CS-BIGS 5(1) : 1-11
http://www.bentley.edu/centers/csbigs/sofronas.pdf

An EM-Algorithm-Based Approach for Predicting Teacher Candidate Success on the Communication and Literacy Skills Test for Educator Licensure

Kimberly S. Sofronas
Emmanuel College, USA

Matthew A. Tom
Division on Addiction, Cambridge Health Alliance, USA

Josh Lederman
Wellesley College, USA

In 1998, the Department of Education in the State of Massachusetts redefined the requirements for teacher licensure, implementing a series of licensure examinations entitled the Massachusetts Tests for Educator Licensure (MTEL). In response, many Massachusetts colleges and universities now offer preparatory support programs to help teacher candidates pass the MTELs. Little research has been conducted to determine the impact, statistical or otherwise, of those measures on students' MTEL scores. This paper outlines the development and analysis of a linear model for predicting the scores of teacher candidates at a small liberal arts college on the Communication and Literacy Skills Test, which all K-12 teacher candidates in Massachusetts are required to pass. Although failing test scores are reported numerically by the Massachusetts Department of Education, passing test scores are reported only as a "Pass." In this paper, a variation on the EM algorithm is applied to address the problem of missing data. The statistical technique is outlined in detail and is followed by a discussion of the effectiveness of the preparatory sessions. This paper is accessible to readers who have an introductory level background in statistics.

1. Background

The State of Massachusetts has played a leading role in defining more rigorous requirements for teacher licensure. Since January 1998, the Massachusetts Department of Education, now called the Massachusetts Department of Elementary and Secondary Education (DESE), has required pre-service teachers to earn a Bachelor's degree, successfully complete an educator preparation program, and pass at least two licensure exams for an initial license. Teacher candidates preparing to teach middle or high school must pass an exam in their content area (e.g., history, mathematics, foreign language), while candidates preparing to teach elementary school must pass two licensure exams: the Foundations of Reading and the General Curriculum. Additionally, all teacher candidates must pass the Communication and Literacy Skills Test [CLST], which consists of a Reading skills subtest and a Writing skills subtest.

That battery of tests is commonly known as the Massachusetts Tests for Education Licensure (MTEL).
The 1998 reauthorization of the federal Higher Education Act requires all states to "report annually the pass rates (on tests the states have chosen or developed) for each cohort of prospective teachers completing training programs" at all teacher training institutions within their state (Center for School Reform, 2009, p. 8). The mandate to report these data has led many educator preparation programs to require that teacher candidates pass all teacher licensure examinations prior to their assignment to a full-time student teaching placement. Since institutions are not required to report failing test results to the state for any candidate who has not completed a student teaching practicum, this generally ensures the reporting of a 100% pass rate by teacher training institutions.

From the perspective of educator preparation programs, the requirement to pass all teacher licensure examinations prior to the assignment of a student teaching placement lessens the likelihood of ethical and legal issues that might otherwise arise. In particular, it precludes a situation in which a teacher candidate graduates from an educator preparation program and is ineligible to teach in public schools because he or she is unable to pass state licensure exams.

A driving motivation for the present study was the need to identify with reasonable accuracy those teacher candidates who have the most difficulty passing the CLST MTEL at a small Massachusetts liberal arts college in order to provide them with targeted training to ensure their initial licensure upon graduation, or advise them to consider other programs of study to preclude the possibility of graduating from college unlicensed to teach in Massachusetts public schools.

Research suggests positive linear relationships between SAT scores and scores on some teacher licensure exams (Blue & O'Grady, 2002; Longwell, 2003; Wakefield, 2003 as cited in Pool et al., 2004; White, Burke & Hodges, 1994). Pool et al. (2004) studied the correlation between the SAT Verbal and Praxis I Reading scores at three types of institutions and found a moderately strong positive linear relationship. The Praxis I Reading exam is a teacher licensure exam that is similar to the CLST.

The present study explores the predictive power of SAT scores along with two additional factors: (a) attendance at a voluntary six-session preparatory course and (b) the number of attempts at either subtest of the CLST. The purpose of this study is to examine the following research questions:

- Did the teacher candidates who performed better on either subtest of the CLST attend the preparatory sessions?
- How well do SAT Verbal and SAT Writing scores predict candidates' performance on the CLST?
- Will teacher candidates' chances for passing improve with each subsequent attempt of either subtest of the CLST?

1.1 The Six-Session Preparatory Course

To address concerns regarding the success rate of the college's teacher candidates on the CLST MTEL, a faculty member in the college's English Department developed a six-session preparatory course offered in fall and spring semesters beginning in 2004. Data collection for this study began in the fall semester of 2006. Table 1 briefly outlines the topics presented in each of the six sessions.

