Copyright
by
Christian Dennis Jackson II
2013

The Report Committee for Christian Dennis Jackson II
Certifies that this is the approved version of the following report:

Modeling Teacher Effectiveness as a Function of Student Ability

APPROVED BY

SUPERVISING COMMITTEE:

Tse-Min Lin, Supervisor

Margaret Myers

Modeling Teacher Effectiveness as a Function of Student Ability

by

Christian Dennis Jackson II, B.S.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2013

Dedicated to my wife Rachel and my mentor Dr. Ben.

Acknowledgments

I wish to thank my wife for helping me, my parents for co-financing my degree, and Ben for walking me through the valley of the shadow of R (and LaTeX). I would also like to thank Dr. Tse-Min Lin for supervising this paper and Dr. Margaret Myers for providing valuable feedback and comments.

Modeling Teacher Effectiveness as a Function of Student Ability

Christian Dennis Jackson II, M.S.Stat.
The University of Texas at Austin, 2013

Supervisor: Tse-Min Lin

In 2010, the L.A. Times newspaper used the test results of Los Angeles County elementary students to assess and rank the elementary teachers. They then published the results on their website. Publicly ranking teachers in this manner has important implications on the careers of the teachers being ranked. It is, therefore, important that any model claiming to rank teachers be as accurate as possible. It seems plausible that a teacher's ability to help a student depends upon that student's prior academic ability. Some teachers might be better at teaching gifted students while others might be better at teaching remedial students. The L.A. Times did not account for this in their model. This paper looks at the results of allowing teacher effect to vary with prior student ability and how that interaction affects the relative rankings of the individual teachers. To assess this, the same Value-Added model the L.A. Times used is employed, with the exception that teacher effect is allowed

to vary with the prior abilities of the students. New teacher ranks are then calculated and compared with the ranks calculated by the L.A. Times. The results of this analysis show a relatively small number of rank changes between the two models. In general, allowing teacher effect to vary results in a 5% to 12% change in the rankings of both the Math and Reading teachers relative to the L.A Times model. Other research on the same data has resulted in a 20% to 55% change in the rankings of the Math teachers and a 40% to 65% change in the rankings of the Reading teachers relative to the L.A. Times model. Although ranking teachers is a popular idea for determining the distribution of funding, the model shown in this paper as well as the other models reviewed, illustrate that a change in the model results in a change in the rankings of the teachers. A model that allows teacher effect to vary with prior student ability results in a better model fit than a model that does not. Whether or not this is a good thing is hard to say. Two examples are provided in this paper. One shows a teacher whose rank appears to be artificially inflated by this model and the other shows a teacher whose rank appears to be artificially lowered by this method. Although the fit of the model proposed by this paper is better than the model used by the L.A. Times, it does not result in radical changes in the rankings of the teachers. Rather, it seems that teacher rankings are sensitive to the particular model used and there are countless numbers of valid models. For this reason it is not wise to release such sensitive information to the public. It is probably true that the weak teachers are ranked relatively low in this analysis and that the truly good teachers are ranked relatively

high. However, these rankings should only be used as one part of a larger metric to rank teachers and too much weight should not be placed on them for the purposes of rewarding or penalizing teachers due to the sensitivity of the model specification.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Education is universally seen as critical to continued economic competitiveness. How to most effectively achieve that education, however, is a matter of intense debate. For this reason, education policy is often a contentious political topic. The policy most recently enacted in the United States is the Race to the Top (RTTT) program. It was enacted by President Obama in 2009 and is funded by the American Recovery and Reinvestment Act of 2009, a government response to the economic recession that occurred in the latter half of the decade spanning 2001-2010. The RTTT program is a national competition for federal education grants that takes place both at the state level and at the local level. The various education departments earn points by complying with the policies set forth in RTTT and federal grant money is awarded to those with the highest point totals.

1.1 The Race To The Top

RTTT ties $4.35 billion in incentive funding to a state's implementation of specific education policies. The policies set forth in RTTT are split into weighted groups. In each of these weighted groups, the policies themselves are

also weighted. Each state receives a score based on whether, or to what degree, it meets each of the various policies set forth the in RTTT. An education district's level of compliance with RTTT policies is thus assessed and scored. The aggregate score is then tallied and higher-scoring states receive more of the incentive funding than lower scoring states.

Of the 500 available points to be won, the heaviest weighted category, with 138 points, is the "Great Teachers and Leaders" category. Within this category, the heaviest weighted policy, with 58 points, is "Improving teacher and principal effectiveness based on performance" [10]. In other words, performance-based teacher evaluation is the heaviest-weighted policy in the heaviest-weighted category in the RTTT. The implementation of policies that allow school districts to measure teacher and principal effectiveness is a big step in collecting enough points to earn the grant money.

This reflects the growing demand in United States education policy to evaluate the effect of individual teachers on student performance. The demand can be seen in the language written into the RTTT program. It can also be seen in individual state legislation, such as Colorado Senate Bill 191 (SB-191), the "Educator Effectiveness Bill," mandates:

> that at least 50 percent of a teacher's evaluation be determined by the academic growth of the teacher's students... The new system will be piloted in 2012-2013, implemented statewide in 2013-2014 and finalized in 2014-2015 [9].

States and districts are moving fast to find ways to evaluate the effects

of their educators and schools. This can, in part, be attributed to the grant money being offered by RTTT. However, RTTT is only an incentive program. It is not intended to be a permanent source of funds for states and districts but to provide financial incentive that spurs rapid change in highly bureaucratic and slow-moving departments of education. This need for rapid change is seen as necessary in an age of limited educational resources. Higher levels of accountability and performance-based analysis are seen as effective tools that can help federal, state, and local communities wisely balance education budgets.

1.2 Value Added Models

Pursuant to these goals, the L.A. Times funded and published a controversial analysis of teacher and school effects in the Los Angeles Unified School District (LAUSD) in 2010. Similarly, in February of 2012, the New York City Department of Education publicly released their own teacher evaluations. In their evaluations, Value-Added (VA) Models were used to isolate teacher effect in the classroom. Once the effects of individual teachers were isolated the teachers were ranked with respect to each other. In the case of the LAUSD these teacher rankings were posted publicly on the L.A. Times website and resulted in a large amount of controversy due to the professional impact of being near the bottom of the rankings.

Evaluation systems are inevitably controversial. Once metrics are proposed that purport to rank teachers, the validity of the metric is immedi-

ately called into question. Not only do teacher evaluations impact the limited amount of federal funding available to the state and local school districts, they have the potential to affect the very teachers that are being ranked. As such, it is important that any system claiming to measure teacher effect be as accurate as possible.

There is evidence that VA models can identify effective teachers. This evidence is both experimental and non-experimental [5][6]. According to the Economic Policy Insitute (EPI), "approaches that measure growth using 'value-added modeling' (VAM) are fairer comparisons of teachers than judgments based on their students test scores at a single point in time or comparisons of student cohorts that involve different students at two points in time [1]."

