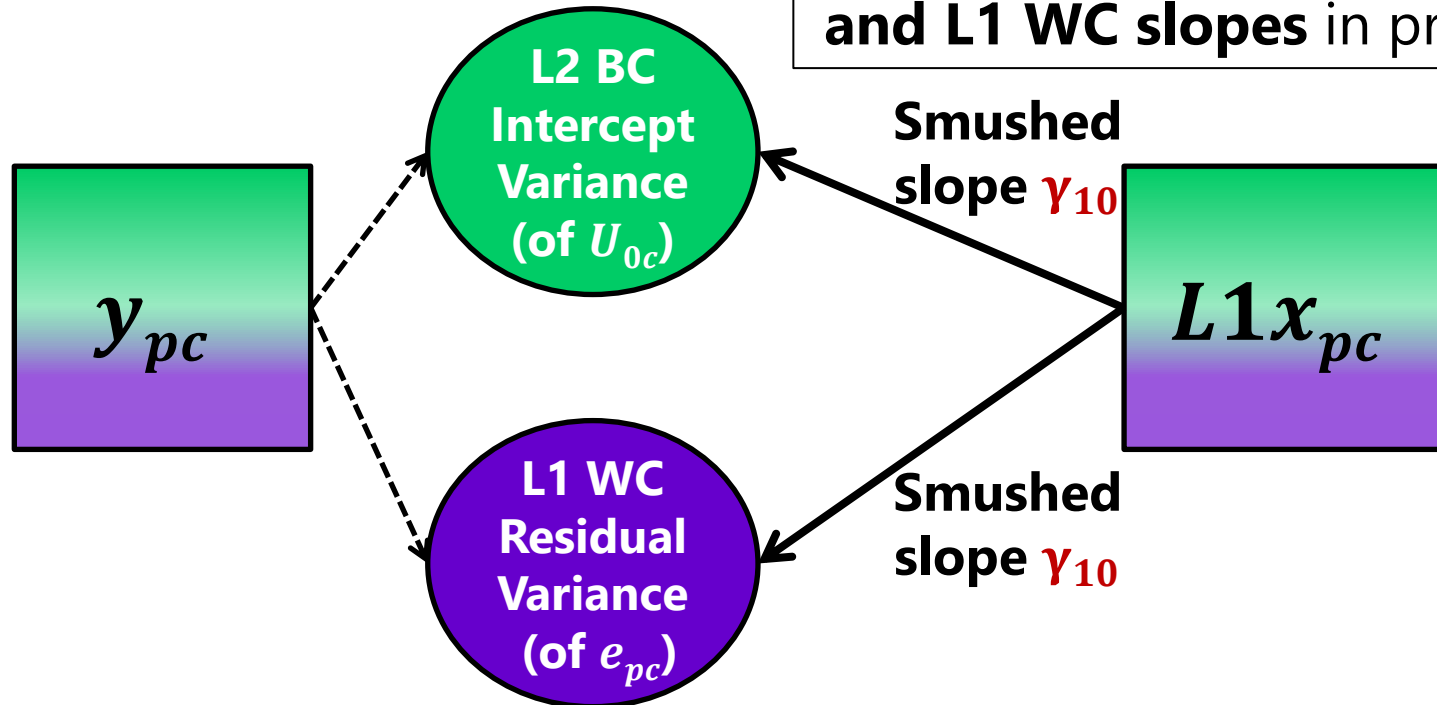
γ_{10} = **smushed effect** (see also *conflated*, *convergence*, or *composite* effect) that assumes equal within- and between-cluster slopes

- If level-1 predictor has both level-2 between and level-1 within variation, then its **one fixed slope has to do the work of two predictors!**
- A **smushed effect** is a **weighted combination of the L1 within and L2 between slopes**, usually **closer to the L1 within slope** (due to larger $L1n$), and thus the L2 between model will be more affected by smushing
- Btw, **smushing** is seen in econometrics (aka, “**endogeneity**” problem) in the context of when to model cluster dependency using fixed effects (i.e., turn cluster ID into a categorical predictor) instead of a random intercept
 - A **smushed effect creates correlation** between the L1 predictor and the L2 random intercept (because the **predictor’s L2 effect is modeled wrong**)
 - Smushing is solved when using **fixed effects for cluster ID**, such that the L2 effect of the L1 predictor is then controlled for in “common” variance
 - But we can still avoid smushed effects when using a cluster random intercept.... Next are the 3 main ways to do so!

Univariate MLM: Adding a Level-1 Predictor Without Addressing Level-2 Part = Smushing

BC and WC variance in the **observed level-1** y_{pc} **outcome** is partitioned by the **model** into estimated **variance components**

Observed level-1 $L1x_{pc}$ predictor still has both BC and WC variance. AND given that $L1x_{pc}$ has only **one fixed slope**, it captures a smushed effect that presumes **equal L2 BC and L1 WC slopes** in predicting y_{pc} !



3 Kinds of Fixed Slopes for L1 Predictors

- **Is there a Level-1 Within-Cluster (WC) slope?**
 - If you have a higher $L1x_{pc}$ predictor value *than others in your cluster*, do you also have a higher (or lower) y_{pc} outcome value *than others in your cluster*?
 - If so, the **level-1 within-cluster part of the L1 predictor** will reduce the level-1 residual variance (σ_e^2) of the y_{pc} outcome
- **Is there a Level-2 Between-Cluster (BC) slope?**
 - Do clusters with higher average $L1x_{pc}$ predictor values *than other clusters* also have higher (or lower) average y_{pc} outcomes *than other clusters*?
 - If so, the **level-2 between-cluster part of the L1 predictor** will reduce level-2 random intercept variance ($\tau_{U_0}^2$) of the y_{pc} outcome
- **Is there a Level-2 Contextual slope: Do the L2 BC and L1 WC slopes differ?**
 - After controlling for the actual value of L1 predictor, is there still **an incremental contribution** from the **level-2 between-cluster part of the L1 predictor** (i.e., does a cluster's general tendency matter beyond a person's $L1x_{pc}$ value)?
 - Equivalently, the **Level-2 Contextual slope** = **L2 BC slope** – **L1 WC slope**, so the Level-2 Contextual slope directly tests **if a smushed slope is ok (pry not!)**

