Random Slopes and Cross-Level Interactions (in General Multilevel Models for Two-Level Nested Data)

- Topics:
  - Random slopes of level-1 predictors
  - Fun with cross-level interactions
    - Using cluster-mean-centered level-1 predictors
    - Using constant-centered level-1 predictors
    - Hybrid models to avoid smushed random effects
    - Level-2 interactions to avoid smushed cross-level interactions
    - Systematically varying effects—a compromise between fixed and random
  - How random slopes create heterogeneity of variance and covariance
  - An overview of model estimation and its practical consequences

MLMs for Clustered Data: A Review

- Multilevel models (MLMs) are used to quantify and predict how much of an outcome's leftover variation is due to each dimension of sampling ***(as people in clusters here)***
- Empty means, two-level model for level-1 person $\boldsymbol{p}$ in level-2 cluster $\boldsymbol{c}$:

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

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

$\boldsymbol{\gamma_{00}}$ = fixed intercept (mean of cluster means)
$\boldsymbol{U_{0c}}$ = level-2 random intercept (with variance across clusters as $\boldsymbol{\tau^2_{U_0}}$)
$\boldsymbol{e_{pc}}$ = level-1 residual (with variance across people as $\boldsymbol{\sigma^2_e}$)

- **Total** outcome variation is partitioned into **two uncorrelated sources**:
  - **Level-2 between**-cluster (BC) mean differences → random intercept $\boldsymbol{\tau^2_{U_0}}$
  - **Level-1 within**-cluster (WC) cluster differences → residual $\boldsymbol{\sigma^2_e}$
  - Dependency effect size via Intraclass Correlation: $\boldsymbol{ICC_1 = \tau^2_{U_0} / (\tau^2_{U_0} + \sigma^2_e)}$
    - ICC_1 = proportion of total variance due to **cluster *mean* differences (another kind soon!)**
    - ICC_1 = average correlation of persons from same cluster
- Fixed slopes of cluster-level-2 predictors explain cluster mean differences, thereby reducing the cluster-level-2 random intercept variance $\boldsymbol{\tau^2_{U_0}}$

MLMs for Clustered Data: Review

- **Level-1 (L1) predictors** are person characteristics, but they **almost always contain cluster mean differences** as level-2 (L2) variance as well
  - **Variance** at each level → **different slope** at each level! (Yes, we must care!)

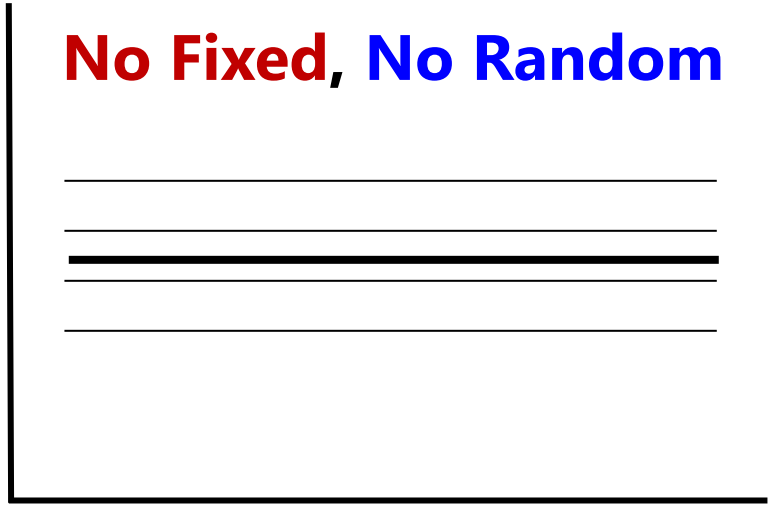
- **3 options** for specifying fixed slopes **of a L1 predictor** in order to distinguish its within and between level-specific effects (i.e., **avoid smushed L1 effects**):
  1. (univariate) **Cluster-Mean-Centering (or variable-centering)**: carve up L1 predictor into L2 between (cluster mean → **L2 Between slope**) and L1 within deviation (→ **L1 Within slope**)
  2. (univariate) **Grand-Mean-Centering (or constant-centering)**: Add cluster mean to become **L2 Contextual slope**, then L1 predictor's unique (unsmsuhed) effect is **L1 Within slope**
  3. (multivariate) **Latent-Centering**: Let model estimate the predictor's (and outcome's) L2 and L1 variance components → latent version of cluster-MC using observed variables

- But cluster-mean-centering or latent-centering is needed instead to prevent a L1 predictor's **random slope** from being smushed… this is a new concept:
  - **Fixed slope** → every cluster gets the **same slope** of the L1 predictor
  - **Random slope** → every cluster gets their **own slope** of the L1 predictor
    - To be explained by "**cross-level" interactions** of a L2 predictor with that L1 predictor!

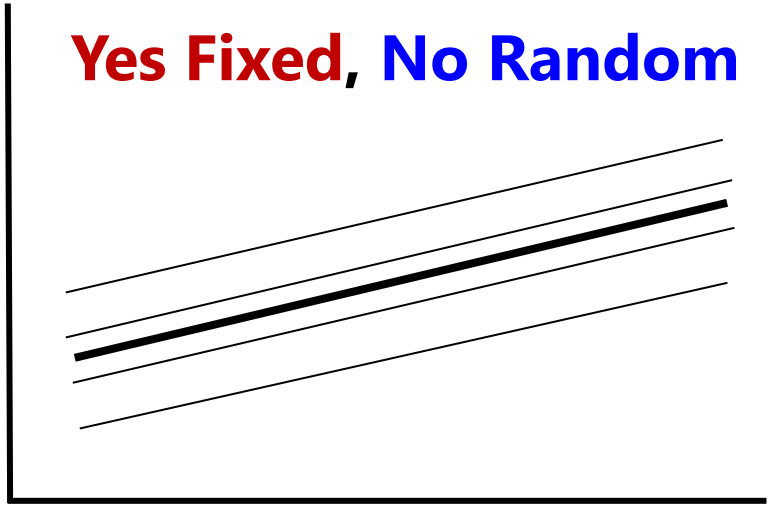
Fixed and Random Slopes of a Level-1 Predictor

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

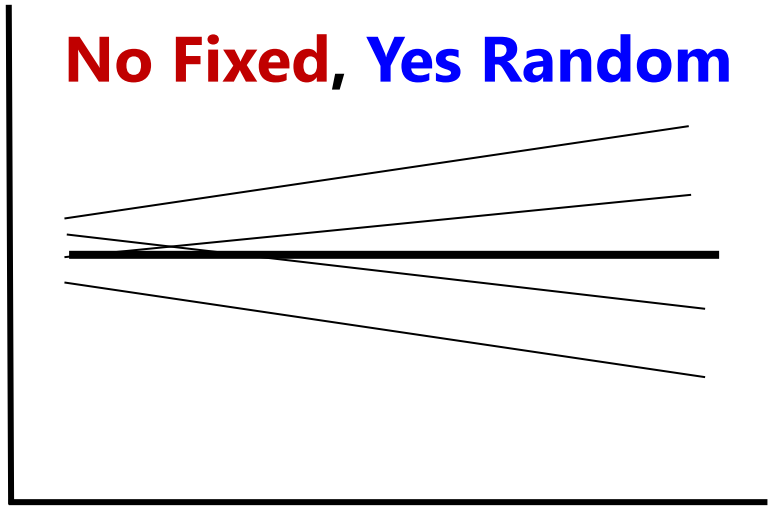
No Fixed, No Random



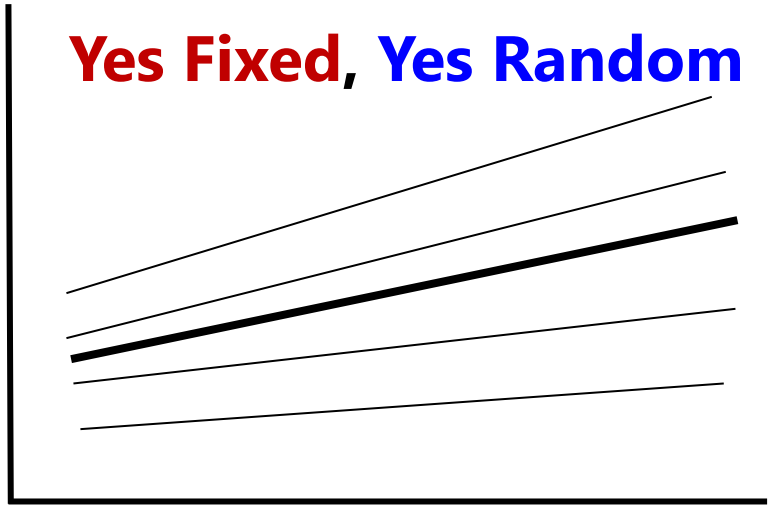
Yes Fixed, No Random



No Fixed, Yes Random



Yes Fixed, Yes Random



Cluster-MC Predictor* with a Random Slope

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

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

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

$\beta_{1c} = \gamma_{10} + U_{1c}$

$WCx_{pc} = L1x_{pc} - \overline{L1x}_c$ → only has L1 within variation

$CMx_c = \overline{L1x}_c - C_2$ → only has L2 between variation

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

Because WCx_{pc} and CMx_c are uncorrelated, each gets the <u>total</u> effect for its level: L1 = within, L2 = between

By adding U_{1c}, we are estimating its variance across clusters as $\tau^2_{U_1}$ *and* its covariances (or correlations in R) with existing random effects at that level

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*

* 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).**

When Cluster-MC ≠ Grand-MC: Random Slopes!

Cluster-MC (variable-centering): $WCx_{pc} = L1x_{pc} - CMx_c$

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

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

$\beta_{1c} = \gamma_{10} + U_{1c}$

Btw, I am using a centering constant = 0 at both levels to simplify the notation so that $\overline{L1x_c} = CMx_c$.

These two models for the means (fixed effects side) are equivalent!

$y_{pc} = \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(L1x_{pc} - CMx_c) + U_{0c} + U_{1c}(L1x_{pc} - CMx_c) + e_{pc}$

$y_{pc} = \gamma_{00} + (\gamma_{01} - \gamma_{10})(CMx_c) + \gamma_{10}(L1x_{pc}) + U_{0c} + U_{1c}(L1x_{pc} - CMx_c) + e_{pc}$

Grand-MC (constant-centering):

