Developing the Vertical Scale for the Connecticut Mastery Test

Hariharan Swaminathan
University of Connecticut
Thanks to

Barbara Beaudin
Bill Congero
Mohammed Dirir
Gil Andrada
Steve Martin
H. Jane Rogers

CT State Department of Education

WWW.STATE.CT.GOV/SDE
Connecticut’s Testing Context

- The Connecticut Mastery Test
  - Criterion-referenced test
  - First administered in 1985 to Grades 4, 6 and 8
  - Subject: Mathematics, Reading, and Writing
  - Performance levels within each grade: Below Basic, Basic, Proficient, Goal and Advanced
Connecticut’s Testing Context

- In 2006,
  - Grades 3 through 8
  - Science was added in grades 5 and 8

- Within each grade and subject area, test forms are equated across each generation of the test.
Testing Context (cont.)

• The content of the Reading and Mathematics tests differs across grades and the standards (cut scores on a scale from 100 to 400) establishing the performance levels were developed independently for each grade.

• Results are reported at the individual child, by grade within school, grade within district, grade within the state.

• Reports are also disaggregated by gender, race/ethnicity, poverty, special education, and ELL status.
Testing Context (cont.)

• Districts could use the results to answer questions like: *How well are this year’s fifth grade students doing in mathematics?*

• If the characteristics of the student populations, such as SES, race/ethnicity, are similar across schools within the district, they could compare performance across schools.
• Districts could compare the performance of subgroups of students: *Are our fifth grade female students performing as well or better than their male counterparts in Mathematics?*

• If last year’s fifth graders are similar to this year’s demographically, for comparative purposes they could answer: *Are this year’s students doing as well or better in Mathematics than last year’s fifth graders?*

• If the demographics of the district’s student population are similar to other districts in the state they could answer the question: *Are we doing as well or better in grade five Mathematics than these other districts with similar populations?*
Testing Context (cont.)

NCLB AYP

• STATUS MODEL
  States are required to set a target - ANNUAL MEASURABLE OBJECTIVE (AMO) – Percent of students achieving Proficiency or above.
  AMO must be set in such a way that ALL students achieve proficiency in ALL subjects by 2014

• “GROWTH” MODEL
  Performance of students in the same grade is compared across years. Requires that the tests are “equated” from year to year – HORIZONTAL EQUATING
Testing Context (cont.)

• NCLB Requirements:
  With NCLB’s heightened emphasis on accountability, school administrators felt that the status measure (% Proficient) used to determine AYP was limited in conveying their student’s performance and the academic progress their students were making.

• This was particularly true for districts with large proportions of low-income students.
Testing Context (cont.)
GROWTH ACROSS GRADES

- Districts were eager to answer questions: *How much have our students progressed or grown in Mathematics from Grade four to Grade five?* 

- As a result, district staff, advocacy groups and the media often used inappropriate comparisons like “an average increase of 23 scale score points from Grade four to Grade five” or “an increase of six percentage points in the proportion of students performing at the goal level from Grade four to five.”
Testing Context (cont.)

• Reporting the scale increase would be inappropriate

  • the fifth grade scale is independent of the fourth grade scale

  • they are not on a single scale.
Testing Context (cont.)

• Reporting an increase in the percentage of students scoring at or above the ‘goal’ level would be inappropriate

  • the fifth grade content is different from the fourth grade content

  • the standards are independent of each other.
    • ‘goal’ does not mean the same for both grades
    • students could have grown between Grade four and five but not met the goal standard.
In 2005, Connecticut instituted a State Assigned Student Identifier (SASID) for every public school student in the state.

- This provided a key for linking student performance across grades and linking information about students across data files.

- 2006 was the first year of the new generation of the CMT administered to students in Grades 3 through 8.
The structures were now in place to create a vertical scale for Mathematics and Reading to provide the Connecticut public with a vehicle to measure ‘growth” or progress over time.
What are the Merits of a Vertical Scale?

The new vertical scale

• ranges from 200 to 700

• allows schools to measure student progress across a body of content knowledge ranging from Grade 3 to Grade 8 for Mathematics and Reading.
What are the Merits of a Vertical Scale?