3 Options to Prevent Smushed Slopes

- Within Univariate MLM framework (predict only one outcome):
 1. **Cluster-mean-centering**: manually carve up L1 predictor into its level-specific parts using observed variables (1 predictor per level)
 - More generally, this is “**variable-centering**” because you are **subtracting a variable** (e.g., the cluster mean here; could use other cluster variables)
 - Will always yield **level-1 within slopes** and **level-2 between slopes**!
 2. **Grand-mean-centering**: do NOT carve up L1 predictor into its level-specific parts, but add level-2 mean to distinguish level-specific slopes
 - More generally, this is “**constant-centering**” because you are **subtracting a constant** while still keeping all levels of variance in the L1 predictor
 - **Choice of constant is irrelevant** (changes where 0 is, not what variance it has)
 - Will always yield **level-1 within slopes** and **level-2 contextual slopes**!
- Within Multivariate MLM framework (i.e., via Multilevel-SEM):
 3. **Latent-centering**: Treat the L1 predictor as another outcome
→ let the model carve it up into **level-specific latent variables**
 - Best in theory, but the type of level-2 slope (between or contextual) depends on model type, syntax type, and the estimator in Mplus! ([Hoffman, 2019](#))

Option 1. Cluster-Mean-Centering (C-MC)

- We partition the L1 predictor $L1x_{pc}$ into two variables that directly represent its **L2 between**-cluster (BC) and **L1 within**-cluster (WC) sources of variation, and **include these variables as the predictors**:
- **Level-2 Between predictor = cluster mean of $L1x_{pc}$**
 - $CMx_c = \overline{L1x_c} - C_2$
 - CMx_c is centered at constant C_2 , chosen for meaningful 0 (e.g., sample mean)
 - CMx_c is positive? Above sample mean → “more than other clusters”
 - CMx_c is negative? Below sample mean → “less than other clusters”
- **Level-1 Within predictor = deviation from cluster mean of $L1x_{pc}$**
 - $WCx_{pc} = L1x_{pc} - \overline{L1x_c}$ (*uncentered cluster mean $\overline{L1x_c}$ is used*)
 - WCx_{pc} is NOT centered at a constant – **we subtract a VARIABLE instead**
 - WCx_{pc} is positive? Above your cluster mean → “more than my cluster”
 - WCx_{pc} is negative? Below your cluster mean → “less than my cluster”

Cluster-MC L1 Predictor + Cluster Mean

→ WC and BC effects directly through separate parameters

$L1x_{pc}$ is cluster-mean-centered into WCx_{pc} , with CMx_c at L2:

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

$WCx_{pc} = L1x_{pc} - \overline{L1x_c} \rightarrow$
only has L1 within variation

Level-2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(CMx_c) + U_{0c}$
 $\beta_{1c} = \gamma_{10}$

$CMx_c = \overline{L1x_c} - C_2 \rightarrow$ only
has L2 between variation

γ_{10} = within effect
of having more
 $L1x_{pc}$ *than others*
in your cluster

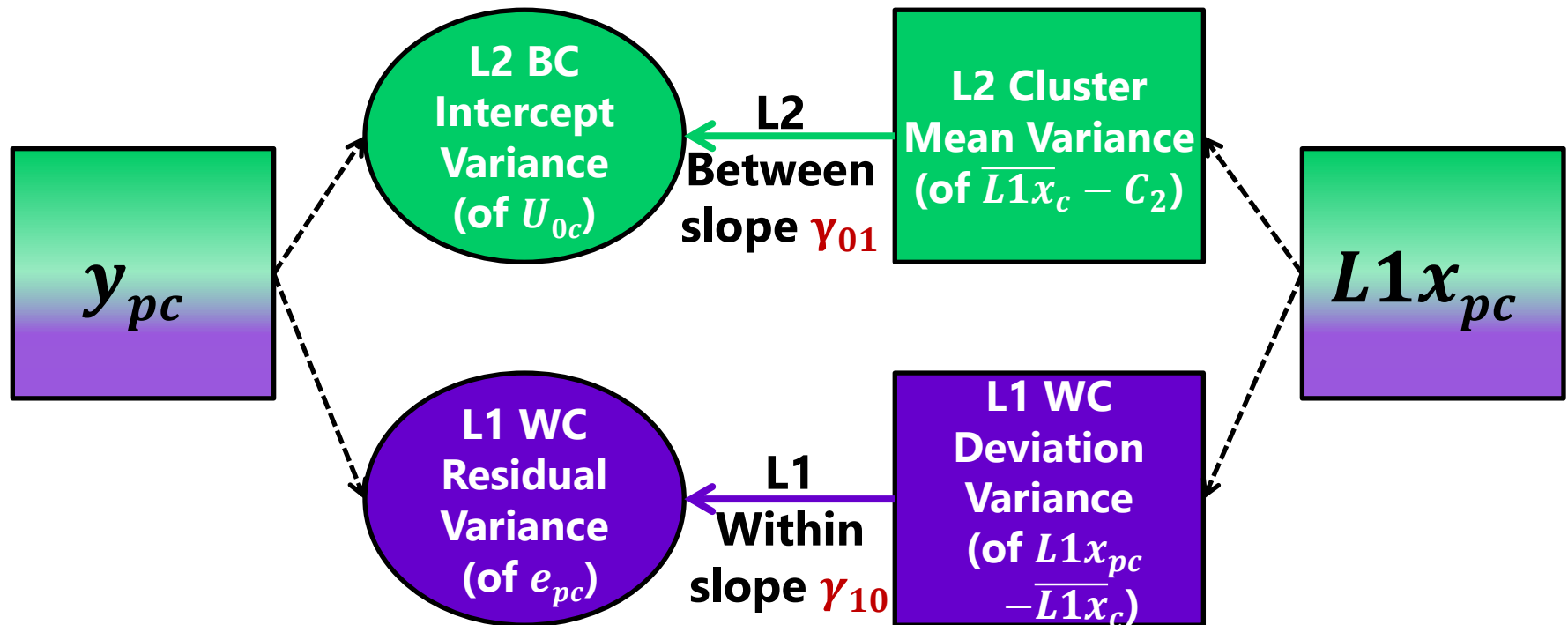
γ_{01} = between
effect of having
more $\overline{L1x_c}$ *than*
other clusters

Because WCx_{pc} and CMx_c
are uncorrelated, each gets
the total effect for its level
(L1 = within, L2 = between)

Univariate MLM: Cluster-Mean-Centering

Model-based partitioning of level-1 y_{pc} outcome into level-specific **latent variables**

Manual partitioning of level-1 $L1x_{pc}$ predictor into level-specific **observed variables**



Why not let the model make variance components for $L1x_{pc}$, too?
That is option 3, multivariate MLM (or "multilevel SEM"): stay tuned...

