Data Connections: Overview & Findings

Dr. Walt Stroup, Dept of Statistics, UNL
Dr. Jennifer Green, Dept of Mathematical Sciences, Montana State University
Dr. Wendy Smith, Center for Science, Mathematics & Computer Education, UNL
Dr. Leslie Lukin, Lincoln Public Schools
Pamela Fellers, Dept of Statistics, UNL
Data Connections

• $1.2 million NSF RETA (Research and Evaluation Technical Assistance), 2011-2014

• Partnership between University of Nebraska-Lincoln (UNL) and Lincoln Public Schools (LPS)

• Focused on developing and sharing statistical models to better estimate value-added teacher effects on student learning
  – Coherent picture of teaching and learning
  – Characterize program effectiveness
Time Line I

- 2004: Math in the Middle; NebraskaMATH; Statistics Department GTA Training; collaboration with Math, TLTE, English
- 2009: At NSF-MSP conference, Dept of Ed in new Obama admin speaks of using data to identify successful MSPs to scale up
- 2009: problem – then existing statistical methods to do so were underdeveloped, controversial, poorly understood
- Much **data-free ideology**
Time Line II

• 2011: received RETA grant
• Back to 1980s
  – value added models (VAMs)
  – origins: W. L. Sanders in Knoxville, TN
  – UTK & Knox County schools
• 1990s to present
  – increased use of VAMs in education
  – many states mandate their use for evaluation
  – close VAM/No Child Left Behind/ Race to the Top connection
Time Line III

• Center for Science, Mathematics and Computer Education
  – Established with NSF funds in the 1990s
• LPS-UNL Professional Development Partnership
  – CCLI grant from NSF to revise mathematics education of preservice elementary teachers
• Math in the Middle Institute Partnership - 2004-2011
  – $5.9 million Math Science Partnership from NSF (grades 5-8)
• NebraskaMATH - 2009-2014
  – $9.2 million from NSF as a Targeted Math Science Partnership
  – Main research focus: Primarily Math (K-3)
Math in the Middle

• Funded by NSF ($5 million + $900,000 supplement)
• Intense master’s degree program
  – 12-course (2 year) masters program that educates and supports teams of outstanding middle level math teachers who will become intellectual leaders in their schools, districts, and ESUs.
  – Compact summer courses + blended AY courses
• Research initiative:
  – Understand how changes in teachers’ mathematics knowledge translate into measurable improvement in student performance
• Six cohorts of teachers, beginning in summer 2004
NebraskaMATH

• University-led teacher professional development
• P-16 partnership across the state
  – 4 core partner districts, over 100 total districts
• University partnerships (5 depts, 3 colleges)
• PD for K-3, algebra, and novice math teachers
• Goal to improve K-12 mathematics achievement for all students and close large achievement gaps
Primarily Math

- Focuses on strengthening the teaching & learning of mathematics in grades K-3
- Six course, 18-credit hour program leading to a K-3 Mathematics Specialist certificate
  - 3 mathematics courses, 3 pedagogy courses
- Optional 7th course focusing on leadership
- Research focus on cohorts 1-3 + comparison group
  - Recruited in 2009 & 2010 to take courses 2009/10, 2010/11, or 2011/12
  - Annual teacher surveys
  - Subset of classrooms also included student data
Three Threads of RETA

• Impact on Teachers’ Attitudes
• Impact on Teachers’ Math Knowledge
• Impact on Students (including but not limited to Value Added Modeling)
Impact on Attitudes & Beliefs

• Premise: before you look at impact on students you need to know if professional development has an impact on teachers

• Math in the Middle
  – Locally developed instrument (bad psychometric properties)

• Primarily Math
  – Fennema-Sherman Mathematics Attitudes Scales (Fennema & Sherman, 1976)
  – Mathematics Beliefs Scales (Fennema, Carpenter, & Loef, 1990; Caprano, 2001)
Impact on Attitudes & Beliefs

Primarily Math K-3 Teachers’ Attitudes toward Mathematics
Impact on Attitudes & Beliefs