- Each point on the scale represents a level of student understanding, independent of the grade the student is in.
  - So, a score of 490 on the mathematics scale for a fourth grader represents the same level of knowledge as 490 for a fifth grader.

- ‘Growth’ can be measured across grades as the difference between the score in Year (n+ 1) and Year (n).
What Practical Solutions Does the Vertical Scale Provide?

- For Accountability, GROWTH
  - provides an additional indicator of school performance (Value Added Models)
  - supplements the status measure (performance level).
  - provides an indicator for measuring the effectiveness of new curriculum or professional development initiatives (Value Added Models)
STATES THAT HAVE VERTICAL SCALES

- ARIZONA (3-8)
- ARKANSAS (3-8)
- FLORIDA (3-10)
- IOWA ITBS (3-8)
- NORTH CAROLINA (3-8)*
- TENNESSEE (3-8)*
Developing the Vertical Scale

- Framework for Developing the Vertical Scale
- Data Collection Design
- Scaling Procedures
- Adjusting the Tails of the Scale
- Value Added Models
Framework

• Constructing a vertical scale requires a different measurement framework from “classical” methods
THE “OLD” WAY :
CLASSICAL TEST THEORY

OBSERVED SCORE  =  TRUE SCORE  +  ERROR

X  T  E
CLASSICAL TRUE SCORE MODEL

- Based on weak assumptions
- Both $T$ and $e$ are unobserved
- Intuitively appealing for its simplicity
- Model can be extended to study different sources of error
INDICES USED IN TRADITIONAL TEST CONSTRUCTION

Item Indices:

• Item difficulty – Proportion of examinees answering the item correctly
• Item discrimination – Correlation between Item Score and Total Score

Examinee Index:

• Test score
CLASSICAL ITEM and EXAMINEE INDICES

- Item difficulty and Item Discrimination depend on the group of examinees
- Test Score depends on the items
What’s wrong with these indices?

They are **SAMPLE DEPENDENT**

i.e., they change as the sample of people or the sample of items change - **NOT INVARIANT**
And what’s wrong with that?

• We cannot compare item characteristics for items whose indices were computed on different subgroups of examinees

• We cannot compare the test scores of individuals who have taken different subsets of test items
Wouldn’t it be nice if ....?

- Our item indices did not depend on the characteristics of the sample of individuals on which the item data was obtained

- Our examinee measures did not depend on the characteristics of the sample of items that was administered
ITEM RESPONSE THEORY solves the problem!*

* Certain restrictions apply. Individual results may vary. IRT is not for everyone, including those with small samples. Side effects include yawning, drowsiness, difficulty swallowing, and nausea. If symptoms persist, consult a psychometrician. For more information, see Hambleton and Swaminathan (1985), and Hambleton, Swaminathan and Rogers (1991).
ITEM RESPONSE THEORY

postulates that

the probability of a correct response to an item depends on the ability of the examinee and the characteristics of the item
ABILITY > DIFFICULTY \[\rightarrow\] PROBABILITY OF CORRECT ANSWER IS HIGH

ABILITY < DIFFICULTY \[\rightarrow\] PROBABILITY OF CORRECT ANSWER IS LOW
The mathematical relationship between the probability of a response, the ability of the examinee, and the characteristics of the item is specified by the

ITEM RESPONSE MODEL
An item may be characterized by its
DIFFICULTY level (usually denoted as $b$),
DISCRIMINATION level (usually denoted by $a$),
“PSEUDO-CHANCE” level (usually denoted as $c$).
One-Parameter Model

\[ P(\text{correct response}) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}} \]

\( \theta \) = trait value

\( b \) = item difficulty
IRT ITEM DIFFICULTY

• Differs from classical item difficulty

• $b$ is the $\theta$ value at which the probability of a correct response is .5

• The harder the item, the higher the $b$

• Hence, $b$ is on the same scale as $\theta$ and does not depend on the characteristics of the sample
One-Parameter (Rasch) Model

\[ b = -1.5 \]

\[ \text{Probability of Correct Response} \]

\[ \text{Theta (Proficiency)} \]
One-Parameter (Rasch) Model

Probability of Correct Response

Theta (Proficiency)