3 Kinds of Fixed Slopes for L1 Predictors

- **2 kinds of slopes Cluster-Mean-Centering tells us *directly*:**
- **Is there a Level-1 Within-Cluster (WC) slope?**
 - If you have higher predictor values than the rest of your cluster, do you also have higher outcomes values than the rest of your cluster, such that the within-cluster deviation of the L1 predictor accounts for L1 residual outcome variance (σ_e^2)?
 - **Given directly by fixed slope of WCx_{pc} regardless of whether CMx_c is there**
 - Note: L1 slope multiplies the **relative** value of $L1x_{pc}$, NOT the **original** $L1x_{pc}$
- **Is there a Level-2 Between-Cluster (BC) slope?**
 - Do clusters with higher predictor values than other clusters (*on average*) also have higher outcomes than other clusters (*on average*), such that the cluster mean of the L1 predictor accounts for level-2 random intercept variance ($\tau_{U_0}^2$)?
 - **Given directly by fixed slope of CMx_c regardless of whether WCx_{pc} is there**
 - Note: BC slope is NOT controlling for the original $L1x_{pc}$ for each person

3rd Kind of Slope for L1 Predictors

- What **Cluster-Mean-Centering DOES NOT** tell us *directly*:
- Is there a **Level-2 Contextual** effect: Do the **BC** and **WC** slopes differ?
 - After controlling for the original value of the L1 predictor per person, is there still **an incremental contribution from having a higher cluster mean** of the L1 predictor (i.e., does a cluster's general tendency for the predictor explain more $\tau_{U_0}^2$ above and beyond just the person-specific value of the L1 predictor)?
 - If there is no contextual effect, then the L1 predictor's **L2 BC** and **L1 WC** slopes show **convergence**, which means their effects are of equivalent magnitude
- To answer this question about the **Level-2 Contextual effect for the incremental contribution of the cluster mean**, we have two options:
 - Still use Cluster-MC, and ask for the **contextual slope = between – within** (via SAS ESTIMATE, R contest1D, SPSS TEST, STATA LINCOM, Mplus NEW...)
 - Use “**constant-centering**” for the L1 predictor: $L1x_{pc} = L1x_{pc} - C_1$
→ **centered at CONSTANT C_1 , NOT A LEVEL-2 VARIABLE**
 - Which constant only matters for the reference point; it could be the grand mean or any (even 0)

Why the Difference in the Level-2 Effect?

Remember Regular Old Regression...

- In this model: $y_p = \beta_0 + \beta_1(x1_p) + \beta_2(x2_p) + e_p$
- If $x1_p$ and $x2_p$ **ARE NOT** correlated:
 - β_1 carries **ALL the relationship** between $x1_p$ and y_p
 - β_2 carries **ALL the relationship** between $x2_p$ and y_p
- If $x1_p$ and $x2_p$ **ARE** correlated:
 - β_1 is **different than** the bivariate relationship between $x1_i$ and y_i
 - “Unique” effect of $x1_p$ *controlling for $x2_p$* (i.e., *holding $x2_p$ constant*)
 - β_2 is **different than** the bivariate relationship between $x2_i$ and y_i
 - “Unique” effect of $x2_p$ *controlling for $x1_p$* (i.e., *holding $x1_p$ constant*)
- **Hang onto that idea...**

Cluster-Mean-Centering vs. Constant-Centering for Level-1 Predictors

Level 2		Original	Cluster-MC Level 1	Grand-MC Level 1
$\overline{L1x_c}$	$\text{CM}x_c = \overline{L1x_c} - 5$	$L1x_{pc}$	$\text{WC}x_{pc} = L1x_{pc} - \overline{L1x_c}$	$L1x_{pc} = L1x_{pc} - 5$
3	-2	2	-1	-3
3	-2	4	1	-1
7	2	6	-1	1
7	2	8	1	3

Same L2 $\text{CM}x_c$ goes into the model given either way of centering the L1 predictor $L1x_{pc}$

In **variable-centering** (C-MC), the level-2 BC mean variation is gone from $\text{WC}x_{pc}$, so it is NOT CORRELATED with $\text{CM}x_c$

In **constant-centering**, the level-2 BC mean variation is still inside $L1x_{pc}$, so it IS STILL CORRELATED with $\text{CM}x_c$

So the effects of $\text{CM}x_c$ and $L1x_{pc}$ when included together under constant-centering will be different than if either predictor were included by itself...

Level-1 Predictor + Cluster Mean

→ Model tests difference of WC vs. BC effects

$L1x_{pc}$ is constant-centered, but WITH CMx_c at Level 2:

Level-1: $y_{pc} = \beta_{0c} + \beta_{1c}(L1x_{pc}) + e_{pc}$

$L1x_{pc} = L1x_{pc} - C_1 \rightarrow$
still has both L2 between
and L1 within variation

Level-2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(CMx_c) + U_{0c}$
 $\beta_{1c} = \gamma_{10}$

