Overview

• Purpose of this presentation
• Description of task at hand
• Overview of various approaches available and considered
• Selected results
• Advantages/limitations of each approach
• Recommendations and lessons learned
• Next steps
Purpose of the presentation

• There is an increased attention and use of growth models in assessment and accountability.

• There are technical difficulties with representing growth that cause limitations.

• There are technical difficulties that occur with performance assessments.

• This presentation shows an example of using performance assessment results to represent growth using a couple of methods with a highlight of advantages and challenges of these approaches.
Description of task at hand

• Performance assessments narrowly defined as essay tests on various topics and grade levels were developed.

• The purpose of these assessments were to capture student performance for use as one of multiple components of an educator evaluation system.
  – The component with the largest weight represented measures of teacher practice.

• The assessments were developed in the disciplines of mathematics, English language arts, science, and history/social studies for grades K to 12.

• The assessments were administered at the beginning and end of a school year with individual student results aggregated to the teacher.
  – Changes in student performance on these essay tests were to be used as one of the components.

• Question: How to represent growth of student performance aggregated to each teacher?
Design of Assessments

- As suggested by a framework for developing and selecting assessments of student growth for use in teacher evaluations (Herman, Heritage, & Goldschmidt, 2011), a series of claims representing the design and validation of assessments for use in teacher evaluations was followed.

- The essays were developed to represent the learning objectives of the courses in the disciplines of English language arts, history/social studies, mathematics, and science in grades K to 12.

- A long-term effort occurred whereby the key learning objectives were identified and the assessments were developed to address them.
  - The learning objectives represented what aspects of the discipline/course were important by the end of the grade.

- Essay prompts and scoring rubrics were developed to represent the learning objectives for each course and grade.
  - The rubrics were based on performance levels anchored in the learning objectives and content standards.
  - Multiple traits were scored using multi-point rubrics.
Design of Assessments (cont’d)

• While the assessments were designed in a principled manner, the way that growth would be represented was considered after the assessments were developed and administered in the two test administrations.

• The essays were administered at the beginning of the academic year and then alternate forms of the essays were administered at the end-of-the academic year.
  
  – The alternate forms were developed by keeping the main features of the essays the same and changing the surface features.
  
  – Evaluated scorer and alternate form reliability
The number of traits varied by discipline.

- English Language Arts: 7
- History/Social Studies: 6 (HS); 7 (ES & MS)
- Mathematics: 14
- Science: 4

Rating values: 0 – 4 or in mathematics (0-2, 0-3 or 0-4)

Rating labels based on degree to which student responses met expected standards:

- 0 → No evidence
- 1 → Attempting the Standards
- 2 → Approaching the Standards
- 3 → Meeting the Standards
- 4 → Exceeding the Standards

Scoring:

- Teachers from the discipline scored the responses after training and interrater reliability checks were done.
- Scores for each trait were added together to form a total score.
Considerations for Growth

• There are fundamental aspects that should be considered in selecting a growth model or representing student performance on assessments over time (for example Castellano & Ho, 20013; Patelis et al., 2013).

• While the users of this assessment had a growth model in-mind, options were not considered in their ultimate decision.

• Various options were considered, however, under a research effort. While all the details and results will not be presented here, an overview of the findings with an indication of the advantages and challenges are provided.

  – Because there are many grades and disciplines, the grade 5 English language arts assessment is used to offer this overview and discussion.
A number of models were considered from the types available. Castellano & Ho (2013) categorized models as follows:

- **Gain Score** - describes growth with simple differences or average gains over time.
- **Trajectory** - Extends gains or average gains in a predictable, usually linear fashion into the future.
- **Categorical** - Defines growth by transitions among status categories (e.g., Basic, Proficient, Advanced) over time.
- **Residual Gain** - Describes growth as the difference between current status and expected status given past scores.
- **Projection** - Uses past scores to predict future scores through regression equations.
- **Student Growth Percentile Model** - Percentile rank of current status in a reference group of students with similar past scores.
- **Multivariate** - Uses entire student score histories, including other subjects and teachers, to detect higher than expected student scores associated with particular teachers.
Overview of Growth Models

- A number of models were considered from the types available. Castellano & Ho (2013) categorized models as follows:

  - **Gain Score** - describes growth with simple differences or average gains over time
  - **Trajectory** - Extends gains or average gains in a predictable, usually linear fashion into the future
  - **Categorical** - Defines growth by transitions among status categories (e.g., Basic, Proficient, Advanced) over time.
  - **Residual Gain** - Describes growth as the difference between current status and expected status given past scores.
  - **Projection** - Uses past scores to predict future scores through regression equations.
  - **Student Growth Percentile Model** - Percentile rank of current status in a reference group of students with similar past scores.
  - **Multivariate** - Uses entire student score histories, including other subjects and teachers, to detect higher than expected student scores associated with particular teachers.