Primarily Math K-3 Teachers’ Beliefs about Mathematics

Cohort 1 Teacher Beliefs Rating

Student-centered Beliefs
Teacher-centered Beliefs
Impact on Attitudes & Beliefs

Primarily Math Teachers’ Beliefs about Mathematics Teaching vs. the Comparison Group
Impact on Teacher Knowledge

- Mathematical Content Knowledge for Teaching Survey (MKT; Hill et al., 2004)
- Separate versions for Elementary (K-6) and Middle Level (4-8)
- Subscales:
  - Number & Operations
  - Patterns, Functions & Algebra
  - Geometry
- 2004 versions did not end up being equated to 2005+ versions
  - Cohorts 1 & 2 could not be compared to later cohorts
Math in the Middle MKT

- Secondary (HS) certification vs. Elementary/Middle/Other Certification shows parallel changes over time

Mathematical Knowledge for Teaching: Math in the Middle Participants

- Elem N&O
- HS N&O
- Elem Alg
- HS Alg
- Elem Geo
- HS Geo

Pre- Post- Followup

Standardized Score

-1 -0.5 0 0.5 1 1.5 2
Primarily Math MKT

Primarily Math K-3 Teachers’ Mathematical Content Knowledge for Teaching

![Graph showing the improvement of Number & Operations, Patterns, Functions, & Algebra, and Geometry over pretest, posttest, and follow-ups.](image-url)
Primarily Math MKT

Primarily Math vs Comparison Group
Mathematical Content Knowledge for Teaching
Impact on Students

• State test data?
  – No state test in mathematics until 2011
  – Prior years: each district created own test & procedures
  – State test data do not tie students to teachers
  – State test validity/reliability?

• Primarily Math impact on K-3 students

• Math in the Middle impact on 5th-8th students
  – Alternative assessment (psychometric fall/spring equating issues) yielded single finding: 5th-8th grade students are very weak at expressing mathematical reasoning in writing
Primarily Math Student Achievement

- Not VAM
- Predictors
  - Teacher-level predictors
    - Mathematical Content Knowledge for Teaching
    - Attitudes towards Learning Mathematics
    - Beliefs about Teaching Mathematics and Student Learning
  - Socio-Economic Status (building level)
- Student-Level Outcomes
  - Change from fall to spring in Math Ability Scores (as measured by the TEMA-3)
  - Not math competence beliefs (too stable)
Analytical Approach: Hierarchical Linear Modeling

\[ M ASD_{ijk} = \beta_{00k} + \beta_{01k} Time_{jk} + \beta_{02k} FMA S_{ijk} + \beta_{03k} SES_{jk} + \sum \beta_{0qk} X_{jk} \]

\[ \beta_{00k} = \gamma_{000} + \gamma_{001} Treatment_k + u_{00k} \]
\[ \beta_{01k} = \gamma_{010} + \gamma_{011} Treatment_k + u_{01k} \]
\[ \beta_{0qk} = \gamma_{0q0} \]
Analytical Approach

Observed and Predicted Mean $MASD$

- $MASD$ vs. Year
- Green line: Observed
- Blue line: Predicted

$MASD$ vs. Year graph showing observed and predicted values.
Results

Observed and Predicted Mean $MASD$

Control

Treatment

$MASD$

Year

Year
Analytical Approach: Hierarchical Linear Modeling

\[ M A S D_{ijk} = \beta_{00k} + \beta_{01k} Time_{jk} + \beta_{02k} FMAS_{ijk} + \beta_{03k} SES_{jk} + \beta_{04k} Num_{jk} + \beta_{05k} Anxiety_{jk} + \beta_{07k} Tbelief_{jk} + r_{ijk} \]

\[ \beta_{00k} = \gamma_{000} + \gamma_{001} Treatment_k + u_{00k} \]
\[ \beta_{01k} = \gamma_{010} + u_{01k} \]
\[ \beta_{02k} = \gamma_{020} \]
\[ \vdots \]
\[ \beta_{07k} = \gamma_{070} \]
### Fixed and Random Effects for Final Model