While it has been shown that VA models can effectively isolate the educational effect of teachers, the literature addresses certain limitations to this method. For example, VA models have been shown to be sensitive to type of achievement measure, student sorting, and test scaling [2][7][8]. Indeed, the EPI goes on to state that "there is broad agreement among statisticians, psychometricians, and economists that student test scores alone are not sufficiently reliable and valid indicators of teacher effectiveness to be used in high-stakes personnel decisions, even when the most sophisticated statistical applications such as value-added modeling are employed [1]." While it is true that VA models are not sufficient indicators of teacher effectiveness, it is also true that they continue to be used in an attempt to assess quality teachers. Therefore, is it imperative that any VA models used be as accurate and unbi-

ased as possible.

1.3 The L.A. Times Analysis of LAUSD Teachers

In the L.A. Times analysis of Los Angeles area teachers, the teachers were ranked using VA modeling. In brief, the model predicted current year test scores using test scores from the previous year. Deviations in actual test scores from the predicted test scores were attributed to the teacher rather than other potential explanatory factors such as family crisis or outside tutoring [4]. Many models, including the L.A. Times model, posit the effect of a teacher as one that is constant irrespective of student traits. In current models, teachers are assumed to affect prior test scores by a constant factor plus or minus some specific teacher effect. In other words, if a teacher received a class of students that all scored in the 30th percentile on standardized prior-year exit exams and those same students all scored in the 40th percentile after being under the tutelage of that teacher for the school year, then that improvement can, in part, be attributed to the quality of that teacher. This result is shown in Figure 1.1.

After ranking the teachers using a VA model, the L.A. Times ranked the teachers by dividing them into quintiles. The first quintile represents the top 20% of the ranked teachers. The second quintile represents the next 20%. This continues to the fifth quintile where the bottom 20% of the ranked teachers, the ones who ranked lowest, can be found. In the L.A. Times model, each teacher was assigned a regression line based upon the prior- and current-

year performance of their students. The lines all have the same slope and teacher rankings were assigned by the relative height at which the lines cross the vertical axis of the graph. Higher lines represent higher ranked teachers.

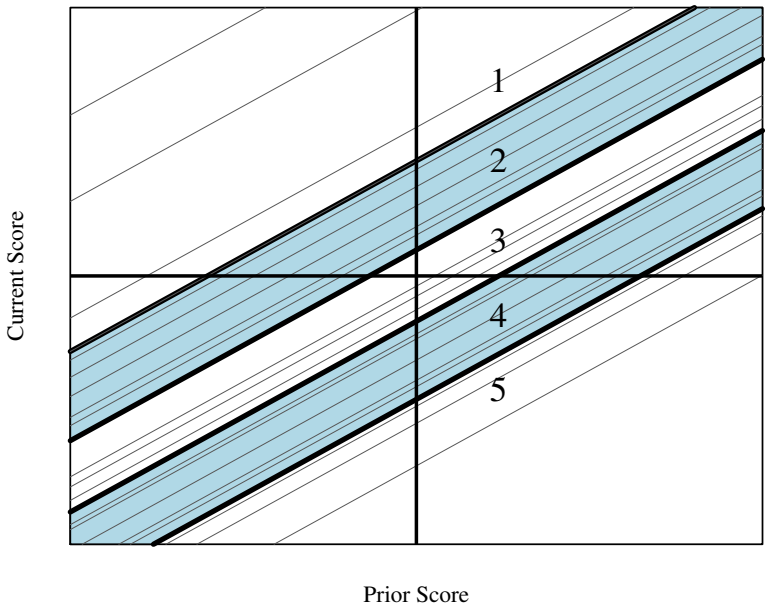


Figure 1.1: L.A. Times model showing quintiles. A sample of 20 regression lines are plotted. The five quintiles are separated with bold lines and shading. They are numbered from 1 to 5. The L.A. Times derived fixed-slope regression lines for each of the approximately 11,000 teachers in the LAUSD. Only the y-intercept of the regression line was allowed to vary from teacher to teacher. These varying y-intercepts were then sorted and separated into quintiles. In this sample for example, the top-most line would be the highest ranked teacher and the bottom-most line would be the lowest ranked teacher.

1.4 The Goal of this Report

The current VA models attempt to account for student-specific variables but fail to acknowledge that some teachers may be particularly good at teaching certain types of students. For example, it is not hard to believe that a certain teacher will be better at teaching remedial students than advanced students, or vice-versa. It is also not hard to believe that a teacher who speaks English as a second language (or who speaks a second language in general) could be more effective with students whose first language is not English. Following this line of reasoning, it is conceivable that a teacher's performance is related to the prior academic ability of their students.

This goal of this paper is to address this alternative approach to modeling. The LAUSD data set used by the L.A. Times and by Briggs and Domingue [4] is used to examine the results of allowing the measured teacher effect to vary along with prior students ability rather than hold it constant as has been done in previous studies. This allows average teacher effectiveness to be modeled along with teacher effectiveness relative to the abilities of their students. This paper posits that the effect or value of a teacher is dependent on student ability and seeks to quantify the variability in ability-based teacher effect modeling.

Chapter 2

A Random Slope Model

The functional result of allowing the prior academic abilities of students to be a factor in assessing a teacher's performance is that the slopes of the lines discussed in Figure 1.1 are no longer constrained to be constant for all teachers. The graphs in this report show prior student scores on the x-axis and current student scores are shown on the y-axis. Each teacher has an individual regression line that predicts current-year scores for any given prior-year scores of their students. In the L.A. Times analysis, every teacher was predicted to improve prior-year scores by the same amount. For this reason, all of the regression lines have the same slope (see Figure 1.1). In this report, the amount a teacher is predicted to help his or her students is allowed to vary from teacher to teacher based on the prior abilities of the students. Therefore, each teacher's regression line will have a different slope when it is modeled. The ranking of teachers is still performed by comparing relative height on the vertical axis in the same manner as the L.A. Times analysis. However, this ranking is now performed at three separate points on the students' prior-ability scale. The primary question is whether allowing these teacher slopes to vary results in significant changes to the L.A. Times rankings.

2.1 The Status Quo Line

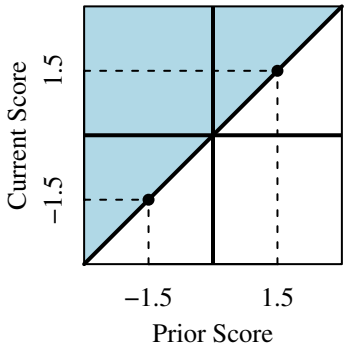


Figure 2.1: Status Quo Line ($y = x$)

Figure 2.1 displays a regression line for a hypothetical teacher. A student who enters this teacher's classroom with a test score that is 1.5 standard deviations above the mean is predicted to leave this teacher's classroom scoring 1.5 standard deviations above the mean. A student who scores -1.5 standard deviations below the mean is predicted to leave this teacher's classroom scoring -1.5 standard deviations below the mean. In this example the teacher is neither predicted to increase nor predicted to decrease a student's test scores over the course of a school year. The students have maintained their level of performance, relative to their peers, over the course of the school year. For this reason, such a line will be referred to as the "status quo" line for the remainder of this report.