*Used these two approaches to illustrate results and highlight the advantages and challenges!*
Gain Score
Descriptive Statistics – Student Performance

**Beginning of the School Year**

- **Shift of distribution toward higher scores**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>25554</td>
<td>0.00</td>
<td>28.00</td>
<td>10.2</td>
<td>5.2</td>
</tr>
<tr>
<td>End</td>
<td>23472</td>
<td>0.00</td>
<td>28.00</td>
<td>17.0</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Was the mean difference adequate?

While there was movement in total scores, the effect was small to moderate.
Descriptive Statistics – Teacher Performance

Beginning of the School Year

End of the School Year

Shift of the teachers’ mean scores

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>1115</td>
<td>5.00</td>
<td>17.00</td>
<td>10.2</td>
<td>1.8</td>
</tr>
<tr>
<td>End</td>
<td>1115</td>
<td>10.00</td>
<td>23.00</td>
<td>17.0</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Was the mean difference adequate?

Large effect seen in the change by teacher

Center for Assessment

Growth and Performance Assessments
Advantages & Challenges

• Intuitive
• Easy to calculate
• Scale represents sum of trait scores.

• Requires judgment about adequate average gain.
  – Need to implement a standard setting
Categorical
Categories

- Traits represented the following:
  1. Focus of Opinion
  2. Development
  3. Reading
  4. Organization – Intro & Conclusions
  5. Organization – Structure
  6. Organization – Transition
  7. Conventions

- Ratings for each score point are as follows:
  - 0 → No evidence
  - 1 → Attempting the Standards
  - 2 → Approaching the Standards
  - 3 → Meeting the Standards (Represents a target for adequate performance)
  - 4 → Exceeding the Standards

- Two scoring models were used:
  - Each of the traits scores at/above 3 represent adequate performance
    - Students who had a 3+ on each of the seven traits were counted as adequate performance
  - Each of the three traits at/above 3 and one of the three organization traits at/above 3 represent adequate performance
    - Students who had a 3+ on four traits and a 3+ on one of the three organization traits were counted as adequate performance
### Descriptive Statistics – Student Performance

<table>
<thead>
<tr>
<th></th>
<th>Beginning</th>
<th></th>
<th>End</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
</tr>
<tr>
<td>All Traits</td>
<td>449</td>
<td>2%</td>
<td>4886</td>
<td>19%</td>
</tr>
<tr>
<td>All Traits with 1 of 3 Org Traits</td>
<td>671</td>
<td>3%</td>
<td>6157</td>
<td>24%</td>
</tr>
<tr>
<td>Total</td>
<td>25758</td>
<td></td>
<td>25758</td>
<td></td>
</tr>
</tbody>
</table>

Using either method one can see larger percentage of students reaching adequate performance.
Descriptive Statistics – Teacher Performance

Shift in the % of students for each teacher meeting adequate performance
Advantages & Challenges

• Categories set by the way rubrics were developed
  – Permit more straightforward interpretation

• Match performance standards

• Performance levels articulated across grades as part of design

• Impact of compensatory vs. conjunctive model in aggregating decision across traits

• Requires judgments about adequate percentage of students within a schools at end-of-year and adequate change of percentage
Recommendations

• Establishing standards and learning objectives important, which form the basis for developing prompts and scoring rubrics

• Build performance levels into rubrics

• Trade-offs in interpretation depending on growth model use.

• Different types of judgments of what is adequate growth must be done depending on growth model used:
  – Gain score: What is adequate gain in scores?
  – Categorical: What is adequate performance across traits over time?
Next Steps

• Develop scale using difficulty of essay prompts and rating accuracy. Calculate and evaluate gain scores.

• Compare results of all models for all teachers to examine similarities of decision classifications across approaches.


Center for Assessment
www.ncriea.org

Thanos Patelis – tpatelis@nciea.org