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>Random Effect</th>
<th>Variance</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.36</td>
<td>0.72</td>
<td>Teacher</td>
<td>19.09</td>
<td>4.37</td>
</tr>
<tr>
<td>SES</td>
<td>0.00</td>
<td>0.01</td>
<td>Teacher*Time</td>
<td>4.76</td>
<td>2.18</td>
</tr>
<tr>
<td>Treatment</td>
<td>1.79</td>
<td>0.98</td>
<td>Residual</td>
<td>69.65</td>
<td>8.34</td>
</tr>
<tr>
<td>FMAS</td>
<td>-0.24</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>0.75</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Num</td>
<td>-0.47</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>1.42</td>
<td>0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tbelief</td>
<td>0.45</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpretation of Findings

• Students in Primarily Math classrooms demonstrated larger than expected growth
  – Initial differences in student growth fall to spring
  – Within classroom variability very large
• Spring variability < Fall variability
• Next steps include more student-level variables (provided by one district, this week); examining differential impact on students who start lowest
What is Value-Added?

Test Scores

Diff between student score & district avg.

Expected growth for student

District avg.

Year

g

g + 1

Ballou, Sanders, & Wright (2004)
What is Value-Added?

Test Scores

Diff between student score & district avg.

Add’tl gain for student
Q: can we call this the “teacher effect”?

District avg.

Year

Ballou, Sanders, & Wright (2004)
What is a Layered Model?

Expected Growth equal to District Avg. Growth

Diff between student score & district avg.

Test Scores

Year

Teacher 1

Teacher 2

District avg.

<table>
<thead>
<tr>
<th>Year</th>
<th>District avg.</th>
<th>Teacher 1</th>
<th>Teacher 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g + 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g + 2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is a Layered Model?

Test Scores

Year

Above district avg. growth

Teacher 1

Equal to district avg. growth

Teacher 2

District avg.

credit to which teacher?

Year

The graph illustrates a layered model in educational data analysis. The x-axis represents the years (g, g+1, g+2) and the y-axis represents the test scores. Each line connects data points for different teachers (Teacher 1, Teacher 2) and years, showing growth patterns. The graph highlights the concept of attributing equal to district average growth and determining the credit to which teacher is responsible.
What is a Layered Model?

- Usual statistical model

\[ \text{score}_{g+1} = \mu + \text{student} + \text{teacher}_1 \]
\[ \text{score}_{g+2} = \mu + \text{student} + \text{teacher}_2 \]

- Layered model

\[ \text{score}_{g+1} = \mu + \text{student} + \text{teacher}_1 \]
\[ \text{score}_{g+2} = \mu + \text{student} + \text{teacher}_1 + \text{teacher}_2 \]
What is a Program Effect?

Test Scores

<table>
<thead>
<tr>
<th>Year</th>
<th>District</th>
<th>Teacher 1</th>
<th>Teacher 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g + 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g + 2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Above district avg. growth

Equal to district avg. growth

program

when?

teacher

District avg.
What is a Program Effect?

- Layered Model with Program Effect
  \[
  \text{score}_{g+1} = \mu + \text{student} + \text{teacher}_{1,P}
  \]
  \[
  \text{score}_{g+2} = \mu + \text{student} + \text{teacher}_{1,P} + \text{teacher}_{2,N}
  \]

- Definition?
  \[
  \text{program effect} = \text{teacher}_{1,P} - \text{teacher}_{1,N}
  \]

- For teachers in the program
  - you need to know their effect **before** as well as **during** the program
  - you need some assurance that their effect is stable
Two Statistical Issues

- Fixed versus Random Model Effects
- Impact of type of effect on how we estimate
  - teacher effect
  - program effect
Types of Model Effects

• Three questions:
  – If multiple studies done independently would all studies use the same *levels* (e.g. in program or not)?
  – Anything special about levels in the study?
  – Do the levels represent a target population?