Table 1. Topics Presented in Six-Session CLST MTEL Preparatory Course

Session	Topic
1	Overview of the preparatory course; Grammar and mechanics
2	Grammar and mechanics
3	Writing composition
4	Writing composition
5	Summary writing
6	Reading comprehension; Adaptation of specific comprehension strategy

1.2.The Data Set

Data were obtained from freshmen or transfer teacher candidates entering the college during the 2006-2007 academic year. There were a total of 78 results on the Reading subtest and 77 results on the Writing subtest, reflecting that some of the 59 teacher candidates attempted one or both subtests multiple times.

Data collected for the purpose of this study include the following, with one record for each attempt:

- SAT Verbal scores (SATV)
- SAT Writing scores (SATW), if available
- Attendance at each of the six CLST preparatory sessions (offered biannually by the college) during the 2006-2007 and 2007-2008 academic years
- Number of the attempt at the CLST Reading subtest or CLST Writing subtest (i.e., 1 for the first attempt, 2 for the second, etc.)
- Test scores on the Reading and Writing subtests of the CLST

While results that follow should not be generalized to all elementary education majors across the state of Massachusetts, others working with data sets obtained from cohorts of teacher candidates taking licensure exams may find the statistical methodology outlined in this paper useful.

2. Statistical Methods

The analysis of our dataset would ordinarily involve the use of two multiple linear regression models: one predicting the *CLST Reading* subtest scores; and one predicting the *CLST Writing* subtest scores. The independent variables for those models would include SATV scores, SATW scores, the number of the attempt at the CLST Reading subtest or CLST Writing subtest, and attendance at each of the six preparatory sessions offered by the college.

In this study, two complicating factors preclude the use of standard multiple linear regression models. First, the Massachusetts DESE no longer reports numerical scores for passing results on any of its teacher licensure exams. Results are reported numerically only for teacher candidates who score below the minimum passing score of 240 on the CLST. Conversely, individuals who score 240 or higher are notified of their P-status (i.e., passing). By treating those P's as missing data, we can use other results to extrapolate the missing values. Second, some individuals completed high school prior to the year the SAT Writing test was introduced and, therefore, had no SATW scores to include in the data set. Again, the best option was also to treat nonexistent SATW scores as missing data and use other candidates' data to extrapolate the missing scores (i.e., their scores on the Reading and Writing subtests of the CLST).

The Reading and Writing subtests are different exams measuring different skills, therefore, the results could not be combined, which necessitated two additional multiple linear regression models to extrapolate the missing SATW scores: one using teacher candidates' CLST Reading subtest scores and another using their Writing subtest scores.The extrapolated SATW values are then combined using weighted averages. Note that it would not be prudent to estimate a student's SATW score at 400 in the CLST Reading subtest regression model and then estimate that same student's SATW score at 600 in the CLST Writing subtest regression model. Likewise, in cases where a student takes the CLST Reading and-or Writing subtest more than once, the same SATW score should apply to each of his or her attempts. Table 2 below summarizes the four models used in the analysis.

Table 2. Four Multiple Linear Regression Models

Model	Formula	Description
1R	$MTELR = \beta_0 + \beta_{SATV}SATW + \beta_{SATW}SATV + \beta_{S1}S_1 + \ldots + \beta_{S6}S_6 + \beta_{ATT}$(# of attempt)	* Predicts MTEL Reading subtest scores. * Response variable: CLST reading subtest score. * Explanatory variables: SATV, SATW, attendance or lack of attendance teach of the six preparatory sessions, number of attempt. * Data set: only the CLST reading subtest results.
1W	$MTELW = \beta_0 + \beta_{SATV}SATW + \beta_{SATW}SATV + \beta_{S1}S_1 + \ldots + \beta_{S6}S_6 + \beta_{ATT}$(# of attempt)	* Predicts MTEL Writing subtest scores. * Response variable: CLST writing subtest score. * Explanatory variables: SATV, SATW, attendance or lack of attendance teach of the six preparatory sessions, number of attempt. * Data set: only the CLST writing subtest results.
2R	$SATW = \beta_0 + \beta_{MTELR}MTELR + \beta_{SATW}SATV + \beta_{S1}S_1 + \ldots + \beta_{S6}S_6 + \beta_{ATT}$(# of attempt)	* Generates missing SAT Writing scores. * Response variable: CLST reading subtest score. * Explanatory variables: MTELR, SATV, attendance or lack of attendance teach of the six preparatory sessions, number of attempt. * Data set: only the CLST reading subtest results.
2W	$SATW = \beta_0 + \beta_{MTELW}MTELW + \beta_{SATW}SATV + \beta_{S1}S_1 + \ldots + \beta_{S6}S_6 + \beta_{ATT}$(# of attempt)	* Generates missing SAT Writing scores. * Response variable: CLST writing subtest score. * Explanatory variables: MTELW, SATV, attendance or lack of attendance teach of the six preparatory sessions, number of attempt. * Data set: only the CLST writing subtest results.