In contrast to this perfectly average teacher, one could imagine two additional types of teachers. One is a low-value teacher who allows students to fall behind their peers. Students in this teacher's class would leave with lower

scores than those with which they entered. The other is a high-value teacher that helps students outperform their peers over the course of an academic year.

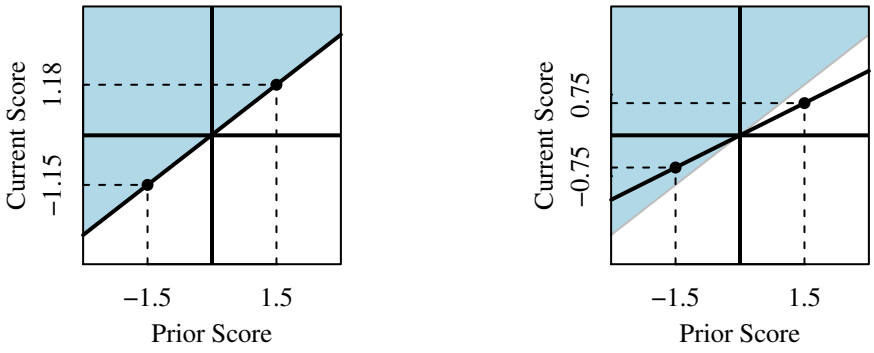
The case shown in Figure 2.1 is an example of a perfectly average teacher and students who perform the same on every test. In reality, there is variation within each student's test scores. A student who performs 1.5 standard deviations above the mean on one test will probably not perform 1.5 standard deviations above the mean on the following test. The student will almost certainly score slightly higher or lower on the next test. In other words, a student will tend to regress toward his or her mean test score. For this reason the ideal model in Figure 2.1 is incorrect. A true status quo line must take into account the variation inherent within individual test scores. In the L.A. Times data, the true status quo line of interest is found by regressing all current year scores on the prior year scores. The result of this regression for the math and reading scores is given in Table 2.1. These are the true slopes and intercepts of the respective status quo lines.

	Intercept	Slope
Math	0.01	0.78
Reading	0.01	0.82

Table 2.1: The true status quo line intercepts and slopes for Math and Reading scores.

According to the results of the regression, a student entering grade 4 with a Math score that is 1 standard deviation above the mean should expect a Math score that is about 0.78 standard deviations above the mean at the end of grade 4, assuming their teacher's regression line is the same as the status

quo line. Any change in this prediction for an individual teacher will alter the slope of the regression line accordingly. The true status quo line for Math achievement can be seen in Figure 2.2a.



(a) Math Status Quo Line
$y \approx 0.78x$

(b) Hypothetical Teacher
Slope < 0.78

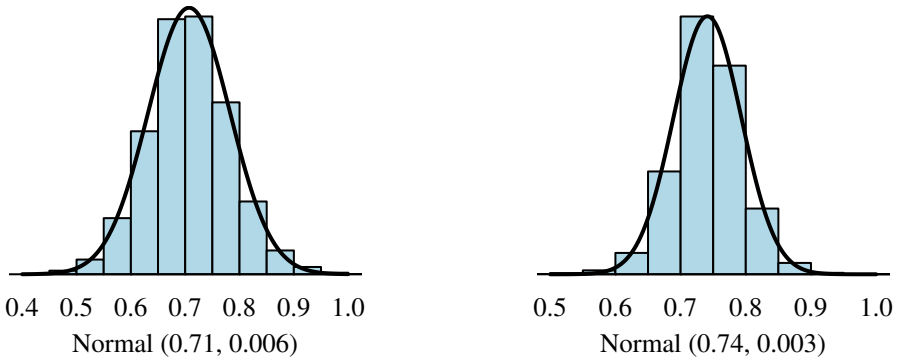
Figure 2.2: Consequences of Random Slopes

Once the slopes of individual teacher's regression lines are allowed to vary according to prior student ability there are a few different theoretical situations that can occur. On one hand, the slope of an individual teacher's regression line could be less than the slope of the status quo line. This case is modeled in Figure 2.2b.[†] On the other hand, the slope could be greater that that of the status quo line. In the former case, the teacher is predicted, based on past student performance, to show an improvement in scores for

[†]Note how the line in Figure 2.2b lies partly in the shaded area and partly in the unshaded area. This distinction in shading will remain throughout relevant figures in this report. The portion of the line lying in the shaded region signifies the prior scores that a teacher is predicted to improve upon; and conversely for the portion of the line lying in the unshaded region.

low-achieving students and a decline in scores for students high-achieving students. The result shown in Figure 2.2b could be achieved in one of three ways: the teacher is simultaneously good at teaching low-achieving students and bad at teaching high-achieving students or, as is more likely, the teacher mainly teaches either low-achieving students or high-achieving students and is, therefore, correspondingly effective or non-effective at their job. Conversely, in the case of a teacher whose regression line is steeper than the status quo line, we have the opposite situation. In this case a teacher has positive score improvement for high-achieving students and negative score change for low-achieving students. This result could be achieved in analogous ways to the result shown in Figure 2.2b.

2.2 The Distribution of Random Slopes



(a) Math **(b)** Reading

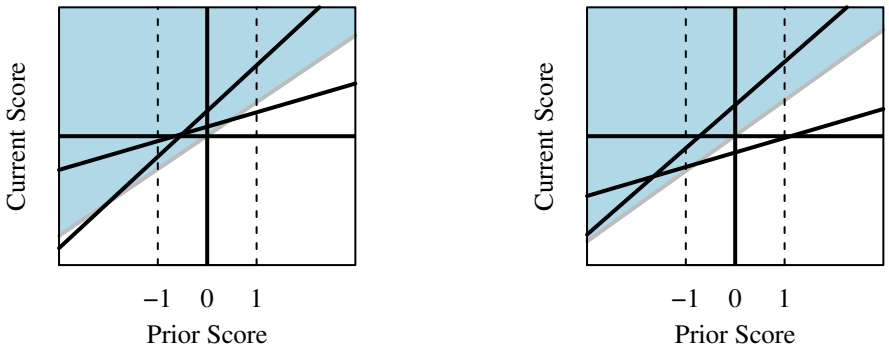
Figure 2.3: Distribution of Slopes

The first step in analyzing a model with differing slopes is to get a

sense of the range and distribution of the slopes themselves. The distribution of slopes for math teachers when slopes are allowed to vary is normal with a mean of 0.77 and a variance of 0.006. The distribution of slopes for reading teachers is normal with a mean of 0.74 and a variance of 0.003. The distributions are shown in Figure 2.3. For visual reference, the maximum and minimum slopes for both the Math and Reading scores are shown in Figure 2.4. It is important to keep in mind that these lines do not represent the "best" and "worst" teachers in the data set. The teachers are not ranked on the value of their regression slopes alone. This is a visual reference to aid in interpreting the slope distribution graphs in Figure 2.3 and the consequences of implementing a random-slope model. Figure 2.4a shows regression lines for two teachers. Note how one teacher's line is above the other at $x = -1$. This results in the teacher whose line is above the other being ranked higher than the teacher whose line is below. If, instead, the relationship between the two teachers is considered at $x = 1$ we come to a different conclusion in terms of teacher rankings. The teacher who was ranked higher at $x = -1$ is now ranked lower.