-5 -4 -3 -2 -1 0 1 2 3 4 5

-5 -4 -3 -2 -1 0 1 2 3 4 5

b = -1.5

b = 0
One-Parameter (Rasch) Model

b = -1.5

b = 0

b = 1.5

Probability of Correct Response

Theta (Proficiency)
Two-Parameter Model

\[ P(\text{correct response}) = \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \]

\[ \theta = \text{trait value} \]

\[ b = \text{item difficulty} \]

\[ a = \text{item discrimination} \]
IRT ITEM DISCRIMINATION

• Differs from classical item discrimination

• \( a \) is proportional to the slope of the ICC at \( \theta = b \)

• The slope of the item response function indicates how much the probability of a correct response changes for individuals with slightly different \( \theta \) values, i.e., how well the item discriminates between them
Two-Parameter Model

The graph illustrates the relationship between the probability of a correct response and theta (proficiency). The x-axis represents theta (proficiency) ranging from -5 to 5, while the y-axis represents the probability of correct response ranging from 0 to 1.

Key elements include:
- **α - Discrimination - Slope**: This parameter affects the steepness of the curve, indicating how well the model can distinguish between different proficiency levels.
- **Difficulty - b**: This parameter affects the intercept of the curve, indicating the baseline difficulty of the task.

The graph shows how the probability of a correct response increases as theta (proficiency) increases, with the curve reflecting the influence of both discrimination and difficulty parameters.
Two-Parameter Model

\[ a = 0.5 \]
\[ b = -0.5 \]
Two-Parameter Model

\[ a = 0.5 \]
\[ b = -0.5 \]

\[ a = 2.0 \]
\[ b = 0.0 \]
Two-Parameter Model

The graph illustrates the probability of correct response as a function of theta (proficiency). The graph shows three curves for different parameter values:

- **Red curve**: $a = 0.5$, $b = -0.5$
- **Yellow curve**: $a = 2.0$, $b = 0.0$
- **Blue curve**: $a = 0.8$, $b = 1.5$

The x-axis represents theta (proficiency), ranging from -5 to 5. The y-axis represents the probability of correct response, ranging from 0.0 to 1.0.
Three-Parameter Model

\[ P(\text{correct response}) = c + (1-c) \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} \]

\[ c = \text{pseudo-chance level} \]
\[ (\text{“guessing”}) \text{ parameter} \]
IRT ITEM GUESSING PARAMETER

• No analog in classical test theory

• $c$ is the probability that an examinee with very low $\theta$ will answer the item correctly
Three-Parameter Model

Theta (Proficiency) vs. Probability of Correct Response

- Lower Asymptote (c)
Three-Parameter Model

\[
\text{Probability of Correct Response} = \frac{e^{a \cdot \text{Theta} + b}}{1 + e^{a \cdot \text{Theta} + b}} + c
\]

- \(a = 0.5\)
- \(b = -0.5\)
- \(c = 0.0\)

Theta (Proficiency)
Three-Parameter Model

Probability of Correct Response

Theta (Proficiency)

- a = 0.5
- b = -0.5
- c = 0.0

- a = 2.0
- b = 0.0
- c = 0.25
Three-Parameter Model

- $a = 0.5$
- $b = -0.5$
- $c = 0.0$
- $a = 2.0$
- $b = 0.0$
- $c = 0.1$
- $a = 0.8$
- $b = 1.5$
- $c = 0.25$
BUT WAIT,

There’s more!
IRT Models for Polytomous Item Responses

- Polytomous IRT models specify the probability that an examinee will respond in (or be scored in) a given response category
Graded Response Model

• For each response category, the probability that an examinee will respond in that category or higher vs. a lower category is given by:

\[
P(u \geq k) = \frac{e^{a(\theta - b_k)}}{1 + e^{a(\theta - b_k)}}
\]
Graded Response Model

• The probability that the examinee will choose a given response category (or be scored in it) is thus

\[ P(u=k) = P(u \geq k) - P(u \geq k + 1) \]
Partial Credit Model

• For every pair of adjacent categories, specify the probability that the examinee will choose the higher of the two (or be scored in the higher):

\[
P(u = k | u = k \text{ or } k+1) = \frac{e^{a(\theta-b_k)}}{1 + e^{a(\theta-b_k)}}
\]
Partial Credit Model

- This results in the model

\[
P(u = k) = \frac{\sum_{r=1}^{k} e^{a(\theta-b_r)}}{1 + \sum_{s=1}^{m-1} \sum_{r=1}^{s} e^{a(\theta-b_s)}}
\]
Polytomous Item Response Model

Graph showing the probability distribution of different proficiency levels (0 to 4) across different proficiency scores.
MORE?