Table 3. Missing SATW scores and missing MTEL scores

CLST Reading Subtest Records	Failing MTEL Score	Passing MTEL Score	Total
SAT Writing Score Available	23	37	60
SAT Writing Score Not Available	2	16	18
Total	25	53	78
CLST Writing Subtest Records	**Failing MTEL Score**	**Passing MTEL Score**	**Total**
SAT Writing Score Available	26	33	59
SAT Writing Score Not Available	4	14	18
Total	30	47	77

Horton and Kleinman (2007) discuss situations in which there is a pattern in the missing data. "If the data matrix can be rearranged in such a way that there is a hierarchy of missingness, so that observing a particular variable X_b for a subject implies that X_a is observed, for $a < b$, then the missingness is said to be *monotone*" (p. 80). According to Horton and Kleinman (2007), monotone patterns of missingness allow for the use of simpler methods. However, the CLST Reading and Writing subtest results did not show any hierarchical structures relating teacher candidates with SATW scores to passing or failing CLST Reading or Writing subtest results. In other words, there were cases in which candidates had no SATW score and passed the CLST, cases in which they had no SATW score and failed the CLST, cases in which they had an SATW score and passed the CLST, and cases in which they had an SATW score and failed the CLST. That held true for both the Reading and Writing subtest results.

Table 3 shows how many records in the data set were missing SAT Writing scores and CLST MTEL scores. We recall that many teacher candidates attempted one or both subtests multiple times. If a candidate without an SATW score had 2 attempts at the Reading subtest and 1 attempt at the Writing subtest, then that would represent 3 records that were missing an SATW score.

2.1. The Iterative Process

According to McLachlan and Krishnan (1997), the "Expectation-Maximization [EM] algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete data problems" (p.1). To address the complications within our data set, we applied a variation of the EM algorithm that consisted of the following three stages:

Stage 1: Obtain initial values for the missing data
Stage 2: Run successive multiple linear regression models as part of an iterative process
Stage 3: Report the final models and results after the iterative process has converged

Before we could begin the iterative process, we needed initial values for the missing SATW scores and the passing scores on the Reading and Writing subtests of the CLST. In Stage 1 of our adaptation of the EM algorithm, we introduced those initial values. Because 300 is a perfect score on all Massachusetts Tests for Educator Licensure and 240 is the minimum passing score (for more information on MTEL scoring, see www.mtel.nesinc.com/MA16_passing.asp), 270, the average of the two scores, is a reasonable initial value for all passing scores on the CLST.

Since there is a fairly strong linear fit between SATV and SATW scores, ($r = 0.62$, see Figure 1), we used the verbal scores to obtain the initial estimates for the missing SATW scores. Running the regression gave us the formula $\widehat{SATW} = 129.05 + 0.7837SATV$. All of the initial estimates fell between 400 and 750.

In Stage 2, we ran and reran our four multiple linear regression models in sequence using a six-step iterative process (see Appendix A).

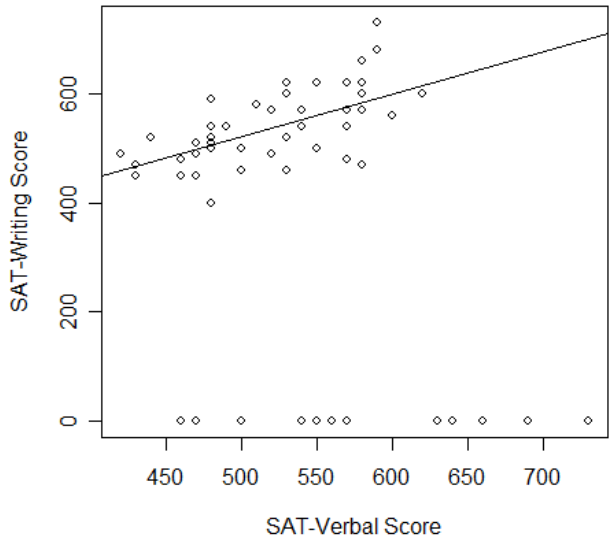

Figure 1. Scatter plot of SAT Verbal and SAT Writing Scores

In Stage 3, we looked at the results of each iteration and confirmed that the iterative process as a whole converged. Standard regression outputs were obtained to determine which variables were significant in each of the four models, and which were not.

2.2. Backwards Elimination

After running an initial 3-stage variation on the EM algorithm to obtain the four models, we identified the coefficient in each model with the highest two-tailed p-value and removed that variable from its associated model. With the new smaller list of variables, we returned to Stage 1, reiterated the whole process and obtained new coefficients for the four models. We continued this backwards elimination cycle, paring down all four models until we were left with coefficients whose p-values were all under 0.10. Any further removal of variables was left to the discretion of the research team.

2.3. The Logistic Regression Alternative

See Appendix B for a discussion of the benefits and drawbacks of using logistic regression to estimate teacher candidates' probabilities of passing the subtests of the CLST and measure the effectiveness of the preparatory sessions.