2.3 Teacher Ranks

The L.A. Times allowed the intercepts of all teachers in the model to vary according to a normal distribution. This means there are as many different intercepts as there are teachers. These individual intercepts were sorted from highest to lowest and then split into five quintiles. The top fifth of the ordered teachers were ranked in the first quintile, the second fifth were



(a) Math **(b)** Reading

Figure 2.4: Maximum and Minimum Slopes

ranked in the second quintile and so on. The same ranking method was used for the random-slopes model.

After creating a VA model that allows the slope for each teacher to vary according to the prior ability of their students, the teachers are ranked in the same manner as they were ranked by the L.A. Times but in three different locations. They are ranked at the intercept ($x = 0$), at $x = 1$, and at $x = -1$ according to the same procedure used by the L.A. Times analysis; the lines are ranked in terms of relative height at these points. These ranks were then compared to the L.A. Times ranks. Figure 2.4 shows how, with varying slopes, two teachers can be ranked much closer together at one location than at another. The results of these comparisons are shown numerically in Tables 2.2 - 2.4 and as percents in Tables 2.5 - 2.7.

Table 2.2a shows, for example, that at $x = 0$ 2070 of the 2177 teachers who were ranked in the first (highest) quintile by the L.A. Times were also

ranked in the first quintile when assessed using the random slopes model at. Additionally, the remaining 107 teachers the L.A. Times ranked in the first quintile were ranked in the second quintile in the random slopes model. In a situation where these rankings are used to determine professional advancement, those 107 teachers would prefer the L.A. Times model to the random slopes model.

L.A. Times	Random Slope Ranks at $x = 0$				
	1	2	3	4	5
1	2070	107	0	0	0
2	107	1942	128	0	0
3	0	128	1920	130	0
4	0	0	128	1942	107
5	0	0	2	105	2071

(a) Math

L.A. Times	Random Slope Ranks at $x = 0$				
	1	2	3	4	5
1	1947	111	2	0	0
2	107	1811	142	1	0
3	4	131	1806	119	1
4	2	6	110	1820	123
5	0	2	1	121	1937

(b) Reading

Table 2.2: Numeric Rank Changes at $x = 0$

L.A. Times	Random Slope Ranks at $x = 1$				
	1	2	3	4	5
1	1947	227	3	0	0
2	226	1640	308	2	1
3	4	306	1551	310	7
4	0	4	316	1619	238
5	0	0	0	246	1932

(a) Math

L.A. Times	Random Slope Ranks at $x = 1$				
	1	2	3	4	5
1	1807	240	11	1	1
2	243	1437	362	18	1
3	10	369	1315	352	15
4	0	13	363	1376	309
5	0	2	10	314	1735

(b) Reading

Table 2.3: Numeric Rank Changes at $x = 1$

L.A. Times	Random Slope Ranks at $x = -1$				
	1	2	3	4	5
1	1898	265	13	1	0
2	259	1558	338	22	0
3	18	324	1482	336	18
4	1	27	330	1566	253
5	1	3	15	252	1907

(a) Math

L.A. Times	Random Slope Ranks at $x = -1$				
	1	2	3	4	5
1	1741	290	25	4	0
2	288	1370	354	47	2
3	21	367	1318	340	15
4	8	27	340	1421	265
5	2	7	24	249	1779

(b) Reading

Table 2.4: Numeric Rank Changes at $x = -1$

L.A.	Random Slope Ranks at $x = 0$				
Times	1	2	3	4	5
1	95.1	4.9	0	0	0
2	4.9	89.2	5.9	0	0
3	0	5.9	88.2	6.0	0
4	0	0	5.9	89.2	4.9
5	0	0	0.1	4.8	95.1

(a) Math

L.A.	Random Slope Ranks at $x = 0$				
Times	1	2	3	4	5
1	94.5	5.4	0.1	0.0	0.0
2	5.2	87.9	6.9	0.0	0.0
3	0.2	6.4	87.6	5.8	0.0
4	0.1	0.3	5.3	88.3	6.0
5	0.0	0.1	0.0	5.9	94.0

(b) Reading

Table 2.5: Percent Rank Changes at $x = 0$

L.A.	Random Slope Ranks at $x = 1$				
Times	1	2	3	4	5
1	89.4	10.4	0.1	0	0
2	10.4	75.3	14.1	0.1	0
3	0.2	14.0	71.2	14.2	0.3
4	0	0.2	14.5	74.4	10.9
5	0	0	0	11.3	88.7

(a) Math

L.A.	Random Slope Ranks at $x = 1$				
Times	1	2	3	4	5
1	87.7	11.7	0.5	0.0	0.0
2	11.8	69.7	17.6	0.9	0.0
3	0.5	17.9	63.8	17.1	0.7
4	0.0	0.6	17.6	66.8	15.0
5	0.0	0.1	0.5	15.2	84.2

(b) Reading

Table 2.6: Percent Rank Changes at $x = 1$

L.A.	Random Slope Ranks at $x = -1$				
Times	1	2	3	4	5
1	87.2	12.2	0.6	0	0
2	11.9	71.6	15.5	1.0	0
3	0.8	14.9	68.0	15.4	0.8
4	0	1.2	15.2	71.9	11.6
5	0	0.1	0.7	11.6	87.6

(a) Math

L.A.	Random Slope Ranks at $x = -1$				
Times	1	2	3	4	5
1	84.5	14.1	1.2	0.2	0.0
2	14.0	66.5	17.2	2.3	0.1
3	1.0	17.8	63.9	16.5	0.7
4	0.4	1.3	16.5	68.9	12.9
5	0.1	0.3	1.2	12.1	86.3

(b) Reading

Table 2.7: Percent Rank Changes at $x = -1$

2.4 A Comparison with Results from Briggs and Domingue

According to Briggs and Domingue, the L.A. Times model

> produces biased estimates because it omits variables that are associated with both student performance and how students and teachers are assigned to one another [4].