$CMx_c = \overline{L1x_c} - C_2 \rightarrow$ only
has L2 between variation

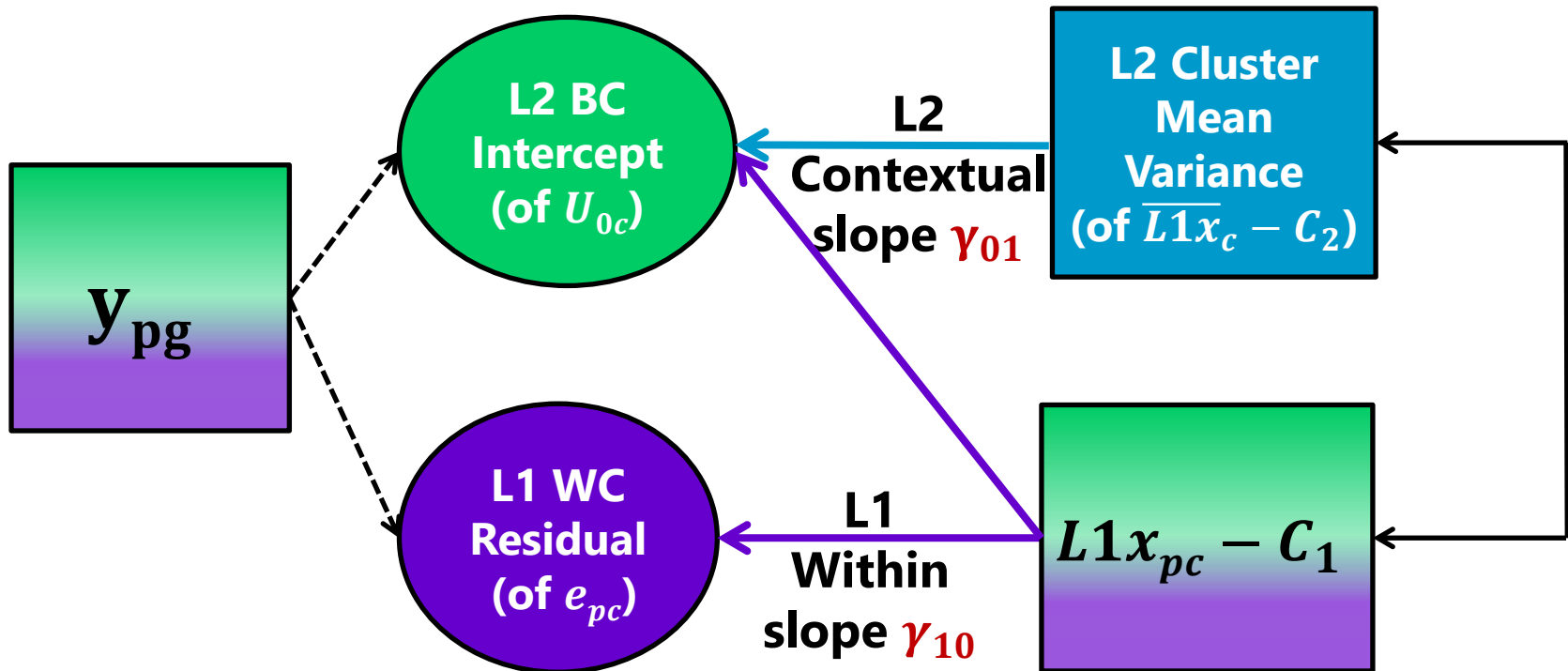
γ_{01} becomes the
within effect →
unique L1 effect
after controlling
for L2 CMx_c

γ_{01} becomes the **L2 Contextual slope** that indicates
how the L2 BC effect differs from the L1 WC effect
→ *unique* level-2 slope after controlling for $L1x_{pc}$
→ does cluster mean matter beyond person value?

Constant-Centering + Cluster Mean

Model-based partitioning of y_{pc} outcome into level-specific **latent variables**

$L1x_{pc}$ is still **NOT** partitioned, but cluster mean $\overline{L1x_c} - C_2$ is added to allow an **incremental L2 effect**



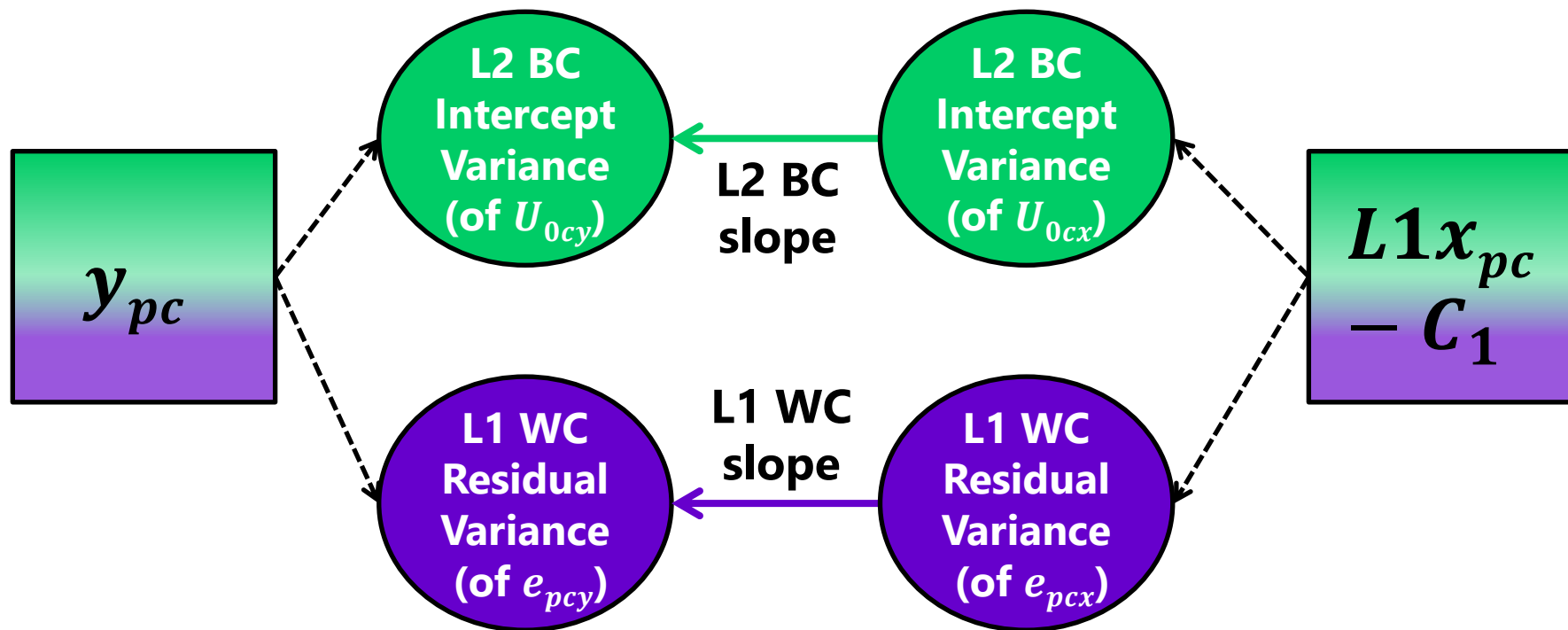
L2 BC slope = **L1 WC slope** + **Level-2 Contextual slope**

Because original $L1x_{pc}$ still has L2 BC variance, it still carries **some** of the L2 BC effect...

Option 3: Latent-Centering in Multivariate MLM

Model-based partitioning of level-1 y_{pc} outcome into level-specific **latent variables**

Model-based partitioning of level-1 $L1x_{pc}$ predictor (= outcome now) into level-specific **latent variables**



Univariate MLM software can be tricked into multivariate MLM if the relationships between X and Y at each level are phrased as covariances, but not if you want directed regressions (or moderators thereof)

I Usually Prefer Variable-Centering (using observed or latent variables)...