3. Results

At each stage of the backward elimination process, the iterative procedure converged. Our final results from Models 1R and 1W gave us the following formulas for teacher candidates' expected Reading and Writing subtest scores:

$$MTELR = 188.4 + \frac{6.1}{100}SATV + 7.4S_2 - 10.4S_3 - 12.3S_4 - 7.1S_5 + 20.0S_6 + 8.5Attempt \quad (1R)$$

$$MTELW = 173.9 + \frac{5.4}{100}SATV + \frac{5.9}{100}SATW + 4.095S_6. \quad (1W)$$

Tables 4 and 5 show the coefficients for the variables included in Model 1R and 1W, along with the two-sided p-values for each of the coefficients.

Table 4. Regression Output for Model 1R

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	188.42253	11.76509	16.02	0.00000	***
SATV	0.06080	0.02029	3.41	0.00110	**
S2	7.41384	4.04414	1.83	0.07102	.
S3	-10.43697	3.34023	-3.13	0.00259	**
S4	-12.32946	3.86958	-3.19	0	**
S5	-7.10014	3.16406	-2.24	0.02800	*
S6	20.02423	3.39374	5.90	1.17E-007	***
Attempt	8.52730	1.70401	5.00	4.01E-006	***

Signif. codes: . $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 5. Regression Output for Model 1W

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	173.94313	9.00993	19.31	0.00000	***
SATV	0.05884	0.02603	2.26	0.02670	*
SATW	0.05445	0.02338	2.33	0.02260	*
S6	4.09514	2.11881	1.93	0.05710	.

Signif. codes: . $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

The residual standard errors for the two models are very close: 9.995 for Model 1R, 9.022 for Model 1W. We can then estimate teacher candidates' probabilities of passing using normal distributions with standard deviations between 9 and 10. With either subtest, a candidate who expects to score approximately 246 has roughly a 75% chance of passing. Likewise, a candidate who expects to score approximately 255 has about a 95% chance of passing.

The sections that follow report the findings of this study as they relate to the three predictors: (a) SATV and SATW, (b) attendance at the six-session preparatory course, and (c) number of attempts at either subtest of the CLST.

3.1. The Preparatory Sessions as Predictors of Success on the CLST

According to Equation 1W, attendance at session 6 had a minor effect on the expected CLST Writing score – adding 4.1 points (standard error: 2.1 points). None of the other sessions had a significant effect. According to Equation 1R, preparatory sessions 2 and 6 were positive indicators for teacher candidates' performance on the CLST Reading subtest, adding 7.4 points and 20.0 points, respectively (standard errors: 4.0 points and 3.4 points, respectively), to a teacher candidate's expected score. Interestingly, preparatory sessions 3, 4, and 5 were negative indicators, lowering a candidate's expected score on the Reading subtest by 10.4, 12.3, and 7.1 points, respectively (standard errors: 3.3, 3.9 and 3.2 points, respectively).

While it is possible that sessions 3, 4 and 5 presented concepts that may have confused some teacher candidates and worked to effectively lower their scores on the Reading subtest, we are not claiming that attendance at sessions 3, 4, and 5 causes candidates to perform worse on the CLST Reading subtest. Two lurking variables offer other possible explanations for this finding. First, it is possible that candidates' weaknesses in *writing*, if any, may negatively affect their performance on the *Reading* subtest of the CLST MTEL. As Table 1 indicates, sessions 3, 4, and 5 emphasize writing skills and the candidates who chose to attend those sessions had either already failed the Writing subtest, or had not yet taken the test. Hence, the sub-pool of individuals attending sessions 3, 4, and 5 included those with weaker writing skills.

Second, many of the writing skills addressed in sessions 3, 4, and 5 (e.g., writing a clear thesis, supporting a thesis with evidence, and drafting conclusions) require significant time to develop. There are essentially no direct testing strategies or other "tricks" that teacher candidates might take away from sessions 3, 4, and 5 and apply to either subtest of the CLST. On the contrary, Session 6 emphasizes a reading comprehension technique that is concrete and immediately applicable and would likely improve a teacher candidate's score.

3.2. SAT Scores as Predictors of Success on the CLST

Teacher candidates with higher SATV scores had both higher expected CLST Reading subtest scores and higher expected CLST Writing subtest scores. With all other factors equal, for example, the expected CLST Reading subtest score for a candidate with an SATV score of 600 is 6.1 points higher than the expected score for a teacher candidate with an SATV score of 500 (standard error: 2.0 points for a 100-point difference in SATV). The expected CLST Writing subtest score for the candidate with the SAT Verbal score of 600 is 5.9 points higher (standard error: 2.6 points for a 100-point difference in SAT Verbal). SATW, on the other hand, was a significant predictor for only the CLST Writing subtest.

With all other factors equal, the expected CLST Writing subtest score for a teacher candidate with an SATW score of 600 is 5.4 points higher than the expected score for a candidate with an SATW score of 500 (standard error: 2.3 points for a 100-point difference in SAT Verbal). If all the other factors are equal, those same two teacher candidates will have the same expected CLST Reading subtest score. Most of SATV and SATW scores in the data set were between 400 and 650. As a result, the expected CLST Reading and Writing subtest scores given by our results are probably not valid for extremely weak teacher candidates (i.e., those with SAT scores under 400) and extremely strong teacher candidates (i.e., those with SAT scores above 700).

