CLUSTERED DATA: ARE MULTILEVEL MODELS REALLY NECESSARY?

Dan McNeish & Laura Stapleton
University of Maryland, College Park
OUTLINE

- Overview of Clustered Data and Multilevel Models
- Proliferation of Multilevel Models
- Details and Differences in Methods for Clustered Data
- Application to a Reading Intervention Study
- Implications
OVERVIEW OF CLUSTERED DATA AND MULTILEVEL MODELS
Clustered data are more often the rule than the exception in education:

- Students clustered within classrooms
- Classrooms clustered within schools
- Schools clustered within districts

If the analysis doesn’t account for the clustered structure, the model estimates will not be trustworthy.
ACCOUNTING FOR CLUSTERING

- Traditional statistical models assume that individuals are not systematically related to one another
  - Not the case if data are clustered
  - Kids in the same school are more like each other than kids in other schools, etc.

![Graph showing relationship between College GPA and HS GPA]
ACCOUNTING FOR CLUSTERING

- Traditional statistical models assume that individuals are not systematically related to one another
  - Not the case if data are clustered
  - Kids in the same school are more like each other than kids in other schools, etc.
ACCOUNTING FOR CLUSTERING II

The Issue:

- Dependencies within clusters reduce the amount of unique information available for a given sample size
  - If I know Timmy’s reading score and know that Tommy is in the same school as Timmy, I already know a little bit about Tommy’s score too
  - The unique contribution of each student is reduced because some information is shared by knowing which school they are in
WHAT TRADITIONAL MODEL ASSUMES

TIMMY

TOMMY
What Happens With Clustering

TIMMY

SCHOOL

TOMMY
ACCOUNTING FOR CLUSTERING III

- The total sample size appears in the denominator of formulas for calculating precision in single-level models
  - Using the total sample size assumes that each individual contributes unique information
  - This is not the case with clustering, so the denominator will be too large

- In English – ignoring the clustering makes the model appear more precise than it is
The predominant method to account for clustering in education research is through multilevel models (MLMs).

MLMs take the clustering into account primarily by estimating random effects for each cluster.

- Random effects allow the prediction model for each cluster to be different from the prediction of the overall sample.
RANDOM EFFECTS

No Random Effects

Random Intercepts

Random Intercepts and Random Slopes

Random Slopes
Overall regression line for 5 clusters of data (ignoring clustering)

5 cluster-specific regression lines with overall regression in black.
MLMs use an overall prediction line but allow the intercept and slope of the line to change for each cluster. The change in the cluster-specific line from the overall line can be quantified and helps to estimate precision more appropriately.

5 cluster-specific regression lines with overall regression in black.
Proliferation of Multilevel Models
DOMINANCE OF MLMS

- MLMs are overwhelming popular to accommodate clustered data in the behavioral sciences

- Found that from 2006 to 2011, 94% of published studies accounted for clustering using MLMs
  - Found only 14 studies (6%) that used design-based methods (DBMs)
A Closer Look at Charter Schools Using Hierarchical Linear Modeling

Measuring School Effects With Hierarchical Linear Modeling: Data Handling and Modeling Issues

Eugene P. Adcock, Ph.D., Prince George's County Public Schools, Maryland

Dropping Out of Middle School: A Multilevel Analysis of Students and Schools
American Educational Research Journal

Using Multi-level Analyses to Study the Effectiveness of Science Curriculum Materials

Vasuki Rethinam
Montgomery County Public Schools, USA
Curtis Pyke and Sharon Lynch
Graduate School of Education and Human Development, The George Washington University, USA

RESEARCH REPORT

Is there a “school effect” on pupil outcomes? A review of multilevel studies

A Multilevel Analysis of Students and Schools on High School Graduation Exam: A Case of Maryland
What is Taught in Graduate School?

- Previous study suggested that MLMs are more relevant to behavioral scientists because they are used most often.

- There is some evidence that they are used because of tradition and lack of exposure to alternatives.