- ...because constant-centering is much easier to screw up! ☺
- Table 1 below from: Hoffman, L., & Walters, R. W. (2022). [Catching up on multilevel modeling](#). *Annual Review of Psychology*, 73, 629-658.

Table 1 Predictor effect type by model specification

Centering strategy for level-1 predictor (constant-centered level-2 predictor)	Fixed effect type by predictors included		
	Level-1 only	Level-2 only	Both levels
Variable-centered level-1			
Level-1 predictor: $L1x_{wb} = x_{wb} - \bar{x}_b$	Within	(= 0)	Within
Level-2 predictor: $L2x_b = \bar{x}_b - C_2$	(= 0)	Between	Between
Constant-centered level-1			
Level-1 predictor: $L1x_{wb} = x_{wb} - C_1$	Smushed	(= 0)	Within
Level-2 predictor: $L2x_{wb} = \bar{x}_b - C_2$	(= Within)	Between	Contextual

Abbreviations: w , within; b , between; C_1 , level-1 centering constant; C_2 , level-2 centering constant.
Parentheses indicate assumptions about the fixed slopes of omitted predictors.

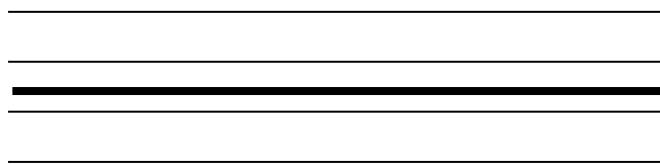
Constant-Centering for L1 predictors (+ Cluster Mean) may be preferable when:

- **You really do want level-2 contextual effects**
 - Directly model the incremental contribution of the cluster mean after controlling for a person's actual (not relative) predictor
- **For *categorical* level-1 predictors**
 - e.g., 0/1 predictors when cluster-MC → impossible values
- **When the cluster mean is not a reliable cluster-level predictor**
 - When the sample of persons within clusters is not complete enough to form a useful cluster mean, using externally-provided info may do a better job of representing the cluster (in which case cluster-MC doesn't really make sense without the cluster mean to go in with it)
- But cluster-MC or latent-centering is needed instead to prevent a L1 predictor's **random slope** from being smushed...
 - **Fixed slope** → every cluster gets the **same**
 - **Random slope** → every cluster gets their **own!**

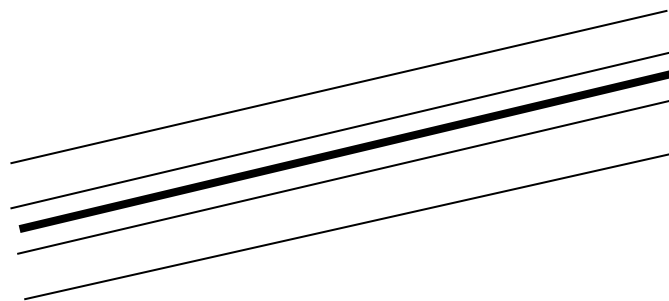
Fixed and Random Effects of L1 Predictor

(Note: The cluster intercept is random in every figure)

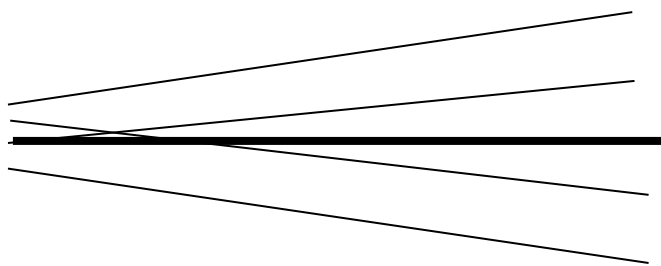
No Fixed, No Random



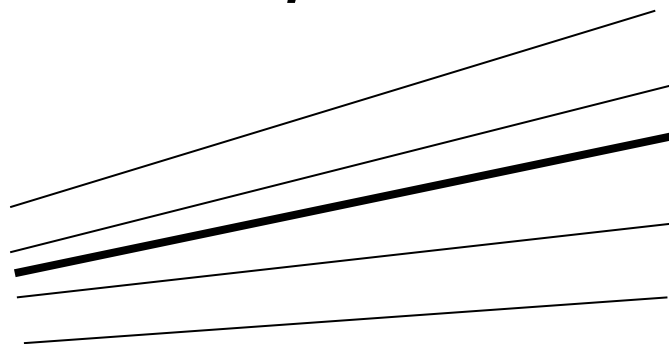
Yes Fixed, No Random



No Fixed, Yes Random



Yes Fixed, Yes Random



Cluster-MC Predictor* with Random Slope

$L1x_{pc}$ is cluster-mean-centered into WCx_{pc} , with CMx_c at L2:

Level-1: $y_{pc} = \beta_{0c} + \beta_{1c}(WCx_{pc}) + e_{pc}$ $WCx_{pc} = L1x_{pc} - \overline{L1x_c} \rightarrow$
only has L1 within variation

Level-2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(CMx_c) + U_{0c}$
 $\beta_{1c} = \gamma_{10} + U_{1c}$ $CMx_c = \overline{L1x_c} - C_2 \rightarrow$ only has L2 between variation

γ_{10} = within effect of having more $L1x_{pc}$ than others in your cluster

γ_{01} = between effect of having more $\overline{L1x_c}$ than other clusters

U_{1c} is a random slope for the WC effect of WCx_{pc}

Because WCx_{pc} and CMx_c are uncorrelated, each gets the total effect for its level (L1 = within, L2 = between)

* If a constant-centered L1 predictor were used instead, the U_{1c} random slope would also multiply its L2 between part, creating bias in the estimated random slope variance. **To avoid such a smushed random slope, we need to use either cluster-MC (in univariate MLM) or latent-centering (in multivariate MLM).**

Implications of Random Slopes

- **L2 random slopes** capture a second, distinct source of cluster **dependency**—differences in **slope of a L1 person predictor**
 - Beyond **constant** covariance for persons from same L2 cluster (as created by the **L2 random intercept**), the **L2 random slope** adds **non-constant** covariance across values of its L1 predictor (e.g., WCx_{pc})
 - Also adds **quadratic heterogeneity** of variance across L1 predictor:
$$Var(y_{pc}) = \tau_{U_0}^2 + (WCx_{pc}^2 * \tau_{U_1}^2) + (2WCx_{pc} * \tau_{U_{01}}) + \sigma_e^2$$
- **Random slopes do NOT*** explain variance (like **fixed slopes** do) because cluster slope differences are still “**error**” conceptually
 - We know **THAT** clusters need different slopes of $L1\ WCx_{pc}$ but not **WHY**
- Therefore, random slopes imply another role for level-2 cluster predictors—to explain cluster differences in slope of $L1\ WCx_{pc}$
 - To do so, we need “**cross-level interactions**” of L2 by L1 predictors!