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

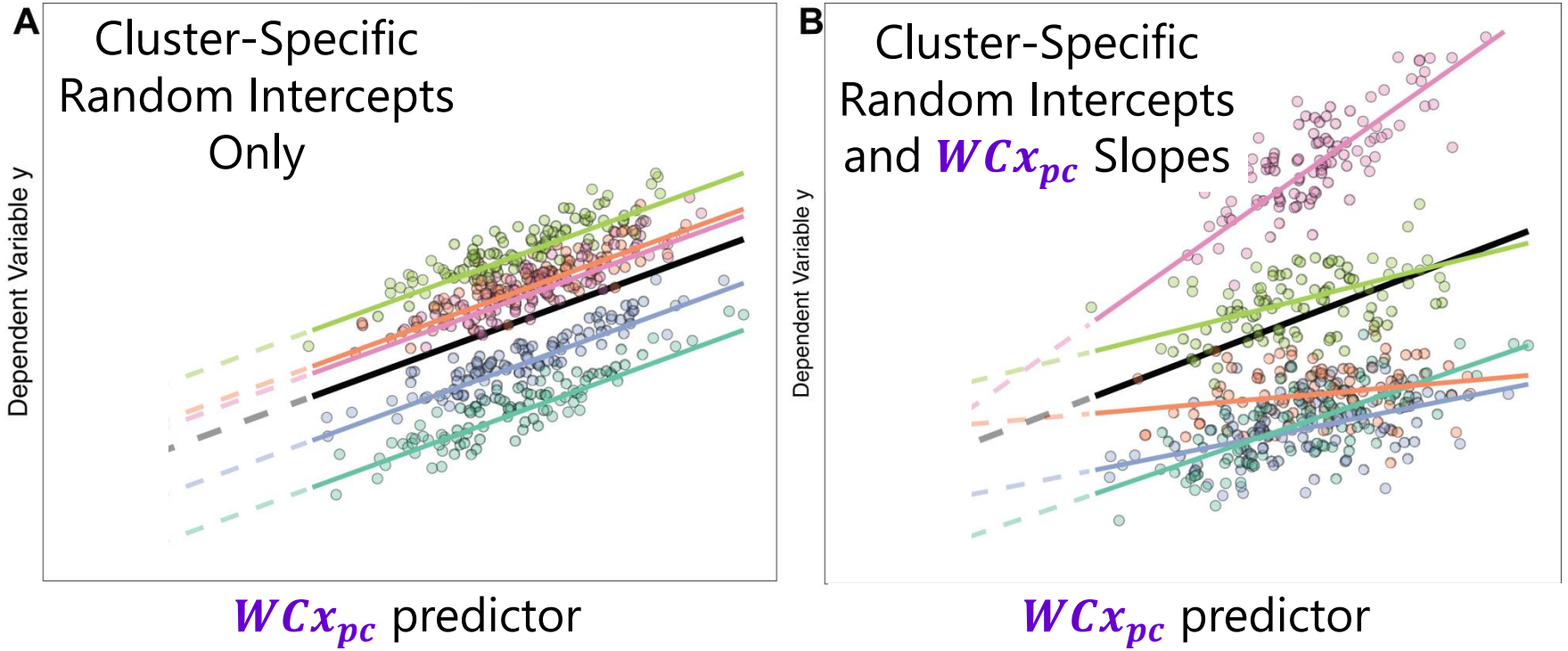
Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(CMx_c) + U_{0c}$

$\beta_{1c} = \gamma_{10} + U_{1c}$

→ $y_{pc} = \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(L1x_{pc}) + U_{0c} + U_{1c}(L1x_{pc}) + e_{pc}$

L2 predictor CMx_c is also multiplied by the random slope in Grand-MC. So these random parts cannot be made equivalent without a separate contextual L2 "random slope" for CMx_c! (Rights & Sterba, 2023)

Random Level-1 Slopes Across Clusters



- Both: the black line conveys the fixed slope for WCx_{pc}, γ_{01}
- Right: deviation for each cluster's WCx_{pc} slope is given by U_{1c}
  - Left: $\beta_{1c} = \gamma_{01}$ Right: $\beta_{1c} = \gamma_{10} + U_{1c}$

How to choose? Likelihood ratio test: $-2\Delta LL$!

Image borrowed from: https://peerj.com/articles/4794/

Example: Cluster-MC Random Slope

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

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + \gamma_{03}\left(\overline{Verbal_c} - 10\right) + U_{0c}$

$\beta_{1c} = \gamma_{10} + U_{1c}$

Adding L2 random slope variance of U_{1c} (as $\tau^2_{U_1}$) and L2 random intercept–slope covariance (as τ_{U01})

L1 WCverb = $Verbal_{pc} - \overline{Verbal_c}$

L2 CMverb10 = $\overline{Verbal_c} - 10$

```
print("Add Random Cluster-MC Verbal Slopes")
Model2 = lmer(data=Lecture4, REML=TRUE,
              formula=langpost~1+hw2+mixgrd+CMverb10+WCverb
                              +(1+WCverb|schoolID))
lmerSummary(Model2, LRT=TRUE, randCI=TRUE, slopeRel=TRUE)
```

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.528	0.358	176.855	116.127	<.001
hw2	-0.095	0.446	177.653	-0.213	0.832
mixgrd	-0.934	0.505	200.896	-1.848	0.066
CMverb10	3.621	0.265	208.962	13.678	<.001
WCverb	2.449	0.068	151.279	35.843	<.001

Model for the Variance:

Groups	Name	Variance	Std.Dev	Corr
schoolID	Intercept	8.468	2.910	
	WCverb	0.224	0.473	-0.505
Residual		39.758	6.305	

Likelihood ratio test of random slope variance (and intercept–slope correlation):
$-2\Delta LL(\sim 2) = 19.29$, $p < .0001$

Example: Cluster-MC Random Slope

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

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + \gamma_{03}\left(\overline{Verbal_c} - 10\right) + U_{0c}$

$\beta_{1c} = \gamma_{10} + U_{1c}$

With random slope U_{1c}:

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.528	0.358	176.855	116.127	<.001
hw2	-0.095	0.446	177.653	-0.213	0.832
mixgrd	-0.934	0.505	200.896	-1.848	0.066
CMverb10	3.621	0.265	208.962	13.678	<.001
WCverb	2.449	0.068	151.279	35.843	<.001

All estimates wiggle after adding U_{1c} because they are solved for after estimating the model for the variance parameters (stay tuned!)

Without random slope U_{1c}:

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.579	0.362	172.098	114.721	<.001
hw2	-0.053	0.459	178.450	-0.115	0.909
mixgrd	-1.121	0.516	197.041	-2.173	0.031
CMverb10	3.660	0.271	207.383	13.508	<.001
WCverb	2.423	0.057	3372.684	42.370	<.001

Without U_{1c}, for the L1 verbal slope γ_{10}, the SE is too small and the DDF is too large

Effect Size via 95% Random Slope CIs

- e.g., "I have a significant fixed $\boldsymbol{WCx_{pc}}$ slope of $\boldsymbol{\gamma_{10} = 2.449}$, so it has a **positive effect on average**. I also have a significant L2 random slope variance for $\boldsymbol{WCx_{pc}}$ of $\boldsymbol{\tau^2_{U_1} = 0.224}$, so clusters need their own $\boldsymbol{WCx_{pc}}$ slope. But how big is a variance of $\boldsymbol{0.224}$ (i.e., besides > 0)?"

- **95% Random Slope Confidence Intervals**
  - Can be calculated for *each L1 term that has a random effect* in your model (intercept or slope)
  - Provide range around the fixed effect within which *95% of YOUR sample* is predicted to fall based on your random effect standard deviation:

$$\text{Random Effect 95\% CI} = \text{fixed effect} \pm \left(1.96 * \sqrt{\text{Random Variance}}\right)$$

$$\text{Random Slope 95\% CI} = \gamma_{10} \pm \left(1.96 * \sqrt{\tau^2_{U_1}}\right) \rightarrow 2.449 \pm \left(1.96 * \sqrt{0.224}\right) = 1.521 \text{ to } 3.376$$

  - $\boldsymbol{WCx_{pc}}$ has a positive effect on average (its fixed slope 2.449), and the individual cluster slopes are predicted to range from 1.521 to 3.376
  - If the random slope CI contained 0, then some clusters would be predicted to have a negative $\boldsymbol{WCx_{pc}}$ slope instead!
  - Is NOT the same as typical CI (for imprecision of the fixed effect based on the fixed effect SE)!

Effect Size via Reliability Indices

How reliable is a given level-2 cluster's random effect?

Intercept Reliability (ICC_2):

$\tau^2_{U_0}$ = random intercept variance
σ^2_e = residual variance
$L1n$ = L1 sample size per L2 unit

$$ICC_2 = \frac{\tau^2_{U_0}}{\tau^2_{U_0} + \frac{\sigma^2_e}{L1n * 1}}$$

Slope Reliability (SR):

$\tau^2_{U_1}$ = random slope variance
σ^2_e = residual variance
$L1n$ = L1 sample size per L2 unit
σ^2_{L1} = variance of L1 predictor

$$SR = \frac{\tau^2_{U_1}}{\tau^2_{U_1} + \frac{\sigma^2_e}{L1n * \sigma^2_{L1}}}$$

Although slope reliability is not commonly reported, it is useful for power analyses! Choose a target slope reliability, and then work backwards to determine what the random slope variance should be →

$$\tau^2_{U_1} = \frac{SR + \frac{\sigma^2_e}{L1n * \sigma^2_{L1}}}{1 - SR}$$

Implications of Random Slopes: More Cluster Differences

- **L2 random slopes** capture a second, distinct source of cluster **dependency**—differences in **slope of a L1 person predictor**
  - Beyond the **constant** correlation for L1 persons from same L2 cluster (created by the **L2 random intercept**), the **L2 random slope** adds **non-constant** correlation varying across the values of its L1 predictor (e.g., WCx_{pc})
  - Also adds **quadratic heterogeneity of variance** across L1 predictor: **Why? Stay tuned!**

$$Var(y_{pc}) = \tau^2_{U_0} + \left(WCx^2_{pc} * \tau^2_{U_1}\right) + \left(2WCx_{pc} * \tau_{U_{01}}\right) + \sigma^2_e$$

- **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** they do
- Therefore, random slopes imply another role for L2 cluster predictors—to explain L2 cluster differences in **L1** WCx_{pc} slope
  - To do so, we need "**cross-level interactions**" of L2 cluster by L1 person predictors!

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

Introducing Cross-Level Interactions

- A **cross-level interaction** is among predictors at different levels
- Below is an example of a "**intra-variable**" cross-level interaction of the L1 within and L2 between parts of the same L1 predictor
- Cross-level interactions explain the random slope variance of L1 person predictor across L2 clusters $\tau^2_{U_1}$—here is a generic example:

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

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

$\beta_{1c} = \gamma_{10} + \gamma_{11}(CMx_c) + U_{1c}$

$WCx_{pc} = L1x_{pc} - \overline{L1x}_c$ → only has L1 within variation

$CMx_c = \overline{L1x}_c - C_2$ → only has L2 between variation

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

γ_{10} = within effect of more $L1x_{pc}$ *than others in your cluster*, now for $CMx_c = 0$

γ_{01} = between effect of more $\overline{L1x}_c$ *than other clusters,* now for $WCx_{pc} = 0$

γ_{11} = difference in within effect of WCx_{pc} per unit CMx_c OR diff in between effect of CMx_c per unit WCx_{pc}

Example: Add 3 Cross-Level Interactions

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

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{01}(HW_c - 2) + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal_c} - 10) + U_{0c}$

$\beta_{1c} = \gamma_{10} + \gamma_{11}(HW_c - 2) + \gamma_{12}(MixGrd_c) + \gamma_{13}(\overline{Verbal_c} - 10) + U_{1c}$

Results—having more verbal ability than your peers matters more for your language in mixed-grade schools:

```
print("Add 3 Cross-Level Interactions")
Model3 = lmer(data=Lecture4, REML=TRUE,
              formula=langpost~1+hw2+mixgrd+CMverb10+WCverb
              +hw2:WCverb +mixgrd:WCverb +CMverb10:WCverb
              +(1+WCverb|schoolID))
lmerSummary(Model3)
PseudoR2(baseModel=Model2, altModel1=Model3)
```

```
Fixed Effects:
 Term                Est    SE      DF t.value p.value
 Intercept        41.583 0.363 172.381 114.560   <.001
 hw2              -0.046 0.459 178.597  -0.100   0.920
 mixgrd           -1.137 0.516 196.788  -2.203   0.029
 CMverb10          3.644 0.271 206.890  13.448   <.001
 WCverb            2.390 0.100 128.552  23.861   <.001
 hw2:WCverb       -0.056 0.130 142.904  -0.429   0.668
 mixgrd:WCverb     0.321 0.159 228.287   2.021   0.044
 CMverb10:WCverb  -0.044 0.078 181.771  -0.559   0.577
```

L1 WCverb slope is now specifically for hw=2, mixgrd=0, and CMverb=10; those 3 slopes are specifically for WCverb=0 (student at school mean)

Relative to the previous model, the 3 new cross-level interactions explained **0.03%** of the **level 2 random WCverb slope** variance, but slope reliability = 0.268 to begin with ☹

L1 WCverb slope is significantly more positive (stronger) in schools with mixed grades (and nonsignificantly weaker in schools with more homework and higher mean verbal ability)

Intermediate Summary: Part 1

- In a two-level model, **cluster-level-2** predictors can have **fixed slopes**
  - For example, does mean school achievement differ between rural and urban schools?