- Reviewed 51 graduate school syllabi for courses specifically focused on methods for clustered data:
  - e.g., Longitudinal, complex survey data, cross-sectionally clustered
  - Convenience sample, so generalizability not fully warranted
  - Spanned 8 academic disciplines
    - Economics notably absent because material is embedded within broader econometrics sequences.
### What is Taught in Graduate School?

<table>
<thead>
<tr>
<th>Discipline</th>
<th>DBMs Only</th>
<th>MLM Only</th>
<th>MLM and DBMs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biostatistics</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Criminology</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Education</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Political Sci.</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Psychology</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Public Health</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Sociology</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Statistics</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

- **DBMs** = Design-Based Methods (GEE, Cluster Robust Errors, Survey Methods)
- **MLMs** = Multilevel Models
## What is Taught in Graduate School?

<table>
<thead>
<tr>
<th>Discipline</th>
<th>DBMs Only</th>
<th>MLM Only</th>
<th>MLM and DBMs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biostatistics</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Criminology</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Education</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Political Sci.</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Psychology</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Public Health</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Sociology</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Statistics</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

- DBMs = Design-Based Methods (GEE, Cluster Robust Errors, Survey Methods)
- MLMs = Multilevel Models
**What is Taught in Graduate School?**

<table>
<thead>
<tr>
<th>Discipline</th>
<th>DBMs Only</th>
<th>MLM Only</th>
<th>MLM and DBMs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biostatistics</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Criminology</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td>0</td>
<td><strong>8</strong></td>
<td>2</td>
<td><strong>10</strong></td>
</tr>
<tr>
<td>Political Sci.</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Psychology</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Public Health</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Sociology</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Statistics</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

- Most disciplines provide coverage of both DBMs and MLMs
- However, Education and Psychology very often exclusively teach MLMs
  - Not surprising to find that 94% of psychology studies use MLMs
Details and Differences in Methods for Clustered Data
RANDOM EFFECTS
IN MLMs

The ability to quantify how much each cluster varies from the overall prediction sounds highly advantageous.

Caveat – Researchers must explicitly select which random effects to include and how they related to each other.
- Do schools start at about the same values or do they differ?
- Do students grow at similar rates or is there a lot of variation?
- Do student characteristics have the same effect in all schools?
- Does where students start affect how they grow over time?
These decisions are not easy and, because of the nature of random effects, there are not well-developed statistical methods to determine if one selected correctly.

Worse yet, incorrect selections will introduce bias into both the values of the estimates and estimates of their precision.

- All relevant random effects must be included
- All irrelevant random effects must be excluded
- The structure (covariance matrix) of the random effects must be correct.
MLM
ASSUMPTIONS

- MLMs also make several other assumptions, including:

1. All relevant predictors are included
2. All relevant random effects are included
3. The covariance structure of the within-cluster residuals is correct
4. The covariance structure of the random effects is correct
5. The within-cluster residuals follow a (multivariate) normal distribution
6. The random effects follow a (multivariate) normal distribution
7. The random effects are not correlated with any predictors in the model
8. The sample size is sufficiently large for asymptotic inferences at all levels
Rather than explicitly model the clustering mechanism, DBMs treat the clustering mechanism as a nuisance:
- Aim to account for clustering rather than explain it

DBMs treat the model as if it were a single level model, then apply statistical corrections to reflect that the data are clustered.

No random effects are included so models require fewer assumptions, are simpler to specify, and have more straightforward interpretations.
**DBM ASSUMPTIONS**

- Because DBMs do not require random effects or explicit modeling of the covariance structure(s), far fewer assumptions are required

---

**DBM Assumptions**

1. All relevant predictors are included
2. The sample size is sufficiently large for asymptotic inferences at the cluster level
3. Observations between clusters are not related
4. The working correlation matrix is "reasonably close" to the population structure
In a single-level model, the regression coefficients are generally interpreted as

“For a one-unit change in X, Y is predicted to change by $\beta$ units, holding all other predictors constant”

- Referred to as “population-averaged” interpretation because every individual in the data shares the same prediction model
- With DBMs, this interpretation is preserved and the coefficients have a population-averaged interpretation
DBM coefficients are interpreted based on the whole sample.
However, this is **not** how coefficients are interpreted with MLMs

- MLMs have a cluster-specific interpretation

Because of the random effects, people with identical data values but who are in different clusters will have different predicted values.