3.3. Repeated Test Attempts as Predictors of Success on the CLST

The findings of this study revealed no increase in expected scores on the CLST Writing subtest with repeated attempts. However, according to the data analysis, we did find an increase of 8.5 points per attempt on the Reading subtest (standard error: 1.7 points). For two reasons, we must exercise some caution in making claims regarding the extent to which candidates can improve their scores through repeated attempts at this subtest.

First, teacher candidates who pass the Reading subtest never take it again. Only those candidates who fail and must take the subtest a second time have the opportunity to show improvement. Candidates who must take the subtest a third time provide two numerical results for comparison. Ideally, in order to truly assess the effects of repeated attempts, we would have at least two or three results for every candidate. Arguably, if teacher candidates who passed the Reading subtest did take it a second time with only minimal or no improvement then our analysis would have yielded a result much lower than 8.5 points gained per attempt.

Second, diminishing returns may be an issue on the fourth, fifth or sixth attempt of the Reading subtest. Our data set does not have enough fourth attempts in it to detect this phenomenon.

4. Using the Results to Identify Teacher Candidates At-Risk of Failing the CLST Subtests

Although we can use the coefficients corresponding to *attendance at the preparatory sessions* (i.e., S_1, S_2, S_3, S_4, S_5, S_6) and to the *number of attempts at the subtests* to draw conclusions about the effectiveness of the preparatory sessions or the tests themselves, we cannot use them to predict the future performance of individual candidates. It does not make sense, for example, to decrease teacher candidates' expected CLST Reading subtest scores because they exercised the motivation to attend preparatory sessions 3, 4 and 5. Similarly, the data collected on *number of attempts at the subtests* might be misleading, as we have no way of knowing how teacher candidates divided their time between the two subtests during the allotted 4-hour testing period.

Anecdotal reports from teacher candidates revealed that some devoted the majority of their time (e.g., 3.5 hours) to one subtest of the CLST leaving very little time for the other subtest. Moreover, adding 8.5 points to an expected score for each additional attempt at the Reading subtest of the CLST as our model suggests is flawed. However, as benchmarks, we can calculate the expected scores for candidates taking the CLST for the first time without any preparation (S_1 through S_6 set to 0, Attempts set to 1). The sections that follow outline procedures for using Models 1R and 1W to estimate the probabilities that a student will pass the Reading and Writing subtests of the CLST using only SAT scores.

4.1. Probabilities of Passing the Reading Subtest of the CLST

Prediction intervals' standard errors can now be used to estimate probabilities of passing the CLST Reading subtest. Equation 1R can be simplified when we consider teacher candidates who are taking the CLST Reading subtest for the first time without any preparation:

$$\widehat{MTELR} = 197.0 + \frac{6.1}{100} SATV \qquad (1R_a)$$

Because SATW was removed from Model 1R (see Eq. 1R), it need not be considered here. When SATV is

between 400 and 700, the standard error on $\widehat{MTELR}$ is between 10.1 and 10.6 points. For SAT Verbal scores within that range, Figure 2 shows the confidence level, α, such that $(240, \infty)$ is an alpha-level right-sided confidence interval for the predicted Reading subtest score. While α is not a direct estimate for a teacher candidate's probability of passing the Reading subtest of the CLST, it is a good indicator. For example, α is approximately 50% when the SAT Verbal is in the 610-640 range. According to our model, even candidates with the highest SATV scores in our data set (i.e., SAT Verbal scores close to 650) have a greater than 40% chance of failing if they take the Reading subtest of the CLST without any preparation.

4.2. Probabilities of Passing the Writing Subtest of the CLST

Similarly, prediction intervals' standard errors can be used to estimate probabilities of passing the CLST Writing subtest. Model 1W (see Eq. 1W) contains both SATV and SATW scores. Again, consider teacher candidates taking the CLST Writing subtest for the first time without any preparation. To get a sense of their chances of passing the CLST Writing subtest, we look at candidates with equal SAT Verbal and SAT Writing scores (e.g., 450, 450). Because the coefficients and standard errors for SATV and SATW are each nearly the same, a teacher candidate with SATV = SATW = 500 should have approximately the same chances of passing the Writing subtest of the CLST as a teacher candidate with an SATV = 400 and SATW = 600. Setting $S_6 = 0$, gives us the simpler equation:

$$\widehat{MTELW} = 173.9 + \frac{5.4}{100}SATV + \frac{5.9}{100}SATW \qquad (1W_a)$$

When SATV is between 400 and 700, the standard error on $\widehat{MTELW}$ is between 9.1 and 9.6 points. For SAT Verbal scores within that range, Figure 2 shows the confidence level, , such that $(240, \infty)$ is an alpha-level right-sided confidence interval for the predicted CLST writing subtest score. In this case, α is approximately 50% when the combined SATV/SATW score is around 1150. In contrast to the results for the CLST Reading subtest, strong students have high chances of passing the Writing subtest on their first attempt. According to our model, the students in the cohort with the highest combined SAT scores in our data set (i.e., approximately 1300) have close to an 80% chance of passing the Writing subtest without any preparation.