- **Person-level-1** predictors refer can have both **fixed slopes AND random slopes** (= fixed-slope random deviations that vary over clusters)
  - For example, does student achievement differ by student SES?
    - **Fixed slope:** Is there an SES difference ("gap") in achievement on average within clusters?
    - **Random slope:** Does the extent of the within-cluster SES gap differ across schools? *(specified to multiply the cluster-MC or latent-centered version of L1 SES)*
  - When a level-1 predictor has both a fixed slope and a random slope, the fixed effect is the average of the level-2 per-cluster slopes (can differ due to different cluster sizes), so always keep it

- Significance tests for **random slope variances** (with covariances) via LRT ($-2\Delta LL$)
  - If using REML, to-be-compared models must have same fixed effects!
  - Random slope CIs and reliability indices can help convey practical effect size
  - A person-level-1 predictor's random slope variance gets explained by **cross-level interactions** of a cluster-level-2 predictor with that person-level-1 predictor (for why clusters need their own slope)

Cross-Level Interactions: Danger Ahead!

- We continue with a simplified version of the prior example without the 3 nonsignificant slopes (γ_{01} and γ_{11} for hw2, and γ_{13} for WCverb*CMverb10) in the level-2 model:
  - **Level 1:** $Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$
  - **Level 2:** $\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal}_c - 10) + U_{0c}$

    $\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c) + U_{1c}$
  - Given **cluster-mean-centered** L1 verbal (→ within info only), the cross-level interaction γ_{12} gives the difference of the **L1 within-school** verbal slope for L2 schools with mixed grades (versus reference = not mixed), *without (yet) moderating the* ***L2 between-school*** *verbal slope*

- **IN WHAT FOLLOWS:** What if we had **constant-centered** L1 verbal (→ variance at both levels still) to get L2 contextual slopes directly as L2 fixed effects instead?
  - L1 fixed "main" verbal effect would still be the unsmushed **L1 within slope** controlling for **L2 contextual** fixed "main" effect of school mean verbal
  - The **random slope of L1 verbal would still be smushed**! To fix it, we need a "**hybrid**" model, where the **fixed slopes** and **random slopes** multiply **different level-1 predictors…!**

Hybrid Model: Fixed ≠ Random Level-1 Predictors!

- Goal: Provide **L2 contextual effects** directly on the fixed side of the model **without smushing the random slope** of L1 predictor
- Solution has been called a "**hybrid**" model (see below): **fixed slope** → constant-centered version of predictor; **random slope**→ cluster-MC version of predictor

Level 1: $$Lang_{pc} = \beta_{0c} + \beta_{1c}\left(Verbal_{pc} - 10\right) + \beta_{2c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$$

Level 2: $$\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}\left(\overline{Verbal}_c - 10\right) + U_{0c}$$

→ ***No random slope!*** $\beta_{1c} = \gamma_{10}$

→ ***No fixed slope!*** $\beta_{2c} = U_{2c}$

Composite:

$$Lang_{pc} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}\left(\overline{Verbal}_c - 10\right) + \gamma_{10}\left(Verbal_{pc} - 10\right) + U_{0c} + U_{2c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$$

Random Slope Variances by Level-1 Centering

Level 1: $Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - 10) + \beta_{2c}(Verbal_{pc} - \overline{Verbal_c}) + e_{pc}$

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal_c} - 10) + U_{0c}$

$\beta_{1c} = \gamma_{10} + (U_{1c})$; $\beta_{2c} = (U_{2c})$

(U) indicates only one or the other

With random slope U_{2c} for **WCverb** (**correct** cluster-mean-centered version):

```
lmer(data=Lecture4, REML=TRUE,
     formula=langpost~1+mixgrd+CMverb10+verb10
     +(1+WCverb|schoolID))
```

```
Model for the Variance:
 Groups    Name        Variance Std.Dev   Corr
 schoolID  Intercept    8.468   2.910
           WCverb       0.224   0.473 -0.505
 Residual              39.758   6.305
Number of obs: 3566 , groups: schoolID, 191
```

With random slope U_{1c} for **verb10** (**smushed** constant-centered version):

```
lmer(data=Lecture4, REML=TRUE,
     formula=langpost~1+mixgrd+CMverb10+verb10
     +(1+verb10|schoolID))
```

```
Model for the Variance:
 Groups    Name        Variance Std.Dev   Corr
 schoolID  Intercept    8.514   2.918
           verb10       0.200   0.447 -0.534
 Residual              39.810   6.310
Number of obs: 3566 , groups: schoolID, 191
```

Hybrid Model: Variance Interpretations

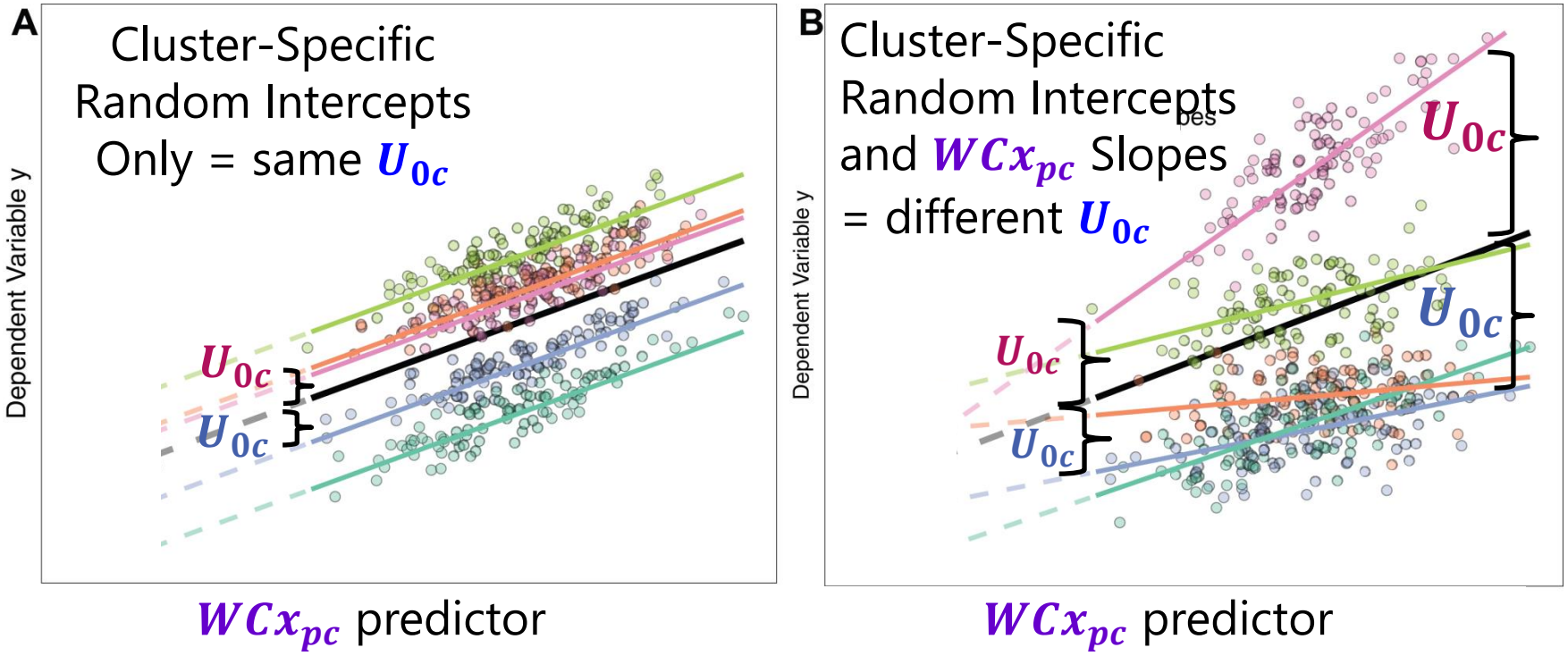
- Goal: Provide **L2 contextual effects** directly on the fixed effects side of the model **without smushing the random slope** of L1 predictor

Composite: $$Lang_{pc} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal}_c - 10) + \gamma_{10}(Verbal_{pc} - 10) + U_{0c} + U_{2c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$$

Interpreting the Model for the Variance:

- U_{0c} = **level-2 random intercept** → remaining deviation from fixed-effect-predicted school mean language for school c, now **specifically when student verbal = their school mean as within-cluster verbal = 0** (with variance = $\tau^2_{U_0}$)
- U_{2c} = **level-2 random slope** → deviation of **L1 within** verbal slope for school c from γ_{10}, its average slope across all schools (with variance = $\tau^2_{U_2}$ and covariance with U_{0c} = $\tau_{U_{02}}$)
  - *If applied to constant-centered student verbal instead, it would be smushed—it would reflect both school differences in the L1 within verbal slope AND L2 intercept heteroscedasticity (bad)*
- e_{pc} = **level-1 residual** = remaining deviation of the observed outcome for student p from their outcome predicted by all fixed and random effects

Random Level-1 Slopes Across Clusters



- Both: the black line conveys the fixed slope for WCx_{pc}, γ_{01}
- After adding a L1 predictor's random slope, the L2 random intercept no longer applies equally along that predictor—**the random intercept is then specifically at a predictor value = 0** (and will differ at a new definition of "0" if centered differently)

Image borrowed from: https://peerj.com/articles/4794/

Hybrid Model: Add a Cross-Level Interaction

- Goal: explain school differences in **L1 within-school** verbal slope (random variance $\tau^2_{U_2}$) using cross-level interaction with **L2** $MixGrd_c$
  - Effect size would be found using pseudo-R^2 for the random slope variance, so **ALWAYS test for L2 random slope variance of the L1 fixed slope first** before examining any of its cross-level interactions—otherwise you'll have high Type I errors for the cross-level interaction if you omit a necessary L2 random slope!

Level 1: $$Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - 10) + \beta_{2c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$$

Level 2:
$$\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal}_c - 10) + U_{0c}$$
$$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c)$$
$$\beta_{2c} = U_{2c}$$

- All good, right? Nope—this model is now VERY LIKELY to be mis-specified at L2
  - Same problem as when adding the fixed main effect of a constant-centered L1 predictor by itself without a fixed main effect of its L2 cluster mean!

Hybrid: Smushed Cross-Level Interaction

Level 1: $Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - 10) + \beta_{2c}(Verbal_{pc} - \overline{Verbal_c}) + e_{pc}$

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal_c} - 10) + U_{0c}$

$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c)$; $\beta_{2c} = U_{2c}$

```
lmer(data=Lecture4, REML=TRUE,
     formula=langpost~1+mixgrd+CMverb10+verb10
     +verb10:mixgrd+(1+WCverb|schoolID))
```

Interpreting Fixed Effects:

- γ_{00} = 41.59 = **intercept**: expected language for a student with verbal = 10 from a school with school mean verbal = 10 and no mixed grades
- γ_{10} = 2.35 = ***conditional* L1 within slope**: difference in student language per unit higher verbal than their school mean, *specifically for schools without mixed grades*
- γ_{03} = 1.04 = **L2 contextual slope**: extra difference in school language per unit higher school mean verbal than other schools (controlling for student verbal; not explicitly conditional on mixed grade)
- $\gamma_{10} + \gamma_{03}$ = 3.39 = **L2 between slope**: difference in school language per unit higher school mean verbal than other schools (NOT controlling for student verbal; not explicitly conditional on mixed grade)
- γ_{12} = 0.37 = **smushed cross-level interaction**: how the **L1 within slope** AND the **L2 between slope** each differ in schools with mixed grades → **assumes equal moderation by mixed grade** of L1 within and L2 between slopes!