** Hill that I will die on, but others disagree (i.e., marginal vs. conditional R^2)*

Explained Variance by Fixed Slopes

- **Fixed slopes of level-2 cluster predictors *by themselves*:**
 - L2 BC main effects or interactions reduce L2 random intercept variance
- **Fixed slopes of *cross-level interactions* (level-1 * level-2):**
 - If the **L1 person predictor also has a random slope**, its cross-level interaction will reduce its corresponding **L2 random slope variance**
 - So make sure you test the random slope before any cross-level interactions!
 - If the **L1 person predictor does NOT have a random slope**, its cross-level interaction will reduce the **L1 residual variance** instead
 - This condition creates a “systematically varying” L1 slope instead, in which the slope varies only by interacting predictors (but not randomly otherwise)
- **Fixed slopes of level-1 person predictors *without L2 variance*:**
 - L1 WC main effects or interactions reduce L1 residual variance
- **Fixed slopes of level-1 person predictors *with L2 variance*:**
 - L1 WC main effects or interactions reduce both L1 residual variance and L2 random intercept variance; need to add corresponding L2 main effects, L2 interactions, or cross-level interactions in order to prevent smushing!
 - See [Hoffman & Walters \(2022\)](#) and [Hoffman \(2019\)](#) for elaboration

Part 2: Summary

- **Level-1 predictors** are person characteristics, but they **almost always contain cluster mean differences** (level-2 variance) as well
 - **Variance** at each level → **different slope** at each level!
- **3 options** for specifying fixed slopes of a L1 predictor in order to distinguish its level-specific effects (i.e., **avoid smushed effects**):
 1. **Cluster-Mean-Centering**: Manually carve up into L2 BC (cluster mean → **L2 Between slope**) and L1 WC deviation (→ **L1 Within slope**)
 2. **Constant-Centering**: Add cluster mean to become **L2 Contextual slope**, then L1 predictor's unique effect is **L1 Within slope**
 3. **Latent-Centering**: Let multivariate MLM estimate L2 and L1 variance components, same as for the outcome → analogous to Cluster-MC
- A **level-2 random slope** variance allows cluster differences in the effect of a L1 person predictor (using only options 1 or 3)
 - Implies **heterogeneity** of variance and covariance across L1 predictor
 - Implies **another way clusters differ** (to be explained by **cross-level interactions** between that L1 predictor and L2 predictors)

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)
- Comparison of between vs contextual effects
- More depictions of level-2 between, level-2 contextual, and level-1 within slopes
 - Example variables: How often students are read to by their parents predict student math outcomes

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 (<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 – $\text{\#level-1 fixed effects}$
 - 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

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 $-2 \times \log \text{likelihood}$ ($-2LL$) 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** = $(\# \text{Parms}_{\text{more}}) - (\# \text{Parms}_{\text{fewer}})$
 3. **Compare $-2\Delta LL$ to χ^2 distribution with numerator $DF = \Delta DF$**
 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 = $-2LL + 2 * (\text{\#parameters})$
 - **BIC** = Bayesian IC = $-2LL + \log(N) * (\text{\#parameters})$
 - 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 <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)

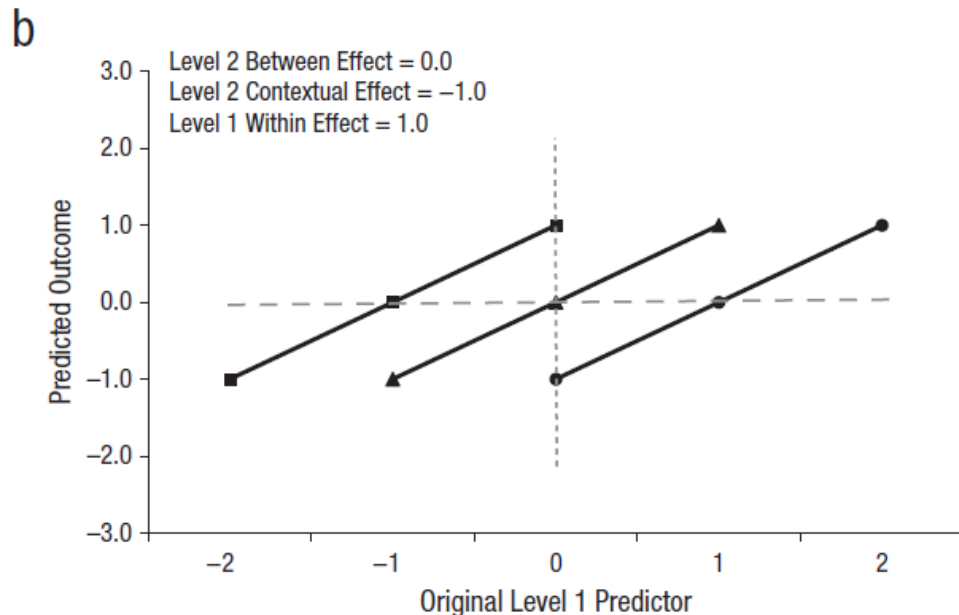
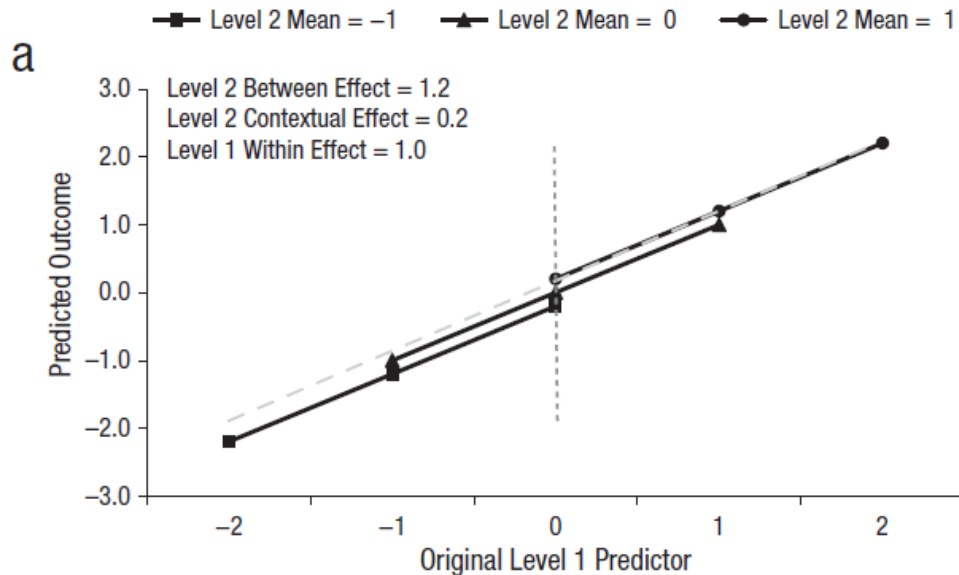
Nested = one model is a direct subset of the other

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

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 $L1n$ perons, 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!