5. Examining the Model

In multiple linear regression, residual plots are often used to detect problems with models and their fits. Before we can construct residual plots for Models 1R and 1W, we must simulate the missing passing scores (note that any simulated score below 240 is re-simulated). Using the simulated passing scores, we can generate sample residual plots, such as those in Figure 3.

Furthermore, we can generate many sets of residuals, and then for each set we can calculate the correlation between the predicted subtest scores and the residuals. Figure 4 shows the correlations for 100 sets of residuals for Models 1R and 1W, respectively. Both histograms are centered to the right of 0, which suggests some positive correlation between the predicted CLST subtest score and the residual. This means Models 1R and 1W likely underestimate, candidates' expected scores. This underestimation may prove to motivate candidates to spend much needed time preparing for the CLST.

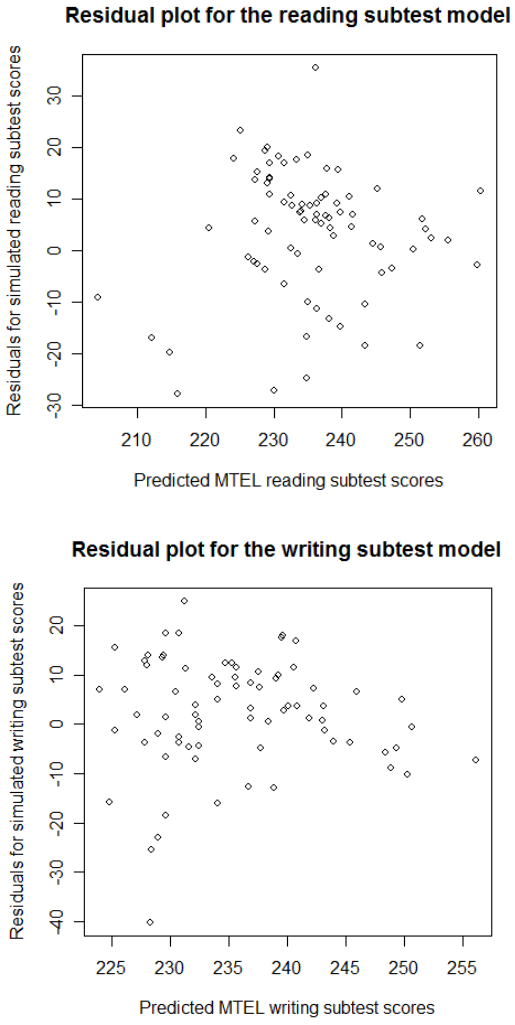

Figure 3. Sample Residual Plots

The missing passing scores and the simulated residuals pose other problems. First, simulated residuals for students who *pass* have minimum values, while simulated

residuals for students who *fail* have maximum values. For example, if a student has an expected score of 250 and passes, the simulated residual will probably be between -10 and +30. If a student has an expected score of 230 and fails, the corresponding range spans from -30 to +10.