No, that’s enough for now
ASSUMPTIONS UNDERLYING IRT

- For the most commonly used models, unidimensionality of the latent trait is assumed.

- The mathematical model is assumed to correctly specify the relationship between the trait and item characteristics and performance on the item.
FEATURES OF IRT

When the model fits the data,

1. Item parameters are independent of the population of examinees, i.e., invariant across sub-groups

2. Ability parameters are independent of the subsets of items administered
IMPLICATIONS

1. Different examinees can take different items. Their ability values will be comparable!

2. Item parameter values will not be affected when different samples of examinees respond to the item.
IMPLICATIONS FOR VERTICAL SCALE

1. We can compare the performance of examinees who take two different tests (e.g., Grade 4 and Grade 5) if ..... 

2. We can somehow *LINK* the items in the two tests and place them on the same scale.
How do we do this?

1. We give Grade 4 and Grade 5 examinees some common items.

2. Since the item parameter values of these common items will be the same for Grade 4 and Grade 5 examinees, we can use this information to place all the Grade 4 and Grade 5 items on a common scale.
3. Once the item parameter values are on a common scale, we can administer any subset of items to examinees in Grades 4 and 5.

4. Since the ability values based on different subsets of items are on the same scale, we can compare the ability values of examinees in the two grades.

5. This design where examinees in two grades are given common items is known as the ANCHOR TEST DESIGN.
VERTICAL SCALING

• IRT is used for all vertical scaling implementations

• Decisions to be made in carrying out vertical scaling with IRT:
  • Data collection design
  • Choice of IRT model
  • Estimation Procedure
  • Scaling procedure
Data Collection Design

- Scaling test design
  - *A test containing items that span all grades is administered to all students*

- Common item design
  - *Examinees in each grade take the on-grade test and some items from adjacent grade tests*
Choice of Model

- 1P Model used in some state assessments (e.g., CT, DE)

- 3P Model used in other states (e.g., MI, FL)
Estimation Procedure

- Different calibration programs use different estimation procedures
  - WINSTEPS fits a 1P/PCM using an unconditional maximum likelihood procedure
  - BILOG-MG, PARSCALE, MULTILOG use marginal maximum likelihood estimation
Scaling Procedures

- Different IRT procedures are available:
  - Separate group calibration
  - Concurrent calibration
  - Fixed parameter calibration
Data Collection Design
THETA DISTRIBUTION FOR MATHEMATICS ACROSS GRADERS

- Grade 3
  - Mean: 0.3371
  - Std Dev: 1.1873

- Grade 6
  - Mean: 1.1977
  - Std Dev: 1.1977

- Grade 5
  - Mean: 1.2862
  - Std Dev: 1.2862

- Grade 3
  - Mean: 2.246
  - Std Dev: 1.246

- Grade 7
  - Mean: 2.5287
  - Std Dev: 1.5287

- Grade 3
  - Mean: 2.5328
  - Std Dev: 2.5328
THETA DISTRIBUTION FOR READING ACROSS GRADES

- Grade 3:
  - Mean: -0.0991
  - Standard Deviation: 1.1313

- Grade 4:
  - Mean: 0.5384
  - Standard Deviation: 1.17

- Grade 5:
  - Mean: 1.1291
  - Standard Deviation: 1.1665

- Grade 6:
  - Mean: 1.4707
  - Standard Deviation: 1.2533

- Grade 7:
  - Mean: 1.7591
  - Standard Deviation: 1.2438

- Grade 8:
  - Mean: 2.0393
  - Standard Deviation: 1.2426
Vertically Scaled Scores

• The theta values are like Z scores – the mean is zero and SD is 1.

• The theta values can be linearly transformed (like Z scores) to span any range or to have a certain mean and SD.

