Evaluating Teachers: The Students Are the Test

Mark D. Reckase
Eun Hye Ham
Michigan State University
A Story

- I have been working with two economists for about four years on a project to evaluate value added models (VAMs).
  - From that research we have learned what works reasonably well.
  - We also know the challenges to the application of VAMs.
  - One challenge is the way that students are assigned to teachers and teachers are assigned to school – there is evidence that it is not a random process and that the assignment process influences results.
- I also was appointed to the Michigan Commission on Educator Effectiveness.
  - We were supposed to devise the educator evaluation system for the state.
  - This was under constraints imposed by the state legislature.
A Story – Part 2

• One day at a meeting of the MCEE, I was sitting next to Joseph Martineau, the head of our state testing office.
  • We were taking about the challenges of using VAMs as part of a system to evaluate teachers.
  • The issue of the assignment of students to teachers to schools came up. We were trying to figure out how to deal with the assignment problem.
  • We talked about how it would be nice if all teachers taught the same students or equivalent groups of students – then an accurate evaluation would be much easier to accomplish.
  • This is like giving all of the teachers the same test.
A Story – Part 3

- In the past, I have done work on adaptive testing and it occurred to me that all persons do not need to take the same test to compare their performance.
- Computerized adaptive tests administer different sets of items to each person, yet the results are on the same reporting score scale.
- This works because the characteristics of test items are on the same scale as the proficiency of people, and the items are calibrated onto that scale before the test is administered. This is an application of item response theory (IRT).
- Can the same idea be used for evaluating teachers using their students as test items?
During the rest of the meeting and the next few days, we thought about this problem.

- A scale was needed for calibrating the students (the test items) and that could be used as the reporting score for teachers.
- The students would need to be calibrated (item calibration).
- The students would need to be scored like test items.
- A method need to be devised to estimate the capability of the teachers on the reporting score scale.

Over the next few weeks we devised such a procedure.

- It was presented at a meeting at the University of Maryland and is due to appear as a book chapter.
- Eun Hye Ham did a more formal evaluation using state data as her doctoral dissertation.
References


Conceptual Framework

• Current educational policy is built around goal of helping students reach educational goals specified by the states.
  • This goal is generally given a label related to performance on a test that is designed to match the educational goals.
  • The label is often “Proficient”.
• School systems are often evaluated by computing the proportion of students who reach the Proficient level.
• Teachers are not usually evaluated using the “Proficient” criterion because students differ on the level of challenge they pose to reaching Proficient.
Background

- Most teacher evaluation procedures based on test scores do not use the concept of Proficiency.
  - Instead, they attribute the difference in observed student performance and predicted student performance based on previous performance and other variables as the effect of the teacher.
  - The results are usually presented in a norm referenced way showing which teachers are above or below average in the difference between observed and predicted performance.
  - But, a teacher could be quite strong working with underachieving students and still have an average that is below the average for most teachers.
  - Or, a teacher can have all students above Proficient but have low average difference between observed and predicted performance.
Background

- It would seem desirable that:
  - The evaluations of teachers be related to the policy requirements of meeting the proficiency standard.
  - The evaluations of teachers take into account the level of challenge posed by working with students with different characteristics.
  - The amount of data required to do the analysis not be burdensome.
- The procedure proposed here is designed to meet these goals.
The Educator Response Function

- The educator response function is a mathematical model that relates the capabilities of teachers and the level of challenge of students to the probability that the combination of teacher and student will reach the Proficient level specified by the state department of education.
  - This model assumes that “teaching capability” is a hypothetical construct and that teachers vary on this construct.
  - The assumption is that teachers vary on “teaching capability” and their level on the construct is the Educator Performance Level (EPL).
  - The goal of a teacher evaluation is to determine the location of the teacher on the teaching capability construct yielding a value for the EPL.
Challenge Index

- A second component of the model is the amount of challenge posed by a student when working with them to reach the Proficient level.
  - The amount of challenge is indicated by a point on the hypothetical construct.
  - The quantification of the point is called the Challenge Index for the student.
  - The location on the hypothetical construct is determined through the use of observable indicators such as
    - (a) previous year’s achievement,
    - (b) attendance record,
    - (c) home language different than the language of instruction,
    - (d) presence of disabilities,
    - (e) low SES level,
    - (f) educational level of parents, etc.
Estimating the Challenge Index

- This is the process of calibrating students (test items).
- **Approach 1**: Related to current VAM approaches
  - Using the **previous cohort of students**, predict the performance in the target grade $G$ from the indicator variables.
  - The predicted level of performance is centered around 0 and then multiplied by -1 to reverse the scale. High values mean high challenge and low values are low challenge.
  - Determine the point on the predicted scale that is equivalent to the Proficient standard. Set that to a fixed value such as 100. Students above 100 are predicted to not meet the Proficient standard.
- **Approach 2**: Direct application of IRT concepts
  - Calibrate the indicator variables as items using an IRT model and estimate a value on the IRT scale for each student in the current cohort.
The conceptual framework for the evaluation of teachers is to evaluate them relative to the CI for students that they can help to be proficient.