Hybrid: Unsmushed Cross-Level Interaction

Level 1: $Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - 10) + \beta_{2c}(Verbal_{pc} - \overline{Verbal_c}) + e_{pc}$

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal_c} - 10)$

$+ \gamma_{04}(MixGrd_c)(\overline{Verbal_c} - 10) + U_{0c}$

$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c)$; $\beta_{2c} = U_{2c}$

```
lmer(data=Lecture4, REML=TRUE,
     formula=langpost~1+mixgrd+CMverb10+verb10
     +verb10:mixgrd+CMverb10:mixgrd+(1+WCverb|schoolID))
```

Interpreting Fixed Effects:

- γ_{10} = 2.36 = ***conditional*** **L1 within slope**: difference in student language per unit higher verbal than their school mean, *specifically for schools without mixed grades*
- γ_{03} = 0.82 = ***conditional*** **L2 contextual slope**: extra difference in school language per unit higher school mean verbal than other schools (controlling for student verbal), *now specifically for schools without mixed grades*
- $\gamma_{10} + \gamma_{03}$ = 3.18 = ***conditional*** **L2 between slope**: difference in school language per unit higher school mean verbal than other schools (NOT controlling for student verbal), *now specifically for schools without mixed grades*
- γ_{12} = 0.33 = **unsmushed cross-level interaction**: how the **L1 within** verbal slope differs in schools with mixed grades relative to unmixed grades
- γ_{04} = 0.42 = **new level-2 interaction = TEST OF SMUSHED INTERACTION**: how the **L2 contextual** verbal slope differs in schools with mixed grades (added to unsmush cross-level interaction γ_{12}) relative to unmixed grades
- $\gamma_{12} + \gamma_{04}$ = 0.75 = **implied level-2 interaction**: how the **L2 between** verbal slope differs in schools with mixed grades relative to unmixed grades

Non-Hybrid: All Cluster-MC Version

Level 1: $Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$

Level 2: $\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal}_c - 10) + \gamma_{04}(MixGrd_c)(\overline{Verbal}_c - 10) + U_{0c}$

$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c) + U_{1c}$

```
lmer(data=Lecture4, REML=TRUE,
     formula=langpost~1+mixgrd+CMverb10+WCverb
     +WCverb:mixgrd+CMverb10:mixgrd+(1+WCverb|schoolID))
```

Interpreting Fixed Effects:

- γ_{10} = 2.36 = ***conditional*** **L1 within slope**: difference in student language per unit higher verbal than their school mean, *specifically for schools without mixed grades*
- γ_{03} = 3.18 = ***conditional*** **L2 between slope**: difference in school language per unit higher school mean verbal than other schools (NOT controlling for student verbal), *for schools without mixed grades*
- $\gamma_{03} - \gamma_{10}$ = 0.82 = ***conditional*** **L2 contextual slope**: extra difference in school language per unit higher school mean verbal than other schools (controlling for student verbal), *for schools without mixed grades*
- γ_{12} = 0.33 = **guaranteed-to-always-be-unsmushed cross-level interaction**: how the **L1 within** verbal slope differs in schools with mixed grades relative to without mixed grades
- γ_{04} = 0.75 = **level-2 interaction**: how the **L2 between** verbal slope differs in schools with mixed grades relative to without mixed grades
- $\gamma_{04} - \gamma_{12}$ = 0.42 = **implied level-2 interaction**: how the **L2 contextual** verbal slope differs in schools with mixed grades (or how moderation differs: BC − WC) relative to without mixed grades

Hybrid vs. Cluster-MC: Different L2 Slopes!

Hybrid Level-1:
$\beta_{1c}(Verbal_{pc} - 10)$
→ **L2 Contextual** Effects

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.591	0.293	167.038	141.856	<.001
mixgrd	-1.107	0.511	197.321	-2.166	0.031
CMverb10	0.820	0.403	182.174	2.037	0.043
verb10	2.361	0.079	123.558	30.055	<.001
mixgrd:verb10	0.336	0.157	238.509	2.146	0.033
mixgrd:CMverb10	0.417	0.552	236.541	0.756	0.451

Cluster-MC Level-1:
$\beta_{1c}(Verbal_{pc} - \overline{Verbal_c})$
→ **L2 Between** Effects

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.591	0.293	167.038	141.856	<.001
mixgrd	-1.107	0.511	197.321	-2.166	0.031
CMverb10	3.181	0.396	171.907	8.025	<.001
WCverb	2.361	0.079	123.558	30.055	<.001
mixgrd:WCverb	0.336	0.157	238.509	2.146	0.033
mixgrd:CMverb10	0.753	0.529	200.554	1.424	0.156

L1 within verbal slope is signif more positive (stronger) by **0.336** in mixed-grade schools
L2 between verbal slope is n.s. more positive (stronger) by **0.753** in mixed-grade schools
L2 contextual verbal slope is n.s. more positive (stronger) by **0.417** in mixed-grade schools

Same Model for the Variance Either Way

Level 1: $Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - 10) + \beta_{2c}(Verbal_{pc} - \overline{Verbal_c}) + e_{pc}$

Level 2:

$$\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal_c} - 10) + \gamma_{04}(MixGrd_c)(\overline{Verbal_c} - 10) + U_{0c}$$

$$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c); \beta_{2c} = U_{2c}$$

Hybrid → $\beta_{1c}(Verbal_{pc} - 10)$
Cluster-MC → $\beta_{1c}(Verbal_{pc} - \overline{Verbal_c})$

Interpreting the Model for the Variance:

- U_{0c} = **level-2 random intercept** → remaining deviation from fixed-effect-predicted school mean language for school *c*, **specifically when student verbal = their school mean** (variance = $\tau^2_{U_0}$)
- U_{2c} = **level-2 random slope** → remaining deviation from fixed-effect-predicted **L1 within** verbal slope for school *c* (variance = $\tau^2_{U_2}$ and covariance with U_{0c} = $\tau_{U_{02}}$)
  - *If applied to constant-centered student verbal instead, it would be smushed—it would reflect both school differences in the L1 within verbal slope AND L2 intercept heteroscedasticity (bad)*
- e_{pc} = **level-1 residual** = remaining deviation of the observed outcome for student *p* from their outcome predicted by all fixed and random effects

Intra-Variable Cross-Level Interactions

- What if we wanted to see how the **L1 within effect** *(of more verbal ability than your peers on your language)* depends on your school's verbal ability *(school mean verbal)*?
  - Back to the **hybrid model** to illustrate:

Level 1: $$Lang_{pc} = \beta_{0c} + \beta_{1c}(Verbal_{pc} - 10) + \beta_{2c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$$

Level 2:
$$\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}(\overline{Verbal}_c - 10) + \gamma_{04}(MixGrd_c)(\overline{Verbal}_c - 10) + U_{0c}$$

$$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c) + \gamma_{13}(\overline{Verbal}_c - 10); \beta_{2c} = U_{2c}$$

- Same potential for a smushed cross-level interaction when using a constant-centered L1 predictor in the intra-variable interaction
  - Slope γ_{13} says the L1 within and L2 between verbal slopes are moderated to the same extent by school mean verbal
  - The solution is the same as before, but it looks strange at first…!

Intra-Variable Cross-Level Interactions

- To unsmush the cross-level interaction, we add the corresponding L2 interaction with the L2 moderator, just as we did before…

Level 1: $$Lang_{pc} = \beta_{0c} + \beta_{1c}\left(Verbal_{pc} - 10\right) + \beta_{2c}(Verbal_{pc} - \overline{Verbal}_c) + e_{pc}$$

Level 2:
$$\beta_{0c} = \gamma_{00} + \gamma_{02}(MixGrd_c) + \gamma_{03}\left(\overline{Verbal}_c - 10\right) + \gamma_{04}(MixGrd_c)\left(\overline{Verbal}_c - 10\right) + \gamma_{05}\left(\overline{Verbal}_c - 10\right)\left(\overline{Verbal}_c - 10\right) + U_{0c}$$

$$\beta_{1c} = \gamma_{10} + \gamma_{12}(MixGrd_c) + \gamma_{13}\left(\overline{Verbal}_c - 10\right); \beta_{2c} = U_{2c}$$

- …the solution is a **quadratic slope** for L2 school mean verbal!
  - γ_{13} = how the **L1 within** verbal slope differs by school mean verbal
  - γ_{05} = how the **L2 contextual** verbal slope differs by school mean verbal
  - $\gamma_{13} + \gamma_{05}$ = how the **L2 between** verbal slope differs by school mean verbal

Hybrid vs. Cluster-MC: Different L2 Slopes!

Hybrid: $\boldsymbol{\beta_{1c}(Verbal_{pc} - 10)}$ → **L2 Contextual** Effects

```
print("Hybrid Model with Intra-Variable Interactions")
Model10 = lmer(data=Lecture4, REML=TRUE,
               formula=langpost~1+mixgrd+CMverb10+verb10+I(CMverb10^2)
               +verb10:mixgrd+CMverb10:mixgrd
               +verb10:CMverb10+(1+WCverb|schoolID))
lmerSummary(Model10)
con1DTable(Model10, L=list("L2 Between Verbal Main Effect"         =c(0,0,1,1,0,0,0,0),
                           "L2 Between Verbal*Verbal Interaction" =c(0,0,0,0,1,0,0,1)))
```

Cluster-MC: $\boldsymbol{\beta_{1c}(Verbal_{pc} - \overline{Verbal_c})}$ → **L2 Between** Effects

```
print("Cluster-MC Model with Intra-Variable Interactions")
Model9 = lmer(data=Lecture4, REML=TRUE,
              formula=langpost~1+mixgrd+CMverb10+WCverb+I(CMverb10^2)
              +WCverb:mixgrd+CMverb10:mixgrd
              +WCverb:CMverb10+(1+WCverb|schoolID))
lmerSummary(Model9)
con1DTable(Model9, L=list("L2 Contextual Verbal Main Effect"         =c(0,0,1,-1,0,0,0,0),
                          "L2 Contextual Verbal*Verbal Interaction" =c(0,0,0,0,1,0,0,-1)))
```

Hybrid vs. Cluster-MC: Different L2 Slopes!

Hybrid:

$\boldsymbol{\beta_{1c}(Verbal_{pc} - 10)}$

→ **L2 Context** Effects

L1 within verbal slope **2.359** is n.s. less positive (weaker) by **0.043** per unit school mean verbal

L2 between verbal slope **3.212** is n.s. less positive by **0.425(*2)** per unit school mean verbal

L2 contextual verbal slope **0.854** is n.s. less positive by **0.382(*2)** per unit school mean verbal

Cluster-MC:

$\boldsymbol{\beta_{1c}(Verbal_{pc} - \overline{Verbal_c})}$

→ **L2 Between** Effects

The **(*2)** is needed because the quadratic slope provides half the rate of acceleration instead!

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.807	0.301	167.169	138.999	<.001
mixgrd	-0.832	0.516	196.971	-1.611	0.109
CMverb10	0.854	0.404	193.024	2.115	0.036
verb10	2.359	0.079	122.803	29.832	<.001
I(CMverb10^2)	-0.382	0.167	345.895	-2.290	0.023
mixgrd:verb10	0.340	0.158	230.599	2.151	0.032
mixgrd:CMverb10	-0.125	0.576	232.197	-0.217	0.828
CMverb10:verb10	-0.043	0.078	179.153	-0.557	0.578

Fixed Effects:

Term	Est	SE	DF	t.value	p.value
Intercept	41.807	0.301	167.169	138.999	<.001
mixgrd	-0.832	0.516	196.971	-1.611	0.109
CMverb10	3.212	0.398	181.255	8.077	<.001
WCverb	2.359	0.079	122.803	29.832	<.001
I(CMverb10^2)	-0.425	0.148	234.193	-2.868	0.005
mixgrd:WCverb	0.340	0.158	230.599	2.151	0.032
mixgrd:CMverb10	0.215	0.553	199.057	0.388	0.698
CMverb10:WCverb	-0.043	0.078	179.153	-0.557	0.578

Prerequisites for Cross-Level Interactions?