• The theta scale is transformed to have a score range from 200 to 700.
The Tale of Two Tails:
Adjustments to the scale
Figure 1. Relationship Between Raw Score and $\hat{\theta}_r$
Relationship of the Vertically Scaled Score (VSS) to $r$

The VSS is a linear transformation of $\hat{\theta}_r$.

Hence

$$r = \sum_{j=1}^{n} P_j \left( VSS \right)$$
Relationship Between Raw Score and Vertically Scaled Score

![Graph showing the relationship between raw score and vertically scaled score. The x-axis represents raw score ranging from 0 to 80, and the y-axis represents vertically scaled score ranging from 0 to 680. The graph indicates a positive correlation between the two scores.](https://example.com/graph.png)
Problem (?)

Small changes in raw score at the extremes result in large changes in the Vertically Scaled Scores

This may cause concern to users when a child progresses from one grade to the next
Figure 2. Relationship Between Vertical Scaled score and Raw Score For Children Progressing from Grade 3 to Grade 4 (Lower End - Reading)
Figure 3. Relationship between Vertical Scaled Score and Raw Score For Children Progressing from Grade 3 to Grade 4 (Upper End - Reading)
Table 1. Gain in Unadjusted Vertical Scale Score Corresponding to an Increase of in Raw Score (LOWER END)

<table>
<thead>
<tr>
<th>Raw Score Grade 3</th>
<th>Gain in Unadjusted Vertical Scale Score Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Score +1</td>
</tr>
<tr>
<td>0</td>
<td>87</td>
</tr>
<tr>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td>3</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
</tr>
<tr>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>14</td>
<td>29</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
</tr>
</tbody>
</table>
Is this a Real Problem?

- This is a natural consequence of a nonlinear transformation
- This is not a problem inherent in the vertical scale
- Increases (or decreases) at the extremes affect only a small percent of test takers
- However, it is an issue of FACE VALIDITY that should be addressed
Solution

- Adjustments are made to the two tails
- Linear interpolation at the extreme ends resolves the problem
Figure 3. ADJUSTMENT TO GRADES 3 AND 4 – (Lower End)
Figure 3. ADJUSTMENT TO GRADES 3 AND 4 – (Lower End)
Figure 3. ADJUSTMENT TO GRADES 3 AND 4 – (Lower End)
Steps

1. Choose a suitable region where the curve is almost linear – where the rate of change is almost a constant.
2. Fit a regression line/ or an analytical line and extend this line.
3. Check to see if the growth (rate of change) is similar to that in the middle of the score range.
4. Carry out the procedure across two grades, e.g., from Grade 3 to 4, checking to make sure the growth from one grade to the next is almost uniform across the raw score range.
5. Keeping the Grade 4 line fixed, carry out the procedure across grades 4 and 5.
6. If the change from grade 4 to 5 is not uniform, reexamine the growth from grades and modify the line(s) until the growths from grades 3 to 4 and 4 to 5 are “reasonable”.
7. Repeat this process across all the grades.
Table 2. Gain in Adjusted Vertical Scale Score Corresponding to an Increase in Raw Score (LOWER END)

<table>
<thead>
<tr>
<th>Raw Score Grade 3</th>
<th>Gain using Adjusted Vertical Scale Score Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Score +1</td>
</tr>
<tr>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td>11</td>
<td>28</td>
</tr>
<tr>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>14</td>
<td>28</td>
</tr>
<tr>
<td>15</td>
<td>28</td>
</tr>
</tbody>
</table>
Figure 4. Adjustments to Grades 3 and 4  (Upper End)
Figure 4. Adjustments to Grades 3 and 4 (Upper End)
Figure 4. Adjustments to Grades 3 and 4 (Upper End)
<table>
<thead>
<tr>
<th>Raw Score</th>
<th>Gain using Adjusted Vertical Scale Score at the upper end of the scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raw Score +1</td>
</tr>
<tr>
<td>70</td>
<td>11</td>
</tr>
<tr>
<td>71</td>
<td>11</td>
</tr>
<tr>
<td>72</td>
<td>11</td>
</tr>
<tr>
<td>73</td>
<td>11</td>
</tr>
<tr>
<td>74</td>
<td>11</td>
</tr>
<tr>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>76</td>
<td>11</td>
</tr>
<tr>
<td>77</td>
<td>11</td>
</tr>
<tr>
<td>78</td>
<td>12</td>
</tr>
<tr>
<td>79</td>
<td>12</td>
</tr>
<tr>
<td>80</td>
<td>12</td>
</tr>
<tr>
<td>81</td>
<td>12</td>
</tr>
<tr>
<td>82</td>
<td>12</td>
</tr>
<tr>
<td>83</td>
<td>12</td>
</tr>
<tr>
<td>84*</td>
<td></td>
</tr>
</tbody>
</table>
A Difficulty