For a one-unit change in $X$, $Y$ is predicted to change by $\beta$ units, holding all other variables **and the random effects** constant
DBM coefficients are interpreted based on the whole sample.

MLM coefficients are interpreted based on the cluster.

One unit change in predictor results in a change in the outcome predicted by the model.
The difference between the cluster-specific and population-averaged interpretations is minimal with continuous outcomes*

With discrete outcomes (variables with a finite number of categories), the differences between population-averaged and cluster-specific coefficients can be very noticeable **

*Fitzmaurice, Laird, & Ware (2011)
** Carlin, Wolfe, Brown, & Gelman (2001) Biostatistics
APPLICATION TO A READING INTERVENTION STUDY
Overview of Research Question

- Data come from a study that was part of an IES grant funded to Dr. Rebecca Silverman

- The overall aim of the study to evaluate the efficacy of a reading intervention to assess whether the growth in reading vocabulary and comprehension was greater for students receiving the treatment compared to students in a control group.

- Will only discuss one of the models for one of the grades
SAMPLE & METHOD

- 203 kindergarten students
  - 53% of students were ELL

- Students were sampled from 12 different urban, Type-I classified classrooms
  - 6 classrooms assigned to the treatment group
  - 6 classrooms assigned to the control group

- Students’ vocabulary was assessed with a pre and post-test
  - Data that will be shown is for the Peabody Picture Vocabulary Test (PPVT)
PPVT Analysis

- Interest is whether PPVT scores are higher for students exposed to the treatment, controlling for pre-test scores and ELL status
  - ICC is 0.21, DEFT is 2.21 so clustering within classrooms is meaningful

- Will demonstrate the difference in running this analysis with a DBM and a MLM
1. All relevant predictors are included
2. The sample size is sufficiently large for asymptotic inferences at the cluster level
3. Observations between clusters are not related
4. The working correlation matrix is "reasonably close" to the population structure

Assumptions are pretty simple to check

(1) the model was theoretically determined
(2) sample size was rather small was a correction was used
(3) there is no meaningful higher level of the hierarchy
(4) the ICC is less than 0.30, so the choice is straightforward
Design-Based Method Results

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>127.43</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>4.88</td>
<td>.04</td>
</tr>
<tr>
<td>ELL</td>
<td>0.04</td>
<td>.99</td>
</tr>
<tr>
<td>Treatment × ELL</td>
<td>-2.08</td>
<td>.38</td>
</tr>
<tr>
<td>PPVT GSV Pre-Test (mean-centered)</td>
<td>0.77</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Residual Variance</td>
<td>71.89</td>
<td></td>
</tr>
<tr>
<td>Model R²</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

- Treatment is significant at 0.05 level
- Treatment does not appear to be different for ELL or non-ELL students
- Scores for ELL students are comparable to non-ELL students
- No concerns about assumptions
After predictor variables are selected, the first step is to determine which random effects to include.
ASSUMPTIONS

MLM Assumptions

1. All relevant predictors are included
2. All relevant random effects are included
3. The covariance structure of the within-cluster residuals is correct
4. The covariance structure of the random effects is correct
5. The within-cluster residuals follow a (multivariate) normal distribution
6. The random effects follow a (multivariate) normal distribution
7. The random effects are not correlated with any predictors in the model
8. The sample size is sufficiently large for asymptotic inferences at all levels

Won’t discuss all of these, but keep in mind that there are a lot of assumptions to check
Assumptions 2, 5, 6, 7 for a model with only 4
### MLM