• Fixed
  – yes
  – yes
  – no

• Random

**Effects in the model**
• Program (P or N)
• Teachers

How do they fit these criteria?
Estimating Model Effects

• Fixed
  – familiar to all
  – compute the mean
• Random
  – they don’t teach this in intro stat
  – key to estimating teacher and program effects
Estimating a Random Effect

- Example: student “mastery”
- Let $M$ denote mastery
- $M$ varies among students
  - mean, denote as $\mu_M$
  - variance, denote as $\sigma_M^2$
- Measure “mastery” by a test, denoted $S$
- $S$ has measurement error
  - mean, denote as $\mu_S$
  - variance, denote as $\sigma_S^2$
Teacher Effect on Mastery

- M varies among students
  - mean, denote as $\mu_M$
  - variance, denote as $\sigma_M^2$
- S has measurement error
  - mean, denote as $\mu_S$
  - variance, denote as $\sigma_S^2$
- Student mastery under teacher T
  - $M+T$
- Teachers in study represent target population
  - mean, denote as $\mu_T$
  - variance, denote as $\sigma_T^2$
Estimating a Random Effect

• We want to estimate teacher effect $T$
• We do so via student mastery $M+T$
• We measure $M+T$ by $S$
• Question: what is the best estimate of $M+T$?
• Hint: it is NOT the test score $S$
• What is it?
  $$E(M+T|S)$$
  - depends on means and variances of $M$, $S$ and $T$
Some Issues Addressed by RETA

• Mixed Model Methodology
  – teacher effects
  – program effects

• Requirements for valid estimates vs real world
  – models assume
    • students randomized to teachers
    • tests do not have ceiling or floor effects
    • vertical alignment of tests
  – in reality
    • student assignment not random (for good reasons)
    • tests often have ceiling / floor effects
Math in the Middle (M²) Research

• **Focus**: Understand how changes in teachers’ math knowledge translate into measurable improvement in student performance

• Extensive data collection from both participating teachers and their students
  – Classroom observations
  – Teacher surveys & interviews
  – Student achievement data on district assessments
Student Achievement Data

- School-based Teacher-led Assessment & Reporting System (STARS)
  - Each school district in state chose tests deemed appropriate
  - Variety of tests administered at various points during school year across districts

- Lincoln Public Schools (LPS)
  - Largest participating school district
  - Grades 5-8 data collected since 2003-2004
  - 317 LPS middle school teachers, 37 M² participants
LPS Data

- Metropolitan Achievement Test (MAT)
- Criterion Referenced Test (CRT)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>MAT</td>
<td>MAT</td>
<td>MAT &amp; CRT</td>
<td>MAT &amp; CRT</td>
<td>MAT &amp; CRT</td>
</tr>
<tr>
<td>6</td>
<td>MAT</td>
<td>MAT</td>
<td>MAT &amp; CRT</td>
<td>CRT</td>
<td>CRT</td>
</tr>
<tr>
<td>7</td>
<td>MAT</td>
<td>MAT</td>
<td>MAT &amp; CRT</td>
<td>MAT &amp; CRT</td>
<td>MAT &amp; CRT</td>
</tr>
<tr>
<td>8</td>
<td>---</td>
<td>---</td>
<td>CRT</td>
<td>CRT</td>
<td>CRT</td>
</tr>
</tbody>
</table>
Research Objectives

• Investigate how to adapt a layered value-added model...
  – For modeling longitudinal student achievement data collected from a mixture of norm- and criterion-referenced assessments
  – To estimate the impact of a professional development program on student learning
Value-Added Methodology

• Estimate program effect
  – Adapted layered VAM to estimate participating & non-participating teacher effects

• Create coherent picture of student growth trajectories
  – Z-scores
  – Parallel Processing
Z-scores

- Calculated within a given year and grade
- Reflects number of standard deviations original score away from average
- Within a cohort of students, changes from year to year reflect differences in relative performance
- Weighted layered teacher coefficient matrix to account for possible variance restrictions when using z-scores with layering
Before Participation Effect Estimates for $M^2$ & non-$M^2$ Teachers

- **non-$M^2$ Mean** = -0.038  
  S.D. = 0.359

- **$M^2$ Mean** = 0.053  
  S.D. = 0.372
Program Effect Estimates for M² Participants

Median Diff = 0.004
Mean Diff = 0.027
S.E. = 0.073
n = 22

Program effect = Participation – Before participation
Limitations

• Z-scores vulnerable to outliers & wide swings in standard deviations
  – Small population of rural districts exacerbates problem