Briggs and Domingue propose a new model they call the altVAM model. It is an attempt to correct the perceived sources of bias in the L.A. Times analysis. It is important to note that the altVAM model and the random-slopes model proposed in this paper are not the same. The altVAM model adds the following variables to the L.A. Times model: student test scores from grades 2 and 3, the mean of grade 4 scores in each grade 5 classroom, and an indication of how similar a student's school is to other schools in Los Angeles County. However, while the random-slopes model and the altVAM model are not the same, both models have been compared to the L.A. Times results and the subsequent change in teacher rankings have been recorded for each. Strictly in terms of quantifying the amount of change in teacher rankings found by the random-slopes model, it is meaningful to look back at the change in teacher rankings that occurred in the altVAM model. The results are shown in Table 2.8. Since Briggs and Domingue ranked the teachers at the intercept, it is most accurate to compare the results in Table 2.8 with the results in Table 2.5. The altVAM model results in significantly more ranking changes in both the Math and Reading teachers than the random-slopes model. In the altVAM analysis the Reading teachers display noticeably more volatility than do their Math

L.A.	Briggs & Domingue Ranks at $x = 0$				
Times	1	2	3	4	5
1	79.5	18.6	1.8	0.2	0
2	19.3	52.8	22.6	4.8	0.5
3	1.2	25.8	45.6	23.4	4.1
4	0	2.9	27.9	50.1	19.2
5	0	0	2.1	21.6	76.3

(a) Math

L.A.	Briggs & Domingue Ranks at $x = 0$				
Times	1	2	3	4	5
1	59.5	21.9	10.9	6.4	1.4
2	31.4	32.9	18.2	10.8	6.7
3	7.3	40.0	35.3	17.5	9.1
4	1.8	13.5	26.4	39.8	18.5
5	0	0.8	9.2	25.6	64.4

(b) Reading

Table 2.8: Briggs & Domingue Percent Rank Changes at $x = 0$

counterparts whereas, in the random-slopes analysis, the Math and Reading teachers show similarly low-levels of change across rankings. This does not imply that any of these models are better than the other. It merely implies that just allowing the slopes to vary by teacher does not cause as much change in etcher rankings as other changes to the model.

2.5 Two Examples

Since the L.A. Times ranked fixed-slope (parallel) lines at the y-intercept ($x = 0$), these rankings are the same at all possible x values. This is not necessarily the case in the random slopes model. As can be seen in Figure 2.4a the higher line at $x = -1$ is, in fact, the lower line at $x = 1$. It clear from this that the random-slope rankings can be different at different points on the x-axis. This suggests that the teacher rankings reported by the L.A. Times are not fixed but could vary based on the prior ability of a student. The three points analyzed in this report are $x \in \{-1, 0, 1\}$ as shown in Figure 2.4. These are the three student ability levels at which the teachers are ranked in the random-slope model. Two individual teachers who showed the largest ranking

change in the random-slopes model are provided in this section as examples of how the two models provide different rankings to the same teacher.

2.5.1 Teacher #67721377681

The reading scores provide an example of the difference between the fixed slope model and the random slope model. Table 2.4b shows there are two teachers the L.A. Times ranked in the fifth (last) quintile that are ranked in the first quintile in the random slopes model at $x = -1$. As can be seen in Table 2.7b, these teachers only make up 0.1% of the teachers ranked in the fifth quintile. However, this outlying example will provide the greatest visual contrast. Of the two teachers, teacher #67721377681 was chosen. The results are shown in Figure 2.5. As can be seen in the figure, this teacher largely receives students whose prior test scores are above the 50^{th} percentile (fall to the right of $x = 0$). Not only that, but this teacher appears to slightly decrease the students' test scores more than would be expected by an average teacher (a slight majority of the points lie in the unshaded area).

As stated earlier, this teacher ranked in the fifth quintile in the L.A. Times model and it is not hard to see why. The y-intercept of the solid L.A. Times line is below zero and the y-intercepts are what the L.A. Times used to rank the teachers. In contrast, the dashed line indicates the best fit line when slopes are allowed to vary. This allows the line to more accurately reflect the nature of the test scores for this particular teacher's students. The slight decrease in the test scores of the students is noticeable. This decrease is what

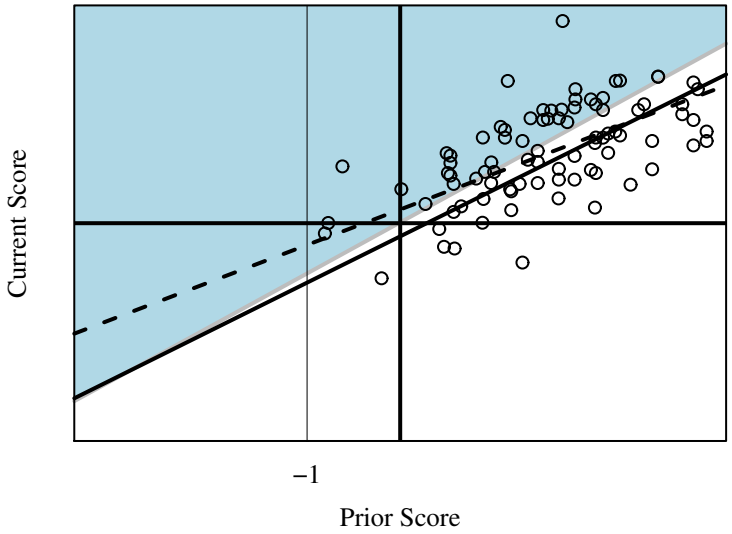


Figure 2.5: Teacher #67721377681 - In this figure, the solid line is the fixed-slope regression line predicted by the L.A. Times and the dashed line is the regression line predicted by the random-slope model.

causes the fixed-slope regression line to have a negative y-intercept and a low overall ranking. When the slope of the regression line is allowed to vary, it more accurately describes the trend of the points using a flatter regression line. The resulting difference in heights between the two lines at $x = -1$ is significant enough to carry this teacher from the last quintile to the first quintile. What we see here is probably a teacher who is a bit below average at teaching high-perfoming students and who was rewarded by the random slope model in the relative rankings at both $x = 0$ and $x = -1$.

Less can be said about this teacher's ability with low-performing students since that is outside the range of the data. This teacher is ranked in the first quintile of the random-slope model at the $x = -1$ level but there is no

reason to believe this is true. The nature of the data is probably artificially inflating the prediction of any results for x values much less than zero. This is a fundamental problem of extrapolation in regression.

2.5.2 Teacher #42559989686

With Teacher #42559989686 we have the opposite circumstance. Again, a teacher has been selected from an extreme group for visual clarity. This teacher, ranked in the first quintile by the L.A. Times, is ranked in the fifth quintile by the random-slope model at $x = 1$. The data is shown in Figure 2.6. The designations of the lines remain the same as in Figure 2.5: the solid line represents the fixed-slope line of best fit and the dashed line represents the random-slope line of best fit.

The solid, fixed-slope line again fits the data fairly well assuming the only thing that can be done to improve fit is to move the line either up or down. However, the dashed, random-slope line truly fits the data much better as its slope is allowed to vary from teacher to teacher. In this particular case, the random-slope line indicates that this teacher is predicted to help low-achieving students. What we see in this case is a teacher who is a well above average at teaching low-perfoming students. This can be seen explicitly by noting the prevalence of data points that lie in the shaded region. The resulting slope and y-intercept for the random-slope line is, though, enough to rank this teacher in the fifth quintile at $x = 1$.