- Let's go back to this generic cluster-MC model for a moment:

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

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

$\beta_{1c} = \gamma_{10} + \gamma_{11}(CMx_c) + U_{1c}$

Can I still include γ_{11} without U_{1c}?

- If the U_{1c} random slope for WCx_{pc} was not initially significant (via $-2\Delta LL$), can I still test cross-level interactions with WCx_{pc}?
  - **"NO"**: If a level-1 slope does not vary randomly over clusters, then it has ~0 variance to predict (so cross-level interactions with that level-1 slope are not necessary); its SE and DDF could be inaccurate SE if $\tau^2_{U_1} > 0$ at all
  - **"YES":** Because power to detect random slope variances is lower than power to detect fixed effects (especially with small L2n), cross-level interactions can still be significant even if there is "no" (~0) variance to be predicted
  - Saying yes requires new vocabulary…

3 Types of Effects: Fixed, Random, and Systematically (Non-Randomly) Varying

Let's say we have a significant L1 fixed slope of $\boldsymbol{WCx_{pc}}$.
What can happen if we test a L2group*$\boldsymbol{WCx_{pc}}$cross-level interaction?

	Nonsignificant $L2group_c$*$\boldsymbol{WCx_{pc}}$ interaction?	Significant $L2group_c$*$\boldsymbol{WCx_{pc}}$ interaction?
Random $\boldsymbol{WCx_{pc}}$ slope was **not** significant when initially tested	Effect of $\boldsymbol{WCx_{pc}}$ is **FIXED**	Effect of $\boldsymbol{WCx_{pc}}$ is **systematically varying**
Random $\boldsymbol{WCx_{pc}}$ was significant initially, became **not** significant after interaction		Effect of $\boldsymbol{WCx_{pc}}$ is **systematically varying**
Random $\boldsymbol{WCx_{pc}}$ was significant initially, **still** significant after interaction	Effect of $\boldsymbol{WCx_{pc}}$ is **RANDOM**	Effect of $\boldsymbol{WCx_{pc}}$ is **RANDOM**

The effects of person-level-1 predictors can be fixed, random, or systematically varying. The effects of cluster-level-2 predictors can only be fixed or systematically varying (not random, *at least in the traditional sense that is not creating intercept heteroscedasticity*).

Explained Variance by Fixed Slopes

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

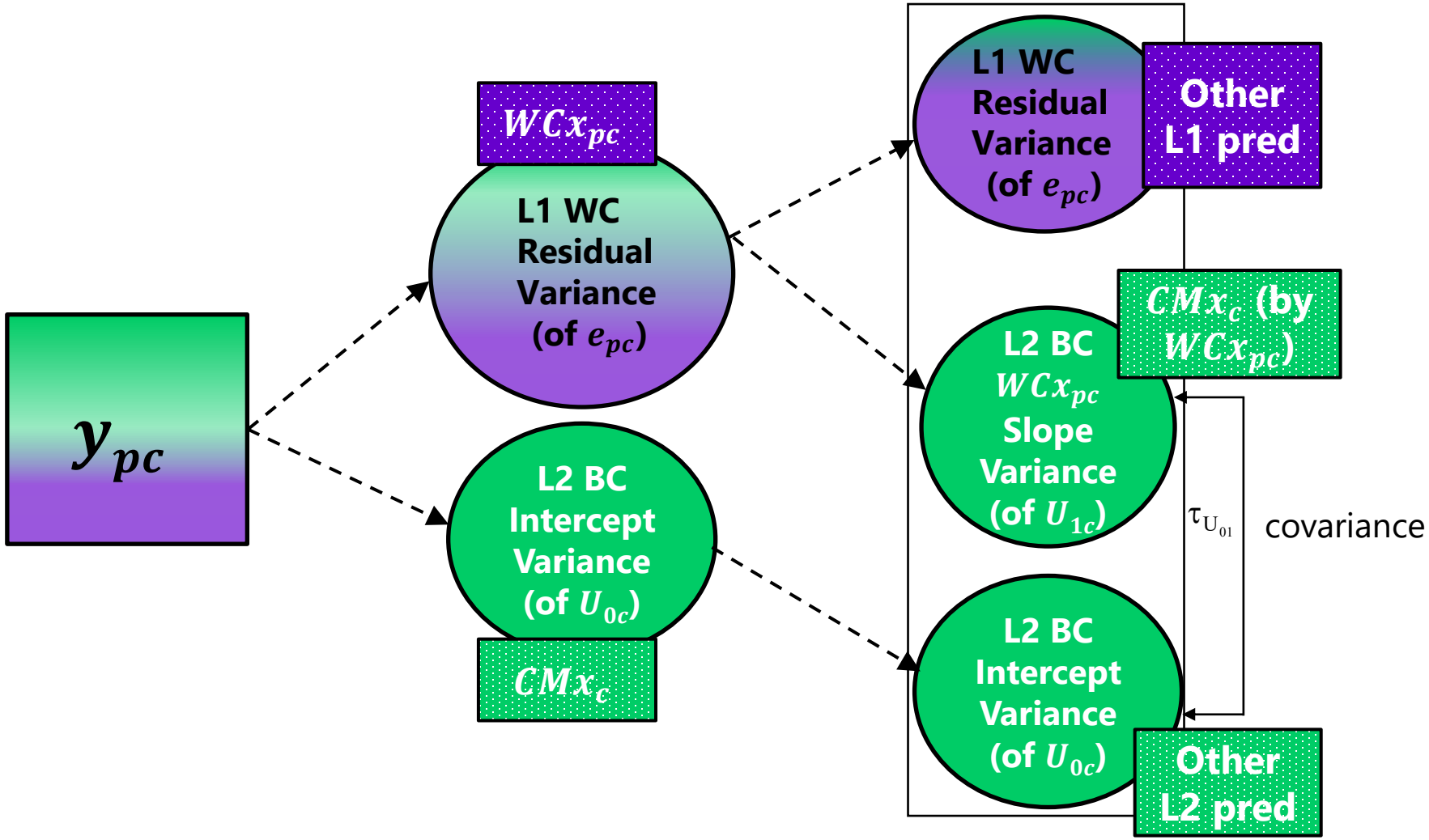
Intermediate Summary: Part 2

- A **level-2 random slope** variance allows cluster differences in the within-cluster effect of a person-level-1 predictor
  - Should be specified to multiply the cluster-MC or latent-centered version of the person-level-1 predictor, otherwise the random slope will be a new kind of smushed!
  - Implies **quadratic heterogeneity of variance** and covariance across the within part of the L1 predictor (and L2 mean part if random slope multiplies both)
  - Implies **another way that clusters differ from each other** (to be explained by **cross-level interactions** between that person-level-1 predictor and cluster-level-2 predictors)

- **Meaning of cross-level interactions vary by type of level-1 predictor:**
  - Cluster-MC: WCx*L2z → L1 within x slope ONLY moderated by L2z
  - Constant-C: L1x*L2z only → L1 within x slope AND L2 between x slope moderated by L2z to the same extent → smushed
- **After adding the corresponding L2 interaction of CMx*L2z:**
  - Cluster-MC: CMx*L2z → How L2 between x slope is moderated by L2z (was 0)
  - Constant-C: CMx*L2z → How L2 contextual x slope is moderated by L2z (was 0); also difference in moderation of L1 within x slope and L2 between x slope by L2z

How MLM "Handles" Dependency

- How does MLM "handle" dependency? By forming a new random effect variance component (or "pile" of variance) for each source of dependency



Model-Implied Variance and Covariance

- So far we've only used scalar equations to describe how the model predicts each person's outcome, but to understand the model-implied pattern of variance and covariance across persons and clusters, we need to show the model using matrix notation instead!

- Example cluster-MC model with a random intercept only:

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

Level 2: $$\beta_{0c} = \gamma_{00} + \gamma_{01}(CMx_{c}) + U_{0c}$$

$$\beta_{1c} = \gamma_{10}$$

Composite: $$y_{pc} = \gamma_{00} + \gamma_{01}(CMx_{c}) + \gamma_{10}(WCx_{pc}) + U_{0c} + e_{pc}$$

Matrices per Cluster: $$\mathbf{Y}_c = \mathbf{X}_c\boldsymbol{\gamma} + \mathbf{Z}_c\mathbf{U}_c + \mathbf{E}_c$$

Example Model for $L1n = 4$ in One Cluster

Random Int Model: $y_{pc} = \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{pc}) + U_{0c} + e_{pc}$

$$\mathbf{Y}_c = \mathbf{X}_c \quad \boldsymbol{\gamma} + \mathbf{Z}_c \quad \mathbf{U}_c + \quad \mathbf{E}_c$$

$$\begin{bmatrix} y_{1c} \\ y_{2c} \\ y_{3c} \\ y_{4c} \end{bmatrix} = \begin{bmatrix} 1 & CMx_c & WCx_{1c} \\ 1 & CMx_c & WCx_{2c} \\ 1 & CMx_c & WCx_{3c} \\ 1 & CMx_c & WCx_{4c} \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [U_{0c}] + \begin{bmatrix} e_{1c} \\ e_{2c} \\ e_{3c} \\ e_{4c} \end{bmatrix}$$

$$\begin{bmatrix} y_{1c} \\ y_{2c} \\ y_{3c} \\ y_{4c} \end{bmatrix} = \begin{bmatrix} \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{1c}) \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{2c}) \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{3c}) \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{4c}) \end{bmatrix} + \begin{bmatrix} U_{0c} \\ U_{0c} \\ U_{0c} \\ U_{0c} \end{bmatrix} + \begin{bmatrix} e_{1c} \\ e_{2c} \\ e_{3c} \\ e_{4c} \end{bmatrix}$$

$$\begin{bmatrix} y_{1c} \\ y_{2c} \\ y_{3c} \\ y_{4c} \end{bmatrix} = \begin{bmatrix} \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{1c}) + U_{0c} + e_{1c} \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{2c}) + U_{0c} + e_{2c} \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{3c}) + U_{0c} + e_{3c} \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{4c}) + U_{0c} + e_{4c} \end{bmatrix}$$

$\mathbf{X}_c$ = $L1n \times k$ values of **predictors with fixed effects**, so can differ by cluster ($k = 3$ here)

$\boldsymbol{\gamma}$ = $k \times 1$estimated **fixed effects**→ same for all clusters ($k = 3$ here)

$\mathbf{Z}_c$ = $L1n \times u$ values of **predictors with random effects**, so can differ by cluster ($u = 1$ here)

$\mathbf{U}_c$ = $u \times 1$ estimated cluster-specific **random effects** (here, just U_{0c})

$\mathbf{E}_c$ = $L1n \times L1n$ person-specific cluster residuals

Same Random Intercept Model: Predicted Marginal Variance–Covariance **V** Matrix per Cluster

$$\mathbf{V}_c = \mathbf{Z}_c \quad \mathbf{G} \quad \mathbf{Z}_c^T + \mathbf{R}_c$$

$$\mathbf{V}_c = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \left[\tau^2_{U_0} \right] \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} \sigma^2_e & 0 & 0 & 0 \\ 0 & \sigma^2_e & 0 & 0 \\ 0 & 0 & \sigma^2_e & 0 \\ 0 & 0 & 0 & \sigma^2_e \end{bmatrix}$$

$$\mathbf{V}_c = \begin{bmatrix} \tau^2_{U_0} + \sigma^2_e & \tau^2_{U_0} & \tau^2_{U_0} & \tau^2_{U_0} \\ \tau^2_{U_0} & \tau^2_{U_0} + \sigma^2_e & \tau^2_{U_0} & \tau^2_{U_0} \\ \tau^2_{U_0} & \tau^2_{U_0} & \tau^2_{U_0} + \sigma^2_e & \tau^2_{U_0} \\ \tau^2_{U_0} & \tau^2_{U_0} & \tau^2_{U_0} & \tau^2_{U_0} + \sigma^2_e \end{bmatrix}$$