• A difficulty arises when test in one grade has a different number of items than the test in another grade
• What happens when Grade 4 test has 84 items but Grade 5 test has 98 items?
• The curves will cross
Figure 4. Relationship between Vertical Scaled Score and Raw Score: Grade 4 to Grade 5 (Upper End - Reading)
• The curves cross at the raw score value of 56.
• A student with a raw score of 70 in Grade 4 and 70 in Grade 5 will show a decline in the VSS of 15 VSS points (490 – 475 = -15).
• A score of 70 in Grade 4 corresponds to percent correct score of 83.3
• A score of 70 in Grade 5 corresponds to percent correct score of 71.4

A Difficulty? – Not really!
Report Percent Correct Scores
Recommendations

- IRT scaling results in a nonlinear relationship between raw score and Vertically Scaled Score
- Small changes in raw scores at the extreme result in large changes in VSS at the extremes
- This is not a theoretical drawback of the VSS, but a perception problem for the test users
- This is an issue of face validity
- Problem is addressed by simple adjustments at the extremes
Recommendations

• We do not recommend describing growth with reference to raw scores

• If observed scores on the test have to be reported, use Percent Correct Scores. This avoids appearances of problems when tests have different number of raw score points
PROBLEMS WITH VERTICAL SCALING

• If the construct or dimension being measured changes across grades/years/forms, scores on different forms mean different things and we cannot reasonably place scores on a common scale.

• Martineau (2006) warns against using vertical scales in the presence of Construct Shift. Above and below grade items with scales for adjacent grades may reduce construct shift.
PROBLEMS WITH VERTICAL SCALING

• Both common person and common item designs have practical problems; items may be too easy for one group and too hard for the other

• Must ensure that examinees have had exposure to content of common items or off-level test
PROBLEMS WITH VERTICAL SCALING

• Scaled scores are not interpretable in terms of what a student knows or can do

• Comparison of scores on scales that extend across several years is particularly risky and not recommended


VALUE ADDED ASSESSMENT MODELS AND ISSUES
Accountability

• Schools and Districts are being held accountable for the performance of students

• State Laws are being proposed to evaluate Teachers, Principals, & Superintendents on the basis of student achievement
Accountability

This trend is not new!

- Travisu, Italy, early 18th century: Teachers were not paid if students did not learn!
- Shogun: ultimate accountability! The teachers (entire village) to be put to death if the student (Blackthorne) does not learn Japanese.
Approaches to Accountability

- **STATUS ASSESSMENT**: Achievement of a pre-set target (AYP)

- **GROWTH ASSESSMENT**:
  - *Cross-sectional (e.g., comparison of performance of students in Year 1 with that of students in the SAME GRADE in Year 2)*

  - *Longitudinal (assessment of growth of students from Year 1 to Year 2)*
Value-Added Assessment

• Is basically an assessment of the effectiveness of a “treatment” by examining growth

• Looks at the growth to determine if it is the result of the “value added” by the “treatment”

• Adjusts for “Initial Status” when assessing the “effect” of the treatment

• In simple terms, it compares actual growth with “expected” growth.
Teacher Effectiveness: Value Added Assessment

“Treatment” $\rightarrow$ Teacher

Can Value-Added Assessment Models untangle teacher effects from non-educational effects?
Value Added Models

- Gain Score Models
- Covariate Adjustment Models
- Layered/Complete Persistence Model
- Variable Persistence Model
  (Lockwood, et al. 2007)
Sources & Resources