Bonus: Between vs. Contextual Effects

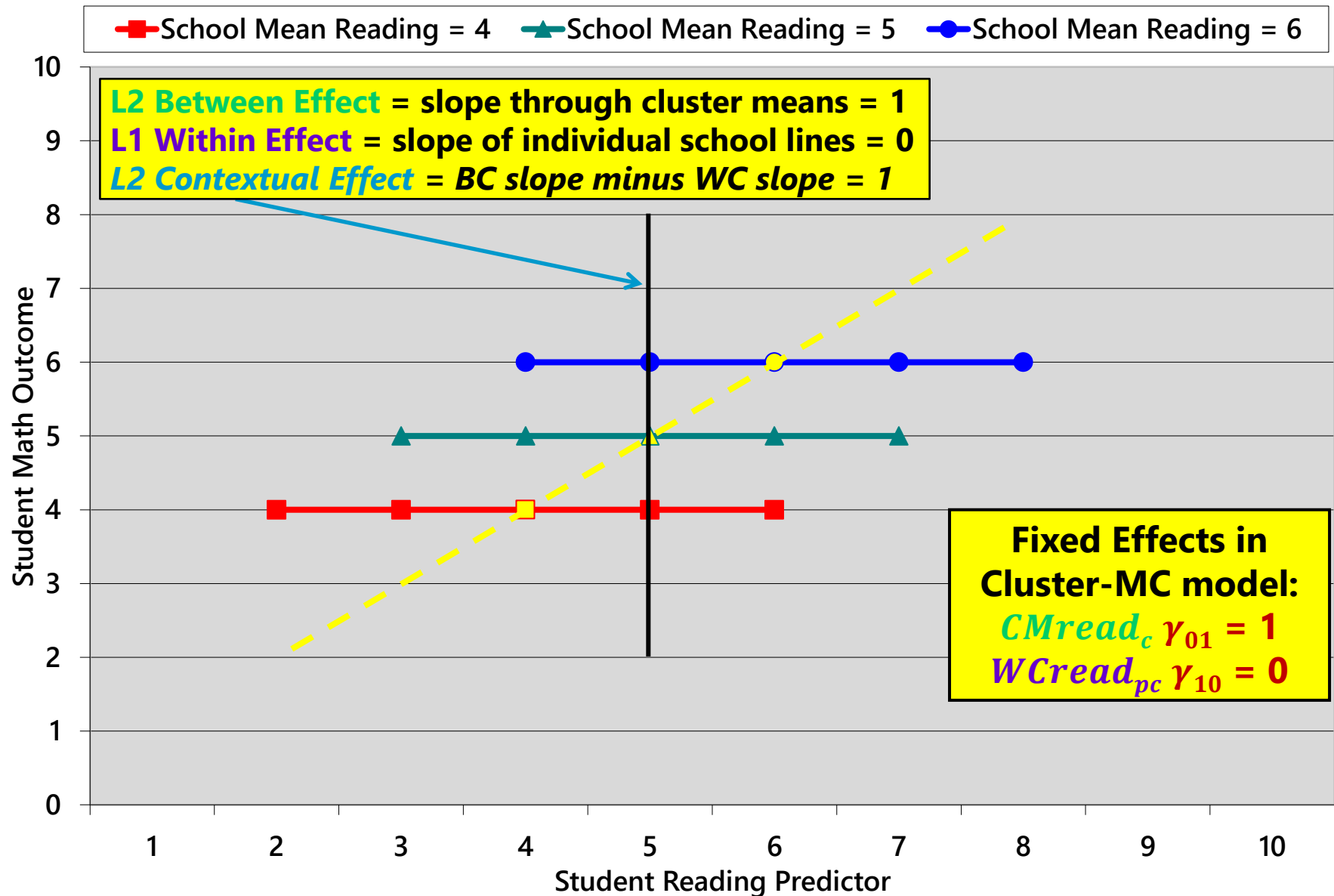


- Image from [Hoffman \(2019\)](#), example using student SES

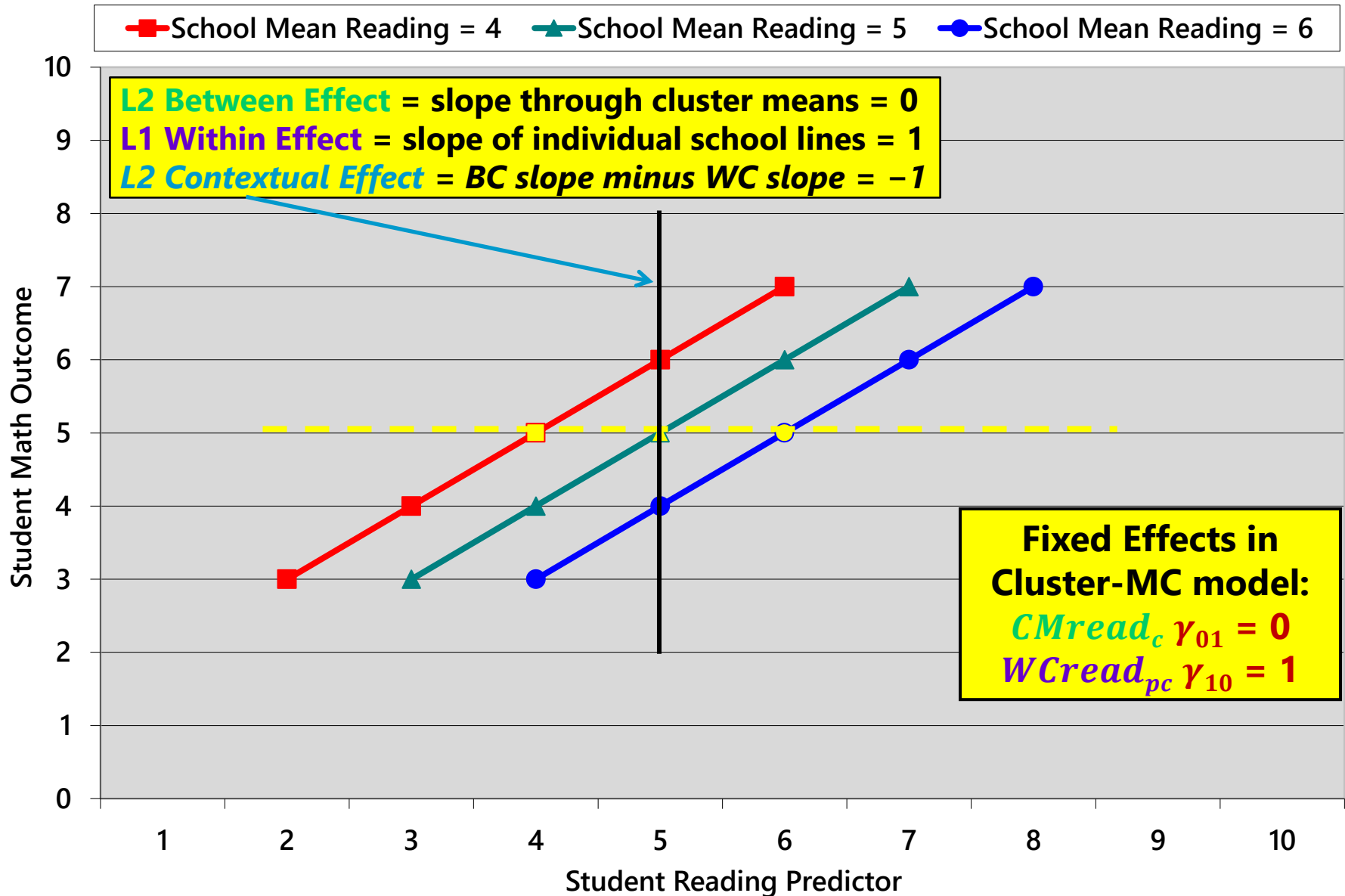
- *Top*: Contextual effect is minimal—there is no added benefit to going to a high-SES school when comparing across schools *at same level of student SES*