$\mathbf{Z}_c$ = $L1n \times u$ values of **predictors with random effects**, so can differ by cluster ($u = 1$ here)

$\mathbf{Z}_c^T$ = $\mathbf{Z}_c$ transposed → on its side

$\mathbf{G}_c$ = $u \times u$ estimated **random effects variances and covariances**, so will be the same for all clusters (here, just $\boldsymbol{\tau^2_{U_0}}$ = intercept variance)

$\mathbf{R}_c$ = $L1n \times L1n$ **person residual variances and covariances**, so will be same for all clusters (here, same $\boldsymbol{\sigma^2_e}$ on the diagonal because persons are exchangeable; all 0 values on the off-diagonals because persons are conditionally independent)

Adding a Random Slope Implies…

- **Clusters differ** from each other randomly **in TWO ways**—in intercept (U_{0c}) and the slope of a person predictor (U_{1c}), which implies **TWO kinds of between-cluster variance**, which translates to **TWO sources of cluster dependency** → TWO reasons for the correlation of outcomes from persons in the same cluster

- Example cluster-MC model adding a **random slope** for WCx_{pc}:

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

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

$$\beta_{1c} = \gamma_{10} + U_{1c}$$

Composite: $$y_{pc} = \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{pc}) + U_{0c} + U_{1c}(WCx_{pc}) + e_{pc}$$

Matrices per Cluster: $$\mathbf{Y}_c = \mathbf{X}_c\boldsymbol{\gamma} + \mathbf{Z}_c\mathbf{U}_c + \mathbf{E}_c$$

Example Model for $L1n = 4$ in One Cluster

Random Slope Model: $y_{pc} = \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{pc}) + U_{0c} + U_{1c}(WCx_{pc}) + e_{pc}$

$$\mathbf{Y}_c = \mathbf{X}_c \quad \boldsymbol{\gamma} \quad + \quad \mathbf{Z}_c \quad \mathbf{U}_c \quad + \quad \mathbf{E}_c$$

$$\begin{bmatrix} y_{1c} \\ y_{2c} \\ y_{3c} \\ y_{4c} \end{bmatrix} = \begin{bmatrix} 1 & CMx_c & WCx_{1c} \\ 1 & CMx_c & WCx_{2c} \\ 1 & CMx_c & WCx_{3c} \\ 1 & CMx_c & WCx_{4c} \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} 1 & WCx_{1c} \\ 1 & WCx_{2c} \\ 1 & WCx_{3c} \\ 1 & WCx_{4c} \end{bmatrix} \begin{bmatrix} U_{0c} \\ U_{1c} \end{bmatrix} + \begin{bmatrix} e_{1c} \\ e_{2c} \\ e_{3c} \\ e_{4c} \end{bmatrix}$$

$$\begin{bmatrix} y_{1c} \\ y_{2c} \\ y_{3c} \\ y_{4c} \end{bmatrix} = \begin{bmatrix} \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{1c}) \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{2c}) \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{3c}) \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{4c}) \end{bmatrix} + \begin{bmatrix} U_{0c} + U_{1c}(WCx_{1c}) \\ U_{0c} + U_{2c}(WCx_{2c}) \\ U_{0c} + U_{3c}(WCx_{3c}) \\ U_{0c} + U_{4c}(WCx_{4c}) \end{bmatrix} + \begin{bmatrix} e_{1c} \\ e_{2c} \\ e_{3c} \\ e_{4c} \end{bmatrix}$$

$$\begin{bmatrix} y_{1c} \\ y_{2c} \\ y_{3c} \\ y_{4c} \end{bmatrix} = \begin{bmatrix} \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{1c}) + U_{0c} + U_{1c}(WCx_{1c}) + e_{1c} \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{2c}) + U_{0c} + U_{2c}(WCx_{2c}) + e_{2c} \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{3c}) + U_{0c} + U_{3c}(WCx_{3c}) + e_{3c} \\ \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{4c}) + U_{0c} + U_{4c}(WCx_{4c}) + e_{4c} \end{bmatrix}$$

$\mathbf{X}_c$ = $L1n \times k$ values of **predictors with fixed effects**, so can differ by cluster ($k = 3$ here)

$\boldsymbol{\gamma}$ = $k \times 1$estimated **fixed effects**→ same for all clusters ($k = 3$ here)

$\mathbf{Z}_c$ = $L1n \times u$ values of **predictors with random effects**, so can differ by cluster ($u = 2$ here)

$\mathbf{U}_c$ = $u \times 2$ estimated cluster-specific **random effects** (here, U_{0c} and U_{1c})

$\mathbf{E}_c$ = $L1n \times L1n$ person-specific cluster residuals

Same Random Slope Model: Predicted Marginal Variance–Covariance **V** Matrix per Cluster

$$\mathbf{V}_c = \mathbf{Z}_c \quad \mathbf{G} \quad \mathbf{Z}_c^T + \mathbf{R}_c$$

$$\mathbf{V}_c = \begin{bmatrix} 1 & WCx_{1c} \\ 1 & WCx_{2c} \\ 1 & WCx_{3c} \\ 1 & WCx_{4c} \end{bmatrix} \begin{bmatrix} \tau^2_{U_0} & \tau_{U_{01}} \\ \tau_{U_{01}} & \tau^2_{U_1} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ WCx_{1c} & WCx_{2c} & WCx_{3c} & WCx_{4c} \end{bmatrix} + \begin{bmatrix} \sigma^2_e & 0 & 0 & 0 \\ 0 & \sigma^2_e & 0 & 0 \\ 0 & 0 & \sigma^2_e & 0 \\ 0 & 0 & 0 & \sigma^2_e \end{bmatrix}$$

$\mathbf{V}_c$ matrix = complicated, but summarized below

$\mathbf{V}_c$ matrix: Marginal Variance at a Given $[WCx]$

$$= \tau^2_{U_0} + \left[\left(WCx^2\right)\tau^2_{U_1}\right] + \left[2\left(WCx\right)\tau_{U_{01}}\right] + \sigma^2_e$$

$\mathbf{V}_c$ matrix: Marginal Covariance at a Given $[WCx_A, WCx_B]$

$$= \tau^2_{U_0} + \left[\left(WCx_A + WCx_B\right)\tau_{U_{01}}\right] + \left[\left(WCx_A * WCx_B\right)\tau^2_{U_1}\right]$$

$\mathbf{Z}_c$ = $L1n \times u$ values of **predictors with random effects**, so can differ by cluster ($u = 2$ here)

$\mathbf{Z}_c^T$ = $\mathbf{Z}_c$ transposed

$\mathbf{G}_c$ = $u \times u$ estimated **random effects variances and covariances**, so will be same for all clusters (here, $\boldsymbol{\tau^2_{U_0}}$, $\boldsymbol{\tau^2_{U_1}}$ and $\boldsymbol{\tau_{U_{01}}}$)

$\mathbf{R}_c$ = $L1n \times L1n$**person residual variances and covariances**, so will be same for all clusters (same $\boldsymbol{\sigma^2_e}$ on the diagonal and 0 values on off-diagonals)

Building a Combined **V** across Clusters: Same Random Slope Model

V for two clusters, both of size $L1n = 4$:

$$\mathbf{V} = \mathbf{Z} * \mathbf{G} * \mathbf{Z}^T + \mathbf{R}$$

$$\mathbf{V} = \begin{bmatrix} 1 & 0.0 & 0 & 0 \\ 1 & 1.0 & 0 & 0 \\ 1 & 2.0 & 0 & 0 \\ 1 & 3.0 & 0 & 0 \\ 0 & 0 & 1 & 0.2 \\ 0 & 0 & 1 & 1.4 \\ 0 & 0 & 1 & 2.3 \\ 0 & 0 & 1 & 3.5 \end{bmatrix} \begin{bmatrix} \tau^2_{U_0} & \tau_{U_{01}} & 0 & 0 \\ \tau_{U_{01}} & \tau^2_{U_1} & 0 & 0 \\ 0 & 0 & \tau^2_{U_0} & \tau_{U_{01}} \\ 0 & 0 & \tau_{U_{01}} & \tau^2_{U_1} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0.0 & 1.0 & 2.0 & 3.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0.2 & 1.4 & 2.3 & 3.5 \end{bmatrix} + \begin{bmatrix} \sigma^2_e & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_e & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_e & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_e & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_e & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_e & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_e & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_e \end{bmatrix}$$

- The combined **V** matrix across all clusters is used in estimation
- It has a "**block diagonal**" structure → predictions are given for each cluster, but 0 values are given for the elements that describe relationships across clusters (because clusters are supposed to be independent in a two-level model!)

Building a Combined **V** across Clusters: Same Random Slope Model

V for a cluster with $L1n = 4$ and a cluster with $L1n = 3$:

$$\mathbf{V} = \mathbf{Z} * \mathbf{G} * \mathbf{Z}^{T} + \mathbf{R}$$

$$\mathbf{V} = \begin{bmatrix} 1 & 0.0 & 0 & 0 \\ 1 & 1.0 & 0 & 0 \\ 1 & 2.0 & 0 & 0 \\ 1 & 3.0 & 0 & 0 \\ 0 & 0 & 1 & 0.2 \\ 0 & 0 & 1 & 1.4 \\ 0 & 0 & 1 & 3.5 \end{bmatrix} \begin{bmatrix} \tau^2_{U_0} & \tau_{U_{01}} & 0 & 0 \\ \tau_{U_{01}} & \tau^2_{U_1} & 0 & 0 \\ 0 & 0 & \tau^2_{U_0} & \tau_{U_{01}} \\ 0 & 0 & \tau_{U_{01}} & \tau^2_{U_1} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0.0 & 1.0 & 2.0 & 3.0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0.2 & 1.4 & 3.5 \end{bmatrix} + \begin{bmatrix} \sigma^2_e & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_e & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_e & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_e & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_e & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_e & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_e \end{bmatrix}$$

- Take home message: Partitioning variance into piles…
  - **Level 2 = Between** → **G** matrix of random effects variances/covariances
  - **Level 1 = Within** → **R** matrix of residual variances/covariances
  - **G** and **R** combine via **Z** into **V** matrix of marginal variances/covariances
  - These flexible options allow the outcome variances and covariances to vary in a predictor-dependent way to better match the actual data

Two Sides of Any Model: Estimation

- **Fixed Effects in the Model for the Means:**
  - How the expected outcome for a given observation varies as a function of values on *known* predictor variables
  - Fixed effects parameters do NOT need to be solved for iteratively in (residual) maximum likelihood estimation for *general* MLMs—they fall out of an equation

- **Random Effects in the Model for the Variance:**
  - How model residuals are related across observations (dependency across persons, clusters, time, etc)—*unknown* things due to sampling
  - Random effects variances and covariances can predict complex patterns of variance and covariance among the outcome residuals
  - Anything besides level-1 residual variance $\boldsymbol{\sigma_e^2}$ must be solved for iteratively—this increases the dimensionality of estimation process
  - Estimation utilizes the predicted $\mathbf{V}$ matrix for each cluster
  - In what follows, $\mathbf{V}$ will be based on the ***previous random slope model:***

$$y_{pc} = \gamma_{00} + \gamma_{01}(CMx_c) + \gamma_{10}(WCx_{pc}) + U_{0c} + U_{1c}(WCx_{pc}) + e_{pc}$$