While this teacher ranks in the random-slope fifth quintile at $x = 1$,

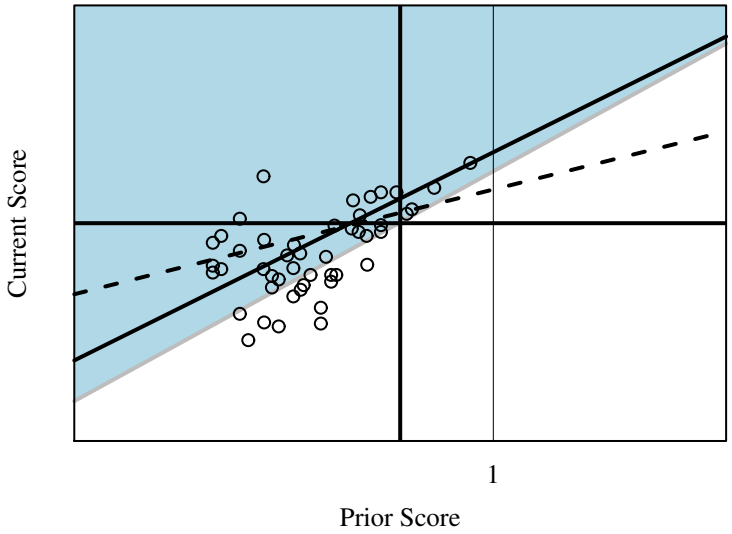


Figure 2.6: Teacher #42559989686 - In this figure, the solid line is the regression line predicted by the L.A. Times and the dashed line is the regression line predicted by the random-slope model.

they rank in the third quintile and the first quintile, respectively, at $x = 0$ and at $x = -1$. Since there is not much data to the right of $x = 0$ the teacher's ranking at $x = 1$ is suspect at best and the ranking at $x = 0$ or $x = -1$ is probably more indicative of his overall performance.

2.6 Conclusion

When comparing two models such as the L.A. Times and random-slopes models, it is useful to have a way of measuring which model fits the data better. There are various ways of doing this and the method used in this report is to compare the AIC values of each model. The AIC value is a measure of model fit. The exact value of the AIC is relatively meaningless, but the relative AIC

value of two models can be used identify which model fits the data better. A model with a lower AIC value is considered to have better fit than a model with a higher AIC value. The random-slope model discussed in this paper has a lower AIC value and thus a better fit than the L.A. Times model for both Math and Reading. This is not surprising since more information is being used to predict student scores in the random-slope model. It does, however, provide empirical proof that the random-slope model does a better job of fitting the data.

	Math Model AIC Value	Reading Model AIC Value
L.A.Times	$1,228,515$	$1,343,828$
Random Slopes	$1,220,612$	$1,086,471$

Table 2.9: The AIC values for the various models mentioned in this Chapter. Low AIC values are considered to show better model fit. Full fit statistics for all four models can be found in the Appendix.

This chapter examined what happens when the L.A. Times model was changed to account for prior student ability in the teacher rankings. The differences in the two groups of rankings are shown numerically in Tables 2.2-2.4 and using percents in Tables 2.5-2.7. Quantifying these changes was done by comparing the results to the ranking changes found by Briggs and Domingue using their altVAM model. The ranking changes found using the altVAM model can be found in Table 2.8. Comparing Table 2.5 with Table 2.8 provides the most accurate comparison in ranking changes between the two models.

As can be readily seen, the altVAM model specified by Briggs and

Domingue produces a greater percentage of ranking changes for both Math and Reading. Based on this ad hoc comparison, it appears that the ranking changes noted in the random-slope model are not as extreme as they could be. What this chapter has shown is that comparing two teachers who primarily teach students of different average prior abilities can be misleading in certain circumstances.

Chapter 3

Adding Teacher Level Variables

Allowing the slopes of the prediction lines for individual teachers has been shown to improve the fit of a regression model. The next question is whether there are any teacher qualities such as experience or education that further predict an improvement in test scores for students. The teacher information available in the dataset is: Education, Ethnicity, Years of Experience, and Certification status. Since the ethnicity of the students themselves is unknown, it seems unhelpful to include the ethnicity of the teacher. The other variables, however, seem plausible in helping to explain the qualities of teachers that predict success when their slopes are allowed to change according to the prior ability of their students.

	Math Model AIC Change (Relative to Baseline)	Reading Model AIC Change (Relative to Baseline)
Adding Teacher Education	−812	−700
Adding Teacher Experience	−4	+4
Adding Teacher Education and Experience	−749	−627
Adding Teacher Gender	−851	−741
Adding Teacher Education and Gender	−805	−696

Table 3.1: The AIC values for models with various teacher-level predictors. Low AIC values are indicate a better model fit. For example, in this table an AIC change of −851 indicates a better fit than −749.

3.1 Adding Teacher Education

There are seven teacher education levels in the dataset. They are given in Table 3.2. When teacher education level is added to the model, none of the

1	2	3	4	5	6
No BA	BA	BA+	MA	MA+	PhD

Table 3.2: Education levels for teachers in the L.A. Times dataset. The '+' symbol indicates the teacher possesses that degree plus at least 30 hours of additional course credit.

levels are significant predictors of future student test scores. Adding education to the model does, however, improve the fit of the model as indicated by the AIC (see Table 3.1). The AIC values for both Math and Reading drop noticeably when education is added to the model. This is to be expected though, because more available information is being used to fit the model.

3.2 Adding Teacher Experience

In the L.A. Times dataset, the values of teacher experience range from 0 years (indicating a first-year teacher) to 59 years. When teacher experience is added to the model it results in a significant coefficient for both Math and Reading scores. The coefficient for Math teacher experience is -0.0006537. This indicates that for every year of experience a teacher has, a student can expect their score to drop by 0.0006537 standard deviations, a negligible amount. The coefficient for Reading teacher experience is -0.0004673 and it has the same interpretation as the Math teacher coefficient. So, while fresher teachers

have a significant impact on test scores, that impact is almost nonexistent. This can be seen as well in the corresponding AIC value where there is only 4 points of difference between the AIC of the baseline Math and Reading models and the AIC of the models that contain teacher experience (see Figure 3.1).

3.3 Adding Both Teacher Education and Experience

When both teacher education and experience are added to the model and allowed to interact none of the resulting coefficients are significant. Furthermore, while the resulting AIC value is still lower than it is in the baseline model, it is now noticeably higher than in the model that includes education only. For this reason it appears that the best model is the one that contains only education as a teacher level predictor.

3.4 Adding Gender

The final relevant teacher-level variable is gender. When a teacher's gender is added to the baseline model it results in noticeably lower AIC values. In this case adding gender also results in a significant gender coefficient. The result is that a male teacher is predicted to increase math scores by 0.006151 standard deviations versus a female math teacher. For the reading scores we have the opposite effect. Here female teachers are predicted to increase reading scores by 0.007203 standard deviations versus male teachers. While these coefficients are significant and larger than the significant coefficients for teacher experience, they are still relatively small and inconsequential.