- A teacher that can help high CI students be proficient is very good teacher.
- If a teacher cannot help low CI students reach proficiency, they are not very good.
- Students are considered as test items and the CI value is the difficulty index for a student.
- The EPL for a teacher is determined from CI levels for students that reach proficiency. Reaching proficiency is getting the item (student) correct.
Estimating the EPL

- Students receive a code of 1 or 0 depending on whether they are proficient or not. These are considered as scores for the students as test items.
- The relationship between EPL and student performance is assumed to follow a two-parameter logistic IRT model in the form of a person characteristic curve:

\[
P(s_{ij} = 1|EPL_j, CI_i, D_j) = \frac{e^{D_j(EPL_j - CI_i)}}{1 + e^{D_j(EPL_j - CI_i)}}
\]

where \( s_{ij} \) is the performance level of Student \( i \) working with Teacher \( j \),
\( EPL_j \) is the Educator Performance Level for Teacher \( j \),
\( CI_i \) is the Challenge Index for Student \( i \),
\( D_j \) is the slope parameter for Teacher \( j \),
and \( e \) is the mathematical constant, 2.718282… .
Estimating the EPL

- The students assigned to a teacher make up the items on a test.
- The proficiency levels are the scores on the items (students).
- Using IRT estimation technology, the EPL for a teacher is the maximum likelihood estimate the location of the teacher on the CI scale that maximizes the pattern of student performance given the CI levels of the students.
- The information from the student proficiency levels can be used to get the standard error of the estimate of the teacher’s location on the EPL construct. Note that the EPL is computed on the CI-scale.
Example: Teacher with 44 Students

CI distribution for students assigned to the teacher.

Note that most of them are below 100. This is not a very challenging group of students.
Example: Teacher with 44 Students

Proficiency levels of students as a function of CI. Most of those with a low CI are proficient.
Estimation of the EPL for the Teacher

The two-parameter logistic model is fit to the data for the teacher.

EPL = 100

This means that the probability of this teacher helping a student with CI = 100 reach proficiency is 0.5.

Standard error is 3.9.
Example: Teacher with 42 Students

Most of these students have CI values above 100. This is a more challenging teaching assignment than the first teacher.
Example: Teacher with 42 Students

EPL estimate is 120 with a standard error of 6.1. The teacher has a higher EPL because students with high CI values were proficient. Error is larger because division is not as distinct.
Real Data Analysis

- Data – Roughly 100,000 students per grade
- Roughly 12,000 teachers
- Analyze the data using IRT calibration of indicators
- Analyze the data using linear regression calibration model.
Indicators

- Prior Math Score
- Prior Math Proficiency
- Economically Disadvantaged
- Free/Reduced Lunch
- Targeted Assistance Program
- Special Education
- Disability
- Limited English Proficiency
- Proportion of Attendance
Results

IRT analysis of indicators showed that there were not enough indicators in the state data.
Results

Linear prediction calibration worked pretty well
Results: Challenge Index Is Related to Performance

Math, Elementary (Grade 4-5)
Teacher Performance Functions

EPL = -1.04 (se = .34); S = -2.85; N = 88

EPL = -0.61 (se = .28); S = -2.85; N = 104

EPL = 0.13 (se = .4); S = -2.92; N = 64

EPL = -0.24 (se = .21); S = -2.59; N = 88

EPL = 0.06 (se = .2); S = -2.59; N = 104

EPL = 0.27 (se = .28); S = -2.74; N = 64
Results for All Teachers

Grade 4

Grade 5

Grade 6

Grade 7
Comparison with VAM Results

- Eun Hye Ham also did a traditional VAM analysis for the data from the state.
  - This was based on a linear regression based VAM model with the teachers included as dummy variables.
  - The coefficient on the teacher dummy variable is the teacher effect.
- The results from the VAM analysis were correlated with the results from the new procedure – how high should the correlation be?
- The rank order correlation was .81 for mathematics and .59 for reading.
Results

- The model works well when there is enough variation in the Challenge Index.
- The results are criterion referenced rather than norm referenced.
- The results are similar to the results from other VAM methods, but not exactly the same.
Implementation Issues

• The critical part of the method is defining the challenge index. In practice, the variables defining the challenge index should be selected in collaboration with teachers and school administrators.

• The procedure only needs the data from the previous cohort of students.

• In principle, CI values can be determined for students assigned to all teachers, but a proficiency standard is needed for all subject matter areas.
Implementation Issues

- The method may have the positive benefit of encouraging teachers to work with challenging students.
- The CI estimation procedure should be updated each year as tests and student characteristics change.
- As always, more research is needed.