End Goals of Maximum Likelihood Estimation

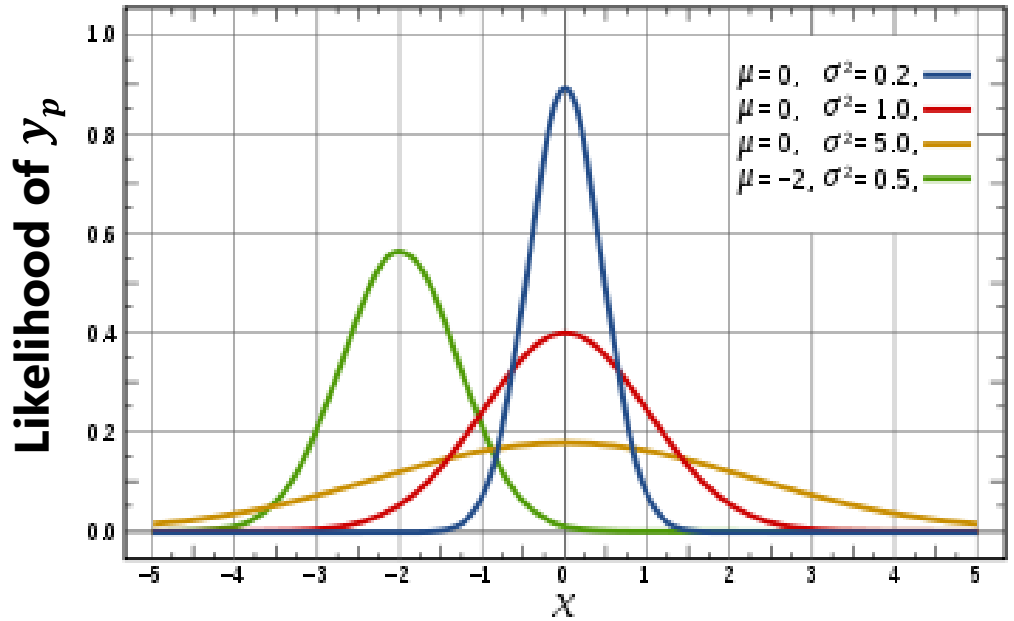
1. Obtain "most likely" values for each unknown model parameter (random effects variances and covariances, residual variances and covariances, which then are used to calculate the fixed effects) → **the estimates**

2. Obtain an index as to how likely each parameter value is (i.e., "really likely" or pretty much just a guess?) → **the standard error (SE) of the estimates**

3. Obtain an index as to how well the model we've specified actually describes the data → **the model fit indices**

How does all this happen? The magic of multivariate normal…
(but let's start with univariate normal first)

Univariate Normal Probability Distribution Function



Univariate Normal PDF (two ways):

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

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

- This PDF tells us how **likely** (i.e., **tall**) any value of y_p is given two things:
  - Predicted value $\hat{y}_p$ → μ on the left
  - Residual variance σ_e^2 → σ^2 on the left

- We can see this work using the NORMDIST function in excel!
  - Easiest for **empty** single-level model:

$$y_p = \beta_0 + e_p$$

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

From Univariate to Multivariate Normal: Joint Height for All $L1n$ Outcomes for Cluster c

$$\text{Univariate Normal PDF: } f(y_i) = \left(2\pi\sigma_e^2\right)^{-1/2} * \exp\left[-\frac{1}{2} * \left(y_i - y_i\right)\left(\sigma_e^2\right)^{-1}\left(y_i - y_i\right)\right]$$

$$\text{Multivariate Normal PDF: } f(\mathbf{Y}_c) = (2\pi)^{-n/2} * |\mathbf{V}_c|^{-1/2} * \exp\left[-\frac{1}{2} * \left(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\gamma}\right)^T\left(\mathbf{V}_c\right)^{-1}\left(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\gamma}\right)\right]$$

- In our example random slope model, three are **three fixed effects** (in $\boldsymbol{\gamma}$) that predict the $\mathbf{Y}_c$ outcomes: intercept $\boldsymbol{\gamma_{00}}$, L2 slope $\boldsymbol{\gamma_{01}}$, and L1 slope $\boldsymbol{\gamma_{10}}$
- Model also gives us $\mathbf{V}_c$ → the model-predicted marginal variance and covariance matrix across clusters, taking into account their $\boldsymbol{WCx_{pc}}$ values
- Uses $|\mathbf{V}_c|$ = determinant of $\mathbf{V}_c$ = summary of *non-redundant* info in $\mathbf{V}_c$

- $(\mathbf{V}_c)^{-1}$ → matrix inverse → analogous to dividing (so can't be 0 or negative)
  - $(\mathbf{V}_c)^{-1}$ must be "positive definite", which in practice means no 0 random variances or covariances that cause out-of-bound correlations between random effects
  - Otherwise, program uses "generalized inverse" → questionable results

Now Try Some Possible Answers...

(e.g., for the 4 variance model parameters in example random slope model)

- Plug $\mathbf{V}_c$ predictions into log-likelihood function, sum over clusters:

$$L = \prod_{c=1}^{L2n} \left\{ (2\pi)^{-n/2} * |\mathbf{V}_c|^{-1/2} * \exp\left[-\frac{1}{2} (\mathbf{Y}_c - \mathbf{X}_c \boldsymbol{\gamma})^T (\mathbf{V}_c)^{-1} (\mathbf{Y}_c - \mathbf{X}_c \boldsymbol{\gamma}) \right] \right\}$$

$$LL = \sum_{c=1}^{L2n} \left\{ \left[-\frac{n}{2} \log(2\pi) \right] + \left[-\frac{1}{2} \log|\mathbf{V}_c| \right] + \left[-\frac{1}{2} (\mathbf{Y}_c - \mathbf{X}_c \boldsymbol{\gamma})^T (\mathbf{V}_c)^{-1} (\mathbf{Y}_c - \mathbf{X}_c \boldsymbol{\gamma}) \right] \right\}$$

- Try one set of possible parameter values to build $\mathbf{V}_c$, compute LL
- Try another possible set to build $\mathbf{V}_c$, compute LL….
  - Different algorithms are used to decide which values to try given that each parameter has its own distribution → like an uncharted mountain
  - Calculus helps the program scale this multidimensional mountain
    - At the top, all first partial derivatives (linear slopes at that point) ≈ 0
    - Positive first partial derivative? Too *low*, try again. Negative? Too *high*, try again.
    - Matrix of partial first derivatives = "score function" = "gradient"

End Goals 1 and 2: Model Estimates and SEs

- Process terminates (the model "converges") when the next set of tried values for $\mathbf{V_c}$ don't improve the LL very much…
  - The values for "how much" is "not very much" are known as "convergence criteria"
  - Those are the values for the parameters that, relative to the other possible values tried, are "most likely" → the variance estimates

- But we need to know how trustworthy all the estimates are…
  - Precision is indexed by the steepness of the multidimensional mountain, where steepness → more negative partial second derivatives
  - Matrix of partial second derivatives = "Hessian matrix"
  - Hessian matrix * −1 = "information matrix"
  - So steeper function = more information = more precision = smaller SE

$$\text{Each parameter SE} = \frac{1}{\sqrt{\text{information}}}$$

What about the Fixed Effects?

- Likelihood mountain does NOT include fixed effects as additional search dimensions (only variances and covariances that make $\mathbf{V_c}$)
- **Fixed effects are computed*** ** given the parameters that build $\mathbf{V_c}$:

$$\boldsymbol{\gamma} = \left\{ \sum_{c=1}^{L2n} \left(\mathbf{X}_c^T \mathbf{V}_c^{-1} \mathbf{X}_c \right) \right\}^{-1} \sum_{c=1}^{L2n} \left(\mathbf{X}_c^T \mathbf{V}_c^{-1} \mathbf{Y}_c \right), \qquad \mathrm{Cov}(\boldsymbol{\gamma}) = \left\{ \sum_{c=1}^{L2n} \left(\mathbf{X}_c^T \mathbf{V}_c^{-1} \mathbf{X}_c \right) \right\}^{-1}$$

All we need is $\mathbf{V_c}$ and the data: $\mathbf{X_c}$, $\mathbf{Y_c}$

γ = fixed effect estimates

$\mathrm{Cov}(\gamma)$ = γ sampling variance (SQRT of diagonal = SE)

- This is actually what happens in regular regression (GLM), too:

$$\text{GLM matrix solution: } \boldsymbol{\beta} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \left(\mathbf{X}^T \mathbf{Y} \right), \qquad \mathrm{Cov}(\boldsymbol{\beta}) = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \sigma_e^2$$

$$\text{GLM scalar solution: } \beta = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \quad \mathrm{Cov}(\beta) = \frac{\sigma_e^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

- **Implication: fixed effects don't cause estimation problems**… (***at least in *general* multilevel models with normal residuals)

What about ML vs. REML?

- **REML** estimates of random effects variances and covariances are **unbiased** because they account for the uncertainty that results from simultaneously also estimating fixed effects (whereas ML estimates do not, so they are too small)

- What does this mean? Remember "population" vs. "sample" formulas for computing variance?

$$\text{Population: } \sigma_e^2 = \frac{\sum_{p=1}^{N}(y_p - \mu)^2}{N} \qquad \text{Sample: } \sigma_e^2 = \frac{\sum_{p=1}^{N}(y_p - \bar{y})^2}{N-1}$$

  - $N - 1$ is used because the mean had to be estimated from the data (i.e., the mean is the fixed intercept in an empty model)…

- Same idea: ML estimates of random effects variances will be downwardly biased by a factor of $(L2n - k) \,/\, L2n$, where k = #fixed effects… it just looks way more complicated

What about ML vs. REML? (N = # obs)

$$\text{ML:} \quad LL = \left[-\frac{N-0}{2}\log(2\pi)\right] + \left[-\frac{1}{2}\sum_{c=1}^{L2n}\log\left|\mathbf{V}_c\right|\right] + \left[-\frac{1}{2}\sum_{c=1}^{L2n}\left(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\gamma}\right)^T \mathbf{V}_c^{-1}\left(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\gamma}\right)\right]$$

$$\text{REML:} \quad LL = \left[-\frac{N-k}{2}\log(2\pi)\right] + \left[-\frac{1}{2}\sum_{c=1}^{L2n}\log\left|\mathbf{V}_c\right|\right] + \left[-\frac{1}{2}\sum_{c=1}^{L2n}\left(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\gamma}\right)^T \mathbf{V}_c^{-1}\left(\mathbf{Y}_g - \mathbf{X}_c\boldsymbol{\gamma}\right)\right]$$

$$+\left[-\frac{1}{2}\log\left|\sum_{c=1}^{L2n}\mathbf{X}_c^T\mathbf{V}_c^{-1}\mathbf{X}_c\right|\right]$$

$$\text{where:} \quad \left[-\frac{1}{2}\log\left|\sum_{c=1}^{L2n}\mathbf{X}_c^T\mathbf{V}_c^{-1}\mathbf{X}_c\right|\right] = \left[\frac{1}{2}\log\left|\left(\sum_{c=1}^{L2n}\mathbf{X}_c^T\mathbf{V}_c^{-1}\mathbf{X}_c\right)^{-1}\right|\right] = \left[\frac{1}{2}\log\left|\text{Cov}(\boldsymbol{\gamma})\right|\right]$$

- Extra part in REML is the sampling variance of the fixed effects…
  it is added back in to account for uncertainty in estimating fixed effects
- **REML maximizes the log-likelihood of the residuals specifically, so the log-likelihoods of models with different fixed effects are not on the same scale and are not comparable**
  - This is why you can't do $-2\Delta LL$ tests in REML when the models to be compared have different fixed effects → the model residuals will be defined differently

End Goal #3: How well do the model predictions match the data?