3.5 Conclusion

In general, all of the models examined in this chapter have results that are either negligible or nonexistent. Of the available information – a teacher's gender, experience, and level of education – the best fit model is the model that includes only gender. This model has significant but negligible results. The second-best model includes only a teacher's level of education and this model returns insignificant results. Though these models show a better fit to the data, their resulting effect on the rankings of the teachers would be very minimal. Interestingly it appears that both a teacher's level of education and years of experience are of no help in predicting whether or not they will be good at teaching students of specific prior abilities.

Chapter 4

Discussion

There is a large financial incentive for both state and local school districts to implement policies that measure teacher effectiveness. VA models are a leading candidate to perform this measurement. The nature of VA modeling constrains all results to a very narrow band of possibilities. In the case of the dataset used in the random-slope model, $10,000+$ teachers are ranked based on an intercept value that ranges approximately from -1 to 1.5. This leaves very little room between the individual teachers and results in unavoidable ranking differences when even the slightest change is made to the model. As an illustration, the value separating the teacher ranked $5,000^{\text{th}}$ and the teacher ranked $5,001^{\text{st}}$ is 0.000018. This observation and the fact that there are numerous valid models that can be proposed, point to the conclusion that any VA model used to rank teachers should not be taken as gospel. It can probably be said that any model would correctly identify the best and worst teachers. However, the professional ramifications of such volatile rankings are too important to ignore. In such circumstances it seems like a VA model is a useful tool for identifying very good or very poor teachers but that its use in policy, funding, or teacher assessment should be limited. It certainly seems rash and ill-conceived to make such information publicly available.

Appendices

Appendix A

Models

The random-slope model presented in this report:

$$\begin{aligned}(\text{CurrentScore})_{ij} &= \beta_{0j} + \beta_{1j}\,(\text{PriorScore})_{ij} + \beta_{2j}\,(\text{Gender})_{ij} \\ &\quad + \beta_{3j}\,(\text{Title1})_{ij} + \beta_{4j}\,(\text{Kinder})_{ij} \\ &\quad + \beta_{5j}\,(\text{ELL})_{ij} + e_{ij} \end{aligned} \tag{A.1}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{A.2}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{A.3}$$

$$\beta_{2j} = \gamma_{20} \tag{A.4}$$

$$\beta_{3j} = \gamma_{30} \tag{A.5}$$

$$\beta_{4j} = \gamma_{40} \tag{A.6}$$

$$\beta_{5j} = \gamma_{50} \tag{A.7}$$

The L.A. Times model[†]:

$$\begin{aligned}(\text{CurrentScore})_{ij} &= \beta_{0j} + \beta_{1j}\,(\text{PriorScore})_{ij} + \beta_{2j}\,(\text{Gender})_{ij} \\ &\quad + \beta_{3j}\,(\text{Title1})_{ij} + \beta_{4j}\,(\text{Kinder})_{ij} \\ &\quad + \beta_{5j}\,(\text{ELL})_{ij} + e_{ij} \end{aligned} \tag{A.8}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{A.9}$$

$$\beta_{1j} = \gamma_{10} \tag{A.10}$$

$$\beta_{2j} = \gamma_{20} \tag{A.11}$$

$$\beta_{3j} = \gamma_{30} \tag{A.12}$$

$$\beta_{4j} = \gamma_{40} \tag{A.13}$$

$$\beta_{5j} = \gamma_{50} \tag{A.14}$$

[†]The only difference between these two models is the presence of u_{1j} in Eq. (A.3) and the absence of u_{1j} in Eq. (A.10)

The variables used in these models are defined as follows:

CurrentScore_{ij}	A student's standardized test score after being in a particular teacher's class
PriorScore_{ij}	A student's standardized prior score
Gender_{ij}	A student's gender
Title1_{ij}	Title 1 identifies students who are failing or are most at risk of failing in schools that have high percentages of children from low-income families
Kinder_{ij}	Whether or not a student joined the school district after kindergarten
ELL_{ij}	A student's status as a native English speaker
e_{ij}	A random component associated with any given student's score; assumed to be distributed $N(0, \sigma^2)$
γ_{00}	A constant intercept term for all predicted teacher effect lines
u_{0j}	A teacher-specific random component to the intercept modifier; assumed to be distributed $N(0, \tau_{00}^2)$
γ_{10}	A constant slope term for all predicted teacher effect lines
u_{1j}	A teacher-specific random component to the slope modifier; assumed to be distributed $N(0, \tau_{11}^2)$
γ_{20}	A constant intercept modifier sensitive to a student's Gender
γ_{30}	A constant intercept modifier sensitive to a student's Title1 status
γ_{40}	A constant intercept modifier sensitive to a student's school district of origin
γ_{50}	A constant intercept modifier sensitive to a student's ELL status

Adding teacher-level variables to the model:

As discussed in Chapter 3, we can add teacher-level variables to the model in an attempt to improve fit. Whenever it was done, it was done according to the model specified here; specifically, in equation A.17 where the variable "Education" was added to the model.

$$\begin{aligned}
(\text{CurrentScore})_{ij} &= \beta_{0j} + \beta_{1j}\,(\text{PriorScore})_{ij} + \beta_{2j}\,(\text{Gender})_{ij} \\
&\quad + \beta_{3j}\,(\text{Title1})_{ij} + \beta_{4j}\,(\text{Kinder})_{ij} \\
&\quad + \beta_{5j}\,(\text{ELL})_{ij} + e_{ij} && \text{(A.15)} \\
\beta_{0j} &= \gamma_{00} + u_{0j} && \text{(A.16)} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}\,(\text{Education})_{j} + u_{1j} && \text{(A.17)} \\
\beta_{2j} &= \gamma_{20} && \text{(A.18)} \\
\beta_{3j} &= \gamma_{30} && \text{(A.19)} \\
\beta_{4j} &= \gamma_{40} && \text{(A.20)} \\
\beta_{5j} &= \gamma_{50} && \text{(A.21)}
\end{aligned}$$

Appendix B

Fit Statistics

Fixed Effects	
(Intercept)	0.20
	(0.00)
cst.lagm	0.71
	(0.00)
factor(gender)F	0.01
	(0.00)
factor(in.title1)1	−0.19
	(0.00)
factor(ell)1	−0.08
	(0.00)
factor(join.after.k)1	0.01
	(0.00)
Random Effects	
Variance: tch_id1.(Intercept)	0.08
Variance: Residual	0.30
Fit Statistics	
AIC	1228515.96
BIC	1228607.98
Log Likelihood	−614249.98
Deviance	1228499.96
Num. obs.	731226
Num. groups: tch_id1	10361

(a) L.A. Times - Math

Fixed Effects	
(Intercept)	0.20
	(0.00)
cst.lagm	0.71
	(0.00)
factor(gender)F	0.01
	(0.00)
factor(in.title1)1	−0.20
	(0.00)
factor(ell)1	−0.08
	(0.00)
factor(join.after.k)1	0.01
	(0.00)
Random Effects	
Variance: tch_id1.(Intercept)	0.09
Variance: tch_id1.cst.lagm	0.01
Variance: Residual	0.30
Fit Statistics	
AIC	1220612.05
BIC	1220727.07
Log Likelihood	−610296.02
Deviance	1220592.05
Num. obs.	731226
Num. groups: tch_id1	10361