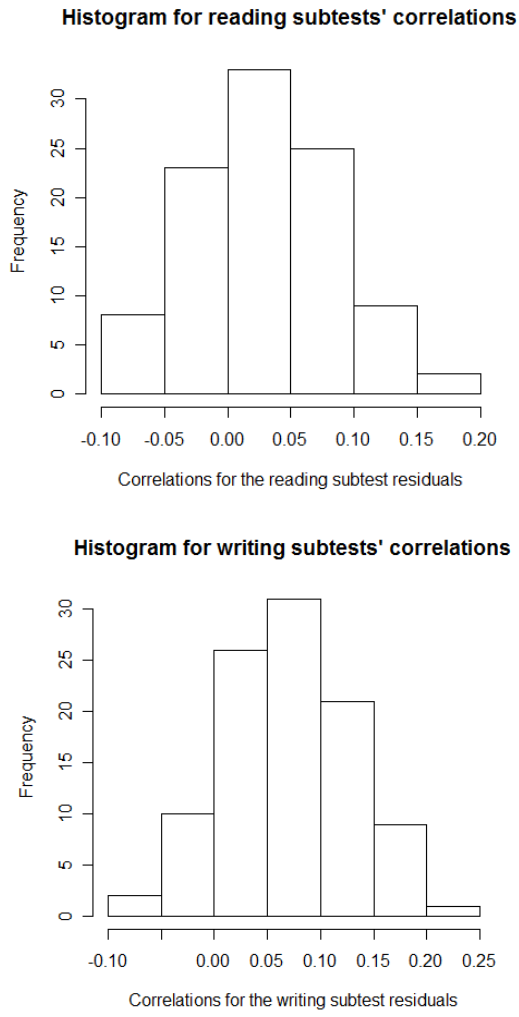

Figure 4. Histograms for Reading and Writing Subtests' Correlations

Since students with lower expected scores have lower chances of passing, we could get more negative residuals on the left side of the residual plot and more positive residuals on the right side. This may produce a positive correlation between the expected scores and the residuals. This problem would be exacerbated if we generated residuals using the larger standard errors computed for confidence or prediction intervals instead of the standard errors given by Models 1R and 1W themselves. Second, the model assumes that for each teacher candidate, the residuals for all of his or her attempts are independent. With multiple attempts, this may not be the case. For example, suppose that the model states that a particular candidate has an expected Writing subtest score of 230, and then suppose that the candidate's score on a first attempt is 215. According to the model, the expected score on a second attempt, assuming no additional preparation, is still 230. If the 215 is a more representative measure of the candidate's potential, then it is likely that the residuals for future attempts will be centered around -15, not 0. Third, the replacement values for the missing data are calculated using the coefficients for Models 1R, 1W, 2R and 2W in that iteration. As the coefficients get updated with each iteration, the resulting residuals decrease. As a consequence, the standard errors in the four models underestimate the actual corresponding standard deviations. It is possible that students' subtest scores have more variability than our models and results would indicate. However, problems with inducing bias and understating variability are not new to missing data methods (Horton and Kleiman, 2007). Because of this potential bias, administrators and advisors should exercise caution when considering counseling students out of the teacher training program.

6. Final Remarks

This study examined the predictive power of three variables on teacher candidates' CLST test scores and presented an adaptation of the EM algorithm as one approach for addressing the problem of missing data across two linked multiple linear regression models. Our algorithm converged after only a few iterations, and we were able to use stepwise elimination to pare down Models 1R, 1W, 2R and 2W. Although the potential bias in the model may not allow us to assign students accurate probabilities of passing the subtests of the CLST, it does identify which students need more or less support relative to each other. The model also offers a way to evaluate and rank the effectiveness of the preparatory sessions.

Since the time this data was collected, the CLST MTEL has undergone considerable revision (See http://www.doe.mass.edu/news/news.aspx?id=4830 for specific details related to those revisions). The open-ended vocabulary definitions have been removed from the revised CLST Reading subtest, leaving only reading comprehension questions. Likewise, two sections have been removed from the revised CLST Writing subtest - a section on open-ended definitions of grammatical terms and a multiple-choice section to identify spelling, punctuation, and-or other grammatical errors. The sections of the Writing subtest that remain (i.e., a section with error-laden sentences that must be rewritten and a section containing short passages followed by multiple-choice questions that address editing issues) have been

expanded to include more of the same kinds of items. We believe the predictive power of session 6 (S_6) may be even greater now that the revised version of the CLST Reading subtest consists of only reading comprehension questions. However, this study should be replicated to determine the impact of these changes on the predictive power of the variables outlined above. While the specific numerical results based on the data collected for this study are no longer useful, the analysis in this paper can be taken as proof of concept.

In light of growing concerns surrounding the use of quantitative test data to determine a qualitative capacity for teaching (Berliner, 2005; Hess, 2005; Pool, Dittrich, Longwell, Pool, and Hausfather, 2004), it is especially critical for educator preparation programs to minimize potential ethical issues that might arise as direct or indirect consequences of teacher licensure examinations. It is unprincipled to accept four years of college tuition money from teacher candidates unlicensed to teach upon graduation. Likewise, prematurely counseling teacher candidates away from the teaching profession because of obstacles related to passing licensure exams is also problematic.

The methodology outlined in this study offers a means to evaluate the effectiveness of MTEL preparatory programs and identify teacher candidates at risk for failing either subtest of the CLST. It is our hope that investigators at other institutions will be able to run this same algorithm on new data sets and obtain similar types of results for the purpose providing candidates at risk of failing with the early support needed to experience success. Investigators at other institutions may choose to include additional variables (e.g., high school and-or college GPA, performance in related coursework, SATM, etc.) in their data set. Since all high school graduates who have taken the SAT now have SATW scores, the need for Models 2R and 2W has been eliminated, simplifying future analyses of this kind.

Acknowledgements: The authors would like to thank the reviewer for helpful comments that have certainly led to a clearer presentation of this paper. The authors would also like to thank Nicholas Horton for his advice on the preparation of this manuscript.

References

Berliner, D. C. 2005. The near impossibility of testing for teacher quality. Journal of Teacher Education, 56(3), 205-213.

Blue, T. & O'Grady, J. 2002. Finding the best teachers: A study of relationships among SAT, GPA, Praxis series test scores, and teaching ratings. Pennsylvania Teacher Educator, 1, 1-12.

Center for School Reform 2009. Why MTEL, not PRAXIS, will maintain teacher quality in Massachusetts. Pioneer Institute Public Policy Research.

Diez, M. E. 2002. The certification connection. Education Next, 2(1), 8-15.

Hess, F. M. 2005. The predictable, but unpredictably personal politics of teacher licensure. Journal of Teacher Education, 56(3), 192-198.

Horton, N. & Kleiman, K. P. 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician, 61(1), 79-90.