• JEBS (2004) Volume 29, Number 1


Layered Model/
Complete Persistence Model

\[ y_{i1} = \mu_1 + u_1 + e_{i1} \]
\[ y_{i2} = \mu_2 + u_1 + u_2 + e_{i2} \]

\[(y_{i2} - y_{i1}) = (\mu_2 - \mu_1) + u_2 + (e_{i2} - e_{i1})\]

\[ y_{ik} \quad \text{Student Score in Grade} \ k \]
\[ \mu_k \quad \text{District Mean for Grade} \ k \ (\text{FIXED}) \]
\[ u_k \quad \text{Teacher "Effect"} \ (\text{RANDOM}) \]
\[ e_{ik} \quad \text{residual} \]
Layered Model (cont’d)

- Teacher effect persists
- Teacher effect is common to all students (class effect?)
- Not necessary to include student variable since each student serves as control
- Gains are weakly correlated with background variables
- Model can be expanded to include student characteristics and information on other subject areas
Layered Model 2/Persistence Model
(McCaffrey et al., 2004)

\[ y_{i3} = \mu_3 + y_{i2} + u_3 + e_{i3} \]
\[ y_{i4} = \mu_4 + y_{i3} + \alpha_{43} u_3 + u_4 + e_{i4} \]
\[ y_{i5} = \mu_5 + y_{i4} + \alpha_{53} u_3 + \alpha_{54} u_4 + u_5 + e_{i5} \]

\[ y_{t}^{k} \quad \text{Student Score in Grade } k, \text{ Year } t \]
\[ \mu_{t}^{k} \quad \text{District Mean in Grade } k, \text{ Year } t \quad \text{(FIXED)} \]
\[ u_{t}^{k} \quad \text{Teacher "Effect" (RANDOM)} \]
\[ e_{t}^{k} \quad \text{error} \]
Persistence Model

- Teacher effects diminish
- Teacher effect is common to all students (class effect?)
- Choice of model (complete persistence/persistence models) an empirical question?
- Complete persistence model may be less sensitive to model misspecifications – but may be inferior if teacher effects dampen quickly.
Gain Score Model

\[ y_{ik} - y_{i,k-1} = \mu_k + \beta_t(Demo)_i + u_k + e_{it} \]

Covariate Adjustment Model

\[ y_{ik} = \mu_k + \gamma_k y_{i,k-1} + \beta_t(Demo)_i + u_k + e_{it} \]
Issues

• Well known that the two analyses may produce different results
• Lord’s Paradox
• Care must be exercised in choosing a model
Multivariate Approaches

- Previous analyses are done within years
- Longitudinal data – “beg” for multivariate approach
- Growth models spanning all the time periods can be incorporated within the multivariate framework
- Although more general, difficult; not guaranteed to improve estimation.
Issues to be considered in Value Added Modeling

- Database structure – number of cohorts, number of years, number of subjects
- Number of levels (hierarchical model)
- Use of covariates/background information
- Problems with omitted variables
- Missing data treatment
- Teacher effects - Random or Fixed
- Causal Inference
Random VS Fixed Effects

- Random effects – empirical/Pure Bayes estimation of teacher effects
- Draws strength from collateral information when sample size is not adequate
Attribution of Cause

- Most troublesome aspect
- Can Value-Added Assessment untangle teacher effects from non-educational effects?
- Are estimated teacher effects causal effects?
- While modeling is sophisticated, it cannot take into account differential interaction between teachers and students
Other Issues

- Use of achievement test scores: Not valid for assessing teacher effectiveness
- Estimates of teacher effects must be linked to other indicators of good teaching
- Effect of scaling – raw scores/IRT scaled scores lead to different estimates of teacher effects
- While vertical scales are not necessary, scores used to assess gains must be scaled.
Other Issues (Cont’d)

- Sensitivity of teacher effects to achievement test scores.
- All subject areas do not provide comparable estimates of teacher effects (Lockwood, et al. 2007)
Conclusion

• We have come a long way with respect to addressing the problem of Teacher Effectiveness. Considerable advances have been made with statistical modeling.

• The issues that have yet to be addressed are:
  Do the achievement test scores have sufficient depth and breadth to serve as indicators of teacher effectiveness?
  How can we establish the valid interpretation of teacher effectiveness measures/indices?