(b) Random Slope - Math

Table B.1: Fit Statistics For Mathematics Models

Fixed Effects	
(Intercept)	0.36
	(0.00)
cst.lagm	0.57
	(0.00)
factor(gender)F	0.18
	(0.00)
factor(in.title1)1	−0.33
	(0.00)
factor(ell)1	−0.31
	(0.00)
factor(join.after.k)1	−0.03
	(0.00)
Random Effects	
Variance: tch_id1.(Intercept)	0.06
Variance: Residual	0.37
Fit Statistics	
AIC	1343828.86
BIC	1343920.77
Log Likelihood	−671906.43
Deviance	1343812.86
Num. obs.	721170
Num. groups: tch_id1	10304

(a) L.A. Times - Reading

Fixed Effects	
(Intercept)	0.18
	(0.00)
cst.lagr	0.74
	(0.00)
factor(gender)F	0.04
	(0.00)
factor(in.title1)1	−0.17
	(0.00)
factor(ell)1	−0.10
	(0.00)
factor(join.after.k)1	0.01
	(0.00)
Random Effects	
Variance: tch_id1.(Intercept)	0.04
Variance: tch_id1.cst.lagr	0.01
Variance: Residual	0.25
Fit Statistics	
AIC	1086471.83
BIC	1086586.72
Log Likelihood	−543225.91
Deviance	1086451.83
Num. obs.	721170
Num. groups: tch_id1	10304

(b) Random Slope - Reading

Table B.2: Fit Statistics For Reading Models

Appendix C

The EM Algorithm

The hierarchical linear models (HLM) in this report are estimated using the lmer package for the statistical software program R. When estimating these models, R uses a reduced maximum likelihood method by default to estimate the parameters of the model. The default method can be changed so that R uses maximum likelihood (ML) to estimate the parameters. In this report, the estimates are generated using the ML method.

The difficulty in estimating the parameters for a hierarchical linear model lies in the fact that there are many unknown parameters. In order to estimate the β_{ij}s it is necessary to know σ^2, the variance of the e_{ij}s. In order to estimate the γs it is necessary to know τ_{00} and τ_{11}, the variances of u_{0j} and u_{1j}, respectively. These quantities (σ^2, τ_{00}, and τ_{11}) are not known and, therefore, estimation of all quantities is problematic.

The Expectation-Maximization (EM) Algorithm is a method used to find maximum likelihood estimates of coefficients when data is either missing or incomplete. In the case of HLMs, the data that is actually observed is considered to be the incomplete data, X. The joint distribution of the incomplete (observable) data and the hidden (unobservable) data, Y, makes up what is

called the complete data. The complete data is assumed to come from some distribution that has a vector, θ, of unknown parameters. For the model specified in this paper, the observed (incomplete) data, X, the hidden data, Y, and the vector of parameters, θ are defined as follows:

$$\begin{aligned} X = \{&\text{PredictedScore}_{ij}, \text{PriorScore}_{ij}, \text{Gender}_{ij}, \\ &\text{Title1}_{ij}, \text{Kinder}_{ij}, \text{ELL}_{ij}, (\text{Teacher Level Variables})_j\} \\ Y = \{&\sigma^2, \tau_{00}, \tau_{11}\} \\ \theta = \{&\gamma_{00}, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{30}, \gamma_{40}, \gamma_{50}, u_{0j}, u_{1j}, e_{ij}\} \end{aligned}$$

The complete data, $P(X, Y|\theta)$, can be expressed in the following way:

$$P(X, Y|\theta) = P(Y|X, \theta)P(X|\theta) \tag{C.1}$$

If random starting values, θ_0, are selected for θ, then both $P(Y|X, \theta_0)$ and $P(X|\theta_0)$ can be estimated. So the probability of the complete data, $P(X, Y|\theta_0)$, can be estimated from arbitrary θ_0 and observed X.

The variables of interest are θ and Y. These variables appear in the likelihood function for the data. The likelihood function for the model in this paper is

$$\begin{aligned} L(X, Y|\theta) = \prod_{ij} &\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(PredictedScore_{ij} - E(PredictedScore_{ij})^2}{2\sigma^2}} \\ &\times \frac{1}{\sqrt{2\pi\tau_{00}^2}} e^{\frac{-(u_{0j})^2}{2\tau_{00}^2}} \frac{1}{\sqrt{2\pi\tau_{11}^2}} e^{\frac{-(u_{1j})^2}{2\tau_{11}^2}} \end{aligned} \tag{C.2}$$

If this function or its logarithm can be maximized for values of θ the problem would be solved. The problem is that the values of Y are unknown. By

definition, $E[g(x)|y] = \int_X g(x)f(x|y)\,dx$. If $g(X,Y|\theta) = \log L(X,Y|\theta)$ then the equation

$$E(\log L(X,Y|\theta)|X,\theta_0) = \int_Y \log L(X,Y|\theta)f(Y|X,\theta_0)\,dy$$

is true by definition. This is the "E" step of the EM algorithm.

Next comes the maximization, "M," step of the algorithm. Since $f(Y|X,\theta_0)$ can be estimated using arbitrary θ_0, the integral

$$\int_Y \log L(X,Y|\theta)f(Y|X,\theta_0)\,dy$$

can now be maximized for values of θ. In almost all applications of HLM, these values cannot be maximized analytically. Once values of θ are found that maximize the expected value given Y,X, and θ_0; those values of θ become the new values of θ_0. This process is repeated until it converges on the true values of θ. Once the values of θ that maximize the expectation are known, they can then be used to determine the true values of Y [3][11].

Bibliography

[1] Eva L. Baker, Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. Problems with the use of student test scores to evaluate teachers. Briefing Paper 278, Economic Policy Institute, Washington, D.C., August 2010.

[2] Dale Ballou. Test scaling and value-added measurement. *Education Finance and Policy*, 4(4):351–383, 2009.

[3] Jeff A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, 1998.

[4] Derek Briggs and Ben W Domingue. Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of los angeles unified school district teachers by the Los Angeles Times, 2011.

[5] Raj Chetty, John N Friedman, and Jonah E. Rockoff. The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Working Paper 17699, National Bureau of Economic Research, Cambridge, MA, 2011.

[6] Thomas J. Kane and Douglas O. Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper 14607, National Bureau of Economic Research, Cambridge, MA, December 2008.

[7] J.R. Lockwood, Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, and Felipe Martinez. The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. Technical report, RAND, Santa Monica, CA, 2006.

[8] Daniel F McCaffrey, T.R. Sass, J.R. Lockwood, and K. Mihaly. The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4):572–606, 2009.

[9] Colorado Department of Education. Colorado educator effectiveness bill, sb 10-191, 2010.

[10] US Department of Education. Race to the top program - executive summary, Nov. 2009.

[11] Lisa M. Sullivan, Kimberly Dukes, and Elena Losina. An introduction to hierarchical linear modeling. *Statistics In Medicine*, 18:855–888, 1999.