- End up with "best" LL from predicting $\mathbf{V_c}$ → so how good is it?
- Absolute model fit assessment is only possible when the $\mathbf{V_c}$ matrix is organized the same for all L2 units and there are no random slopes
  - If items are treated as fixed, we can get absolute fit (as in CFA and SEM) → χ^2 test is based on match between actual and predicted variance–covariance matrix
  - No absolute fit provided by default in univariate MLM programs (or in SEM or multilevel SEM when using random slopes), as a "saturated model" as the answer key of cluster dependency is not really possible

- Relative model fit is given as $\boldsymbol{LL}$ (bigger is better) or $\boldsymbol{-2LL}$ (smaller is better)
  - $-2 *$ needed to conduct "likelihood ratio" or "deviance difference" tests
  - Information criteria use $\boldsymbol{-2LL}$, in which smaller is always better:
    - **AIC** = Akaike IC = $-2LL$ + 2 *(#parameters)
    - **BIC** = Bayesian IC = $-2LL$ + $\log(N)$*(#parameters)

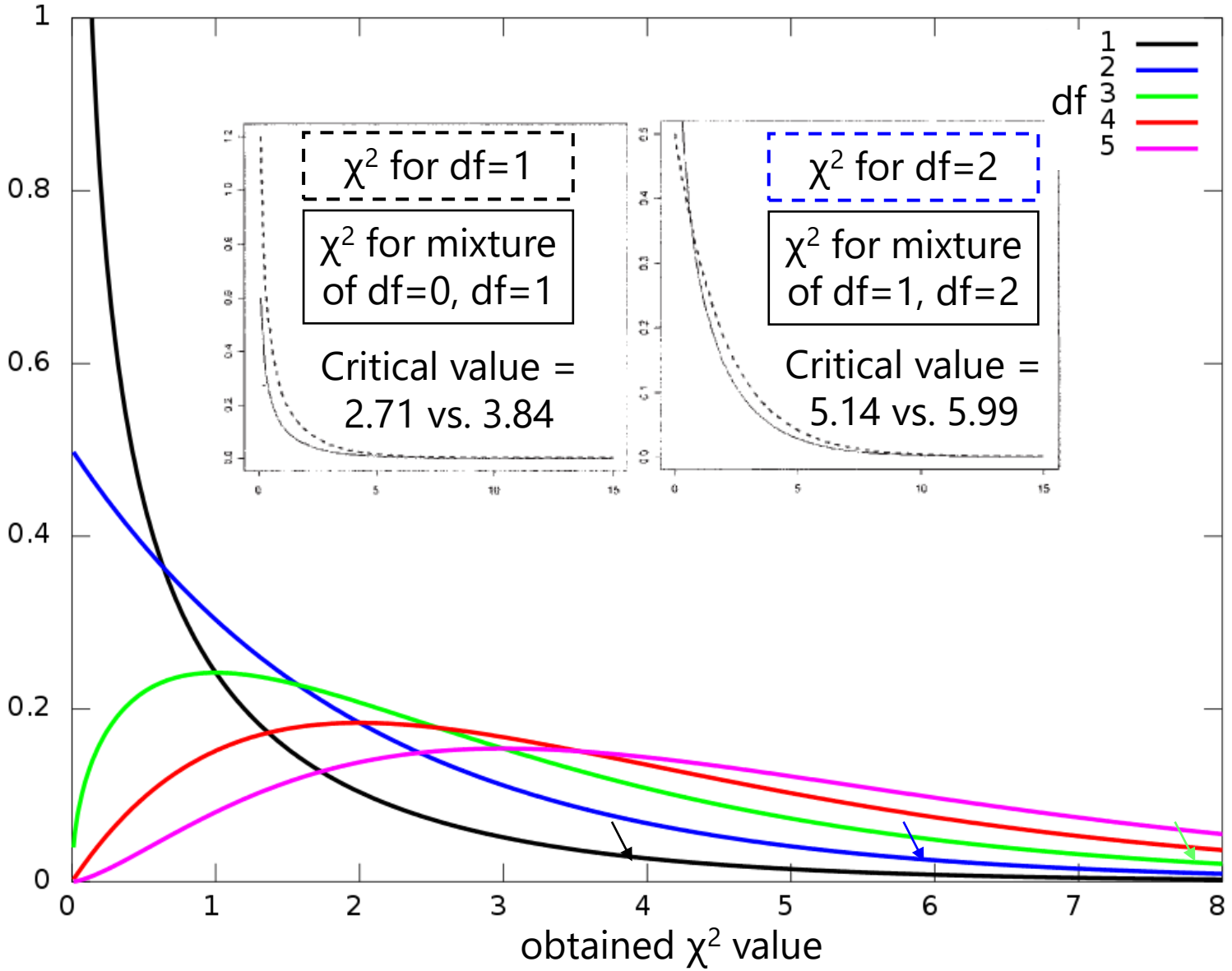
What about testing variances > 0?

- $-2\Delta LL$ between two nested models is distributed as χ^2 only when added parameters do not have a boundary (like 0 or 1)
  - Is ok for fixed effects using ML (could be any positive or negative value)
  - Is NOT ok for ML or REML tests of random variances (must be > 0)
  - Is ok for ML or REML tests of heterogeneous variances and covariances (because extra parameters can be phrased as unbounded deviations)

- When testing the addition of parameters with a boundary, $-2\Delta LL$ will follow a **mixture** of χ^2 distributions instead
  - e.g., when adding random intercept variance (test > 0?)
    - When estimated as positive, will follow χ^2 with df=1
    - When estimated as negative… can't happen, will follow χ^2 with df=0 (= 0)

- End result: **$-2\Delta LL$ will be too conservative in boundary cases** (which I can live with)

χ^2 Distributions small pictures from Stoel et al., 2006



Mixture Chi-Square Critical Values

	Significance Level				
df (q)	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
0 vs. 1	1.64	2.71	3.84	5.41	6.63
1 vs. 2	3.81	5.14	6.48	8.27	9.63
2 vs. 3	5.53	7.05	8.54	10.50	11.97
3 vs. 4	7.09	8.76	10.38	12.48	14.04
4 vs. 5	8.57	10.37	12.10	14.32	15.97
5 vs. 6	10.00	11.91	13.74	16.07	17.79
6 vs. 7	11.38	13.40	15.32	17.76	19.54
7 vs. 8	12.74	14.85	16.86	19.38	21.23
8 vs. 9	14.07	16.27	18.35	20.97	22.88
9 vs. 10	15.38	17.67	19.82	22.52	24.49
10 vs. 11	16.67	19.04	21.27	24.05	26.07

This may work ok if only one new parameter is bounded … for example:

+ Random Intercept
df=1: 2.71 vs. 3.84

+ Random Slope #1
df=2: 5.14 vs. 5.99

+ Random Slope #2
df=3: 7.05 vs. 7.82

Critical values such that the right-hand tail probability = 0.5 x Pr ($\chi^2_q > c$) + 0.5 x Pr ($\chi^2_{q+1} > c$)

Source: Appendix C (p. 484) from Fitzmaurice, Laird, & Ware (2004). *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley

Solutions for Boundary Problems when using $-2\Delta LL$ (LRTs)

- If adding random intercept variance only, use $p < .10$; $\chi^2(df=1) > 2.71$
  - Because χ^2 (df=0) = 0, can just cut p-value in half to get correct p-value

- If adding ONE random slope variance (and covariance with random intercept), can use mixture p-value from $\chi^2(df=1)$ and $\chi^2(df=2)$

$$\text{Mixture } p\text{-value} = 0.5 * \text{prob}(\chi^2_1 > -2\Delta\text{LL}) + 0.5 * \text{prob}(\chi^2_2 > -2\Delta\text{LL})$$

so critical χ^2 = 5.14, not 5.99

- However—using a 50/50 mixture assumes a diagonal information matrix for the random effects variances (i.e., it assumes each value is arrived at independently, which is not likely to be true)
- Two options for more complex cases:
  - Simulate data to determine actual mixture for calculating p-value
  - Accept that -2ΔLL is conservative in these cases, and use it anyway
    → I use ~2 to acknowledge this: e.g., -2ΔLL(df~2) > 5.99, $p < .05$

Predicted Level-2 $\mathbf{U_c}$ Random Effects
(aka Empirical Bayes or BLUP Estimates)

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

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

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

What about the $\mathbf{U_c}$ random effects?

- Level-2 unit $\mathbf{U_c}$ values are NOT estimated in the likelihood function
  - $\mathbf{G}$ matrix variances and covariances are sufficient statistics for the estimation process assuming multivariate normality of $\mathbf{U_c}$ values
  - Level-2 $\mathbf{U_c}$ random effects are **predicted** [blups <- ranef(model, condVar=TRUE) in R] as: $\mathbf{U_c} = \mathbf{G_c}\mathbf{Z_c^T}\mathbf{V_c^{-1}}(\mathbf{Y_c} - \mathbf{X_c}\boldsymbol{\gamma})$
    - Which then create cluster estimates as: $\boldsymbol{\beta_{0c}} = \boldsymbol{\gamma_{00}} + \boldsymbol{U_{0c}}$ and $\boldsymbol{\beta_{1c}} = \boldsymbol{\gamma_{10}} + \boldsymbol{U_{1c}}$

- What isn't obvious: the composite $\boldsymbol{\beta}_c$ values are weighted combos of the fixed effects ($\boldsymbol{\gamma}$) and their level-2 OLS estimates ($\boldsymbol{\beta}_{\mathbf{OLS_c}}$) :

$$\text{Random Effects: } \boldsymbol{\beta}_c = \mathbf{W}_c\boldsymbol{\beta}_{OLS_c} + (\mathbf{I} - \mathbf{W}_c)\boldsymbol{\gamma} \qquad \text{where: } \mathbf{W}_c = \mathbf{G}_c\left[\mathbf{G}_c + \sigma_e^2\left(\mathbf{Z}_c^T\mathbf{Z}_c\right)^{-1}\right]^{-1}$$

  - The more "true" variation in intercepts and slopes in the outcome (in $\mathbf{G}$), the more the $\boldsymbol{\beta}_c$ estimates are based on level-2 unit OLS estimates
  - The more "unexplained" residual variation around the level-2 slopes (in $\mathbf{R}$), the more the fixed effects are heavily weighted instead
    - = **SHRINKAGE** (more so for clusters with fewer persons, too)

What about the $\mathbf{U}_c$ random effects?

- Point of the story: $\mathbf{U}_c$ values are NOT single scores!
  - They are the mean of a distribution of possible values for each person (i.e., as given by the SE for each $\mathbf{U}_c$)
  - These "best estimates" of the $\mathbf{U}_c$ values are shrunken anyway

- Good news: you don't need those $\mathbf{U}_c$ values in the first place!
  - Goal of MLM is to estimate and predict the variance of the $\mathbf{U}_c$ values (in $\mathbf{G}$) with cluster-level characteristics directly in the model
  - If you want your $\mathbf{U}_c$ values to be predictors instead, then you need to estimate your model using multivariate MLM ("M-SEM")
  - You could use the predicted $\mathbf{U}_c$ values to examine potential violations of model assumptions, though…

Estimation: The Grand Finale

- Estimation in MLM is all about finding the most likely estimates for the random effects variances and covariances
  - The more of them there are, the harder it is to find them (the more dimensions of the likelihood mountain there are to scale)
  - "Non-positive-definite" $\mathbf{G}$ matrix means "broken model" (usually because a variance went to 0 or a correlation went out of bounds)
  - Fixed effects are solved for given $\mathbf{V}$ in general MLMs, so they rarely cause estimation problems
  - Individual random effects are not model parameters, but can be predicted after-the-fact (but try never to use these as data)

- Estimation comes in two flavors:
  - ML → maximize the data; use $-2\Delta LL$ to compare **any** nested models
  - REML → maximize the residuals; use $-2\Delta LL$ to compare models that **differ in their model for the variance ONLY**