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>127.16</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>4.53</td>
<td>.07</td>
</tr>
<tr>
<td>ELL</td>
<td>0.49</td>
<td>.83</td>
</tr>
<tr>
<td>Treatment $\times$ ELL</td>
<td>-1.65</td>
<td>.54</td>
</tr>
<tr>
<td>PPVT GSV Pre-Test (mean-centered)</td>
<td>0.77</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept Variance</td>
<td>6.11</td>
<td></td>
</tr>
<tr>
<td>Within-Cluster Variance</td>
<td>66.32</td>
<td></td>
</tr>
<tr>
<td>Total Residual Variance</td>
<td>72.43</td>
<td></td>
</tr>
</tbody>
</table>

### DBM

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>127.43</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>4.88</td>
<td>.04</td>
</tr>
<tr>
<td>ELL</td>
<td>0.04</td>
<td>.99</td>
</tr>
<tr>
<td>Treatment $\times$ ELL</td>
<td>-2.08</td>
<td>.38</td>
</tr>
<tr>
<td>PPVT GSV Pre-Test (mean-centered)</td>
<td>0.77</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual Variance</td>
<td>71.89</td>
<td></td>
</tr>
<tr>
<td>Effect</td>
<td>MLM</td>
<td>DBM</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Intercept</td>
<td>127.16</td>
<td>127.43</td>
</tr>
<tr>
<td>Treatment</td>
<td>4.53</td>
<td>4.88</td>
</tr>
<tr>
<td>ELL</td>
<td>0.49</td>
<td>0.04</td>
</tr>
<tr>
<td>Treatment × ELL</td>
<td>-1.65</td>
<td>-2.08</td>
</tr>
<tr>
<td>PPVT GSV Pre-Test (mean-centered)</td>
<td>0.77</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MLM</th>
<th>DBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept Variance</td>
<td>6.11</td>
<td>---</td>
</tr>
<tr>
<td>Within-Cluster Variance</td>
<td>66.32</td>
<td>---</td>
</tr>
<tr>
<td>Total Residual Variance</td>
<td>72.43</td>
<td>71.89</td>
</tr>
</tbody>
</table>
IMPLICATIONS
MLM/DBM COMPARISON

- Main point is that the DBM analysis was much simpler, involved far fewer assumptions, and, in this case, allowed calculating an OLS-type $R^2$
  - With these data, the cluster-specific and population-averaged interpretation are interchangable

- The $p$-value changing was a coincidence in these particular data
  - Not a typical result
  - Strong assumptions of MLMs could have contributed to this though*

MLM/DBM Comparison II

- DBMs are also much less sensitive to researcher specifications
  - Fewer decisions need to be made explicitly by the researchers
  - These decisions are often based on sub-optimal statistical theory
  - Results for the same data with the same predictors can vary based on who is conducting the analysis

- Quantifying the random effects sounds great in theory
  - In realistic models with many predictors, it can become a very complex task that can have discernible effects on the results
MLDS CENTER’S GOALS

- Internal discussion has been devoted to making MLDS’ findings broadly interpretable to the public.

- DBMs make fewer assumptions, are interpreted the same as far simpler statistical models, and give simpler interpretations:
  - The model building and assumption made by MLMs alone can be difficult to understand.
  - MLM coefficient interpretation is also far less intuitive.
As a general guideline, researchers strive for the simplest statistical method that will accommodate their data and handle their research question

- Much criticism is directed toward using models that are too simple and don’t meet assumptions/data requirements

With clustered data in education, the exact opposite occurs in practice

- MLMs are almost universally used to account for clustering
- Cluster-specific information is not always of interest
- Researchers are making strong assumptions and tackling difficult problems to get random effects they don’t need or even care about
This talk in one sentence

Use methods that address the needs of your research questions.
CONCLUDING RECOMMENDATIONS

- **MLMs are useful if:**
  - Directly modeling the covariance structures/ explaining the clustering mechanism is relevant to the research question
  - Inferences about specific clusters are desired
  - The partitioning of variance is substantively important

- **DBMs are useful if:**
  - The clustering is just an aspect of the data to accommodate, especially with discrete outcomes
  - The correlation within-clusters is not a modeling interest
  - Interest does not pertain to specific clusters
Thank You!

dmcneish@umd.edu