• Changes in mobility rates, curricula and/or test content

• Possibly arbitrary scaling of layered teacher coefficients
Parallel Processing

- Latent growth curve models
- Model growth on multiple parallel processes
- Curve-of-factors model (Little et al., 2006)
  - Estimate teacher’s effect on longitudinal changes in a latent trait (e.g., math ability) measured by multiple instruments in each year
Curve-of-Factors Model (Level 1)

* Freely estimated (i.e., unconstrained) parameter
Simulation Study

- Compared predicted to true teacher percentiles
- Estimated root mean square error (RMSE) is largest at middle percentiles for all models
  - Difficult to predict average teacher’s true rank
- Curve-of-factors model loses its statistical advantages (e.g., smaller RMSE and absolute bias) when tests missing in some years
- Z-score models’ results similar in both complete and missing tests cases
Key Insights...at the Time

• Z-scores are an alternative to using raw data when analyzing less-than-ideal achievement data

• Weighting considerations for variance components may arise when using a layered VAM to analyze gains in z-scores

• Parallel processing utilizes information from multiple measurements of a common latent trait over time

• Simulation tentatively suggests negligible benefits with the complexity of parallel processing
Some Issues

• Team teaching
• High mobility rates
• Curricula and test content vary across grades
  – VAM teacher effects sensitive to ways achievement is measured (Lockwood et al., 2007)
• Ceiling effects
• Uncertainty of program / “teacher” effects
  – Link VAM estimates to alternative, valid measures of teacher effectiveness
Program Effects

Investigated model behavior under identified data characteristics through simulation

– Non-random assignment of students to classrooms
– Assessment ceiling effects
– Student growth trajectories
Randomization

- Previous studies address non-randomization to “game” the VAM
- We address non-randomization through the tracking process commonly used in schools
- No significant impact on accuracy of PD program effect estimates
- Small (but possibly important) impact on precision
Ceiling Effects

• Numerous cited examples of state assessments with ceiling effects
• Explored ceiling effects in combination with non-random student assignment
• Sufficient impact to invalidate PD program estimates
• Assessing teacher & program effect requires tests with adequate “stretch”
Student Trajectories

Recall...

Test Scores

<table>
<thead>
<tr>
<th>Year</th>
<th>Diff between student score &amp; district avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td></td>
</tr>
<tr>
<td>g + 1</td>
<td></td>
</tr>
</tbody>
</table>

Additional gain for student

Additional Student-specific growth

District avg.

Additional
Student Growth Trajectories

• Students do not grow at the same rate

\[ y_{ij} = \beta_0 + \beta_1 G + \sum Teachers + e_{ij} \]

\[ y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1j})G + \sum Teachers + e_{ij} \]

• Evidence of positive correlation between initial status and rate of growth

• Exacerbates problems associated with
  – Non-randomization
  – Ceiling effects
Implications

• VAMs can help inform quality improvement in education
• Help inform re: how are we doing
• Variability
  – Estimates of teacher / program effects involve a mean AND a standard error
  – Standard errors tend to be large enough so that precise statements about individual teachers require extreme caution
  – Help improve: yes; high stakes evaluation: caution
Dead Ends – Things We Tried

• Binning
  – classify score by rank
  – binning $\rightarrow$ divide into quantiles
  – works in some applications, but not for VAM
  – too much lost information, modeling headache

• Historical data with strong ceiling effect
  – e.g. LPS achievement data pre-state test
  – attempted to use for Math in the Middle
  – in view of ceiling effect finding, lack of success no surprise
Quality Improvement
W Edwards Deming

- Preeminent figure / founding father of QI
- “Not enough to do your best. You have to know what to do, then do your best.”
- “Profound Knowledge” – understanding and working with variation
- 14 Points
  - 3: cease dependence on inspection
  - 11: eliminate management by numbers & numeric goals
- 85/15
Deming, QI and VAM

• Deming advocated data-informed quality improvement
• Deming deplored merit evaluation in any form
• VAM can be effectively used in a manner consistent with guidelines Deming articulated
• more ... Sharon Lohr’s talk tomorrow ...
Data Connections is supported by the National Science Foundation grant DUE-1050667, with additional support from UNL’s Center for Science, Mathematics and Computer Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.