- *Bottom*: Contextual effect is negative—at the same student SES level, relatively high students from low-SES schools do better than relatively low students from high-SES schools

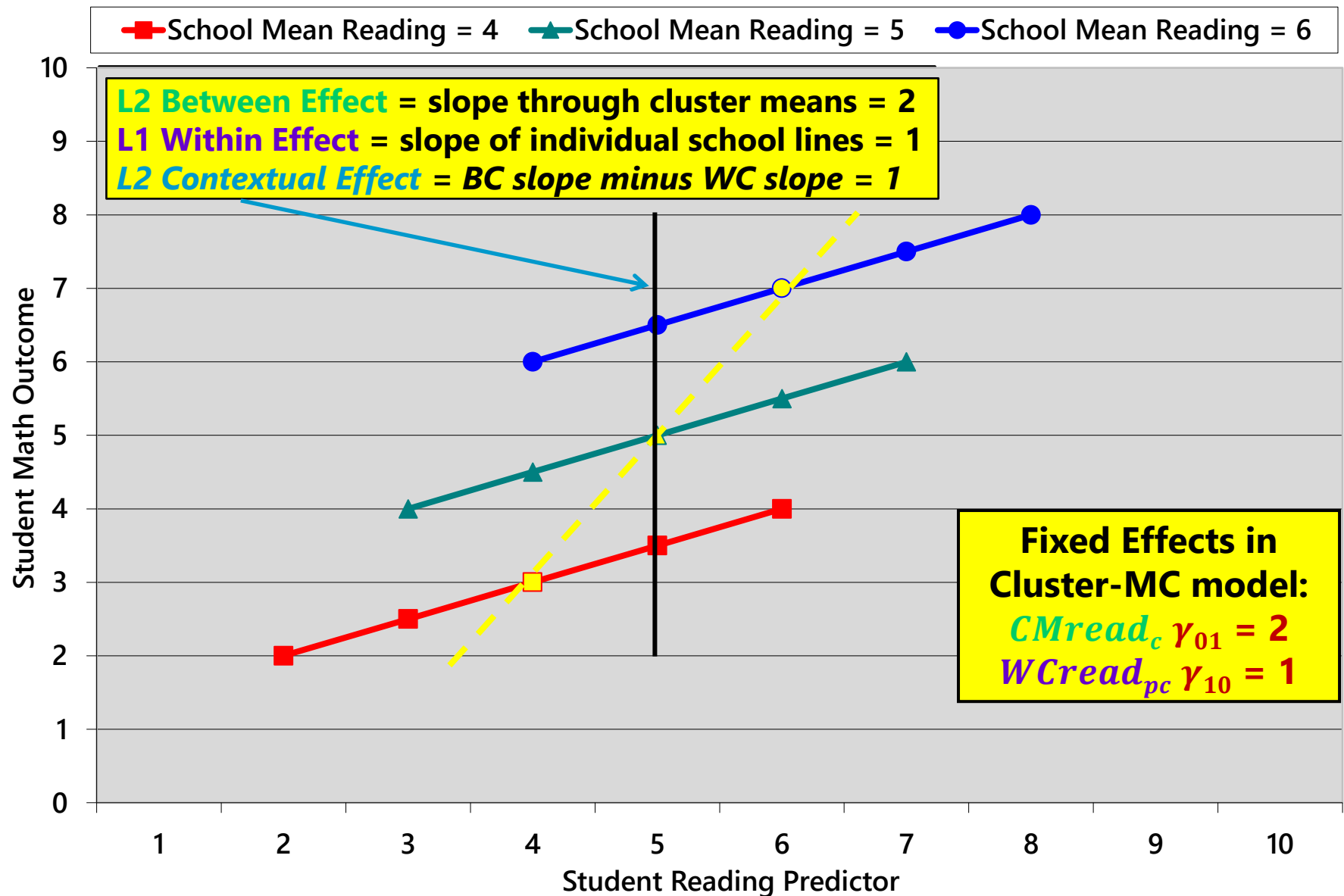
ALL Between Effect, NO Within Effect



NO Between Effect, ALL Within Effect



Between Effect > Within Effect



Between, Within, and Contextual Effects