McLachlan, G. & Krishnan, T. 1997. The EM algorithm and extensions. Wiley series in probability and statistics. John Wiley & Sons.

Pool, J., Dittrich, C., Longwell, E., Pool, K., & Hausfather. 2004. An analysis of SAT and PRAXIS I: Performance of teacher education candidates at three difference types of institutions. Action in Teacher Education, 26(2), 60-68.

White, W. F., Burke, C. M., & Hodges, C. A. 1994. Can Texas teacher certification be predicted from SAT scores and grade point averages? Journal of Instructional Psychology, 21, 298-299.

Appendix A

Step 1: Fit models 1R and 1W.
Step 2: Use those results to obtain refined estimates for the passing CLST Reading and Writing subtest scores. Substitute estimates into our data set as the new passing scores.
Step 3: Run models 2R and 2W with this new data set.
Step 4: Obtain improved estimates for the missing values in SATW. For each student who did not take the SAT Writing test, models 2R and 2W will give multiple estimates for the missing SATW score - one per attempt at each of the Reading and Writing subtests. Average estimates to obtain a single value for the missing SATW score.
Step 5: Substitute these updated values in for all of the missing entries in SATW.
Step 6: Run Steps 1 - 5 until all of the models and values for missing data converge.

Appendix B

One alternative approach is to reduce all the failing grades to F's and use logistic regression. This approach still requires two separate models: one for the CLST Reading subtest and one for the CLST Writing subtest; however, the response variable is binary – pass or fail – instead of numerical. While that simplifies the problem of missing data, discarding the actual failing scores may represent a significant loss of information.

The missing SAT Writing scores remain problematic and still need to be interpolated. One possibility is to use another four-model variation on the EM-algorithm. Instead of the two linear regression models 1R and 1W, we have two logistic regression models 3R and 3W.The independent variables are the same SAT scores, session attendance, and number of attempts. The response variable is whether the candidate passes or fails. To re-estimate the missing SATW scores, we use two linear regression models 4R and 4W.The only difference between models 2R and 2W and 4R and 4W is that we replace the CLST subtest scores with the estimated log-odds:$\ln\left(\frac{p_{pass}}{1-p_{pass}}\right)$. Here, p_{pass} is the estimated probability of an attempt yielding a passing score. Just as the original algorithm cycles through Models 1R, 1W, 2R and 2W, this new algorithm cycles through Models 3R, 3W, 4R and 4W. Backwards elimination is again used to prune the four models.

Another approach involves first interpolating the missing SATW scores and then running the logistic regression a single time without iteration. We start with Models 5R and 5W, which model SAT Writing score as a function of attendance at the preparatory sessions, SATV scores, number of attempts at the CLST subtest and whether the attempt yields a passing or failing score. Those two models are then used to get estimates for the missing SATW scores. From there, we introduce two logistic regression models, 6R and 6W, to estimate the effects of the different factors on the chances of passing the subtests. Just as in Models 3R and 3W, the independent variables are the SAT scores, session attendance, and the number of attempts. The response variable is whether or not the candidate passes or fails. Similarly, backwards elimination is used to prune all the models.

Unfortunately, there are a number of drawbacks to both the iterative and the one-pass approaches. First, interpreting the coefficients and results of the logistic regression models is more complicated. We learned from out original approach that attendance at a given session will raise or lower the expected subtest score by a fixed number of points and the expected score on a second attempt at the CLST Reading subtest is a fixed number of points higher than the expected score on a first. Interpreting the coefficients of the logistic regression models involves examining changes in odds of success to get to the [non-linear] changes in candidates' probabilities of passing. For many, the interpretations of the results in the logistic regression models are less accessible than the interpretations of the results in the models from our originally proposed method.

Second, more variables were removed during backwards elimination with both logistic approaches, limiting the possible conclusions (for reference, Tables 6, 7, 8 and 9 contain the output for models 3R, 3W, 6R and 6W, respectively). For example, the number of attempts is a significant component of the model for the CLST Reading subtest using our original approach; however, this variable (p-value = 0.0945) could be eliminated at the discretion of the statistician using the iterative logistic regression approach. Number of attempts was not a factor in the case of the one-pass logistic regression approach. Variables S3 and S5 – both significant indicators in the models from our original approach - did not survive the backwards elimination with either logistic regression approach. It is possible that, for many of the teacher candidates, changes in the values of these three variables represented the difference between failing and failing by less. Because the logistic regression models use a binary response variable, such an effect cannot be detected by these procedures.

A benefit of the logistic regression approaches is that they offer a more optimistic, and possibly more realistic, view

of teacher candidates' chances of passing the subtests of the CLST MTEL. Figure 5 shows the three procedures' estimates for the probabilities of passing the test the first time without preparation, for students with different SAT Verbal scores and matching SAT Writing scores. According to the logistic regression models, students with SAT scores in the 500's have better chances of passing than our original models predicted and students with SAT Scores in the 700's should pass more than 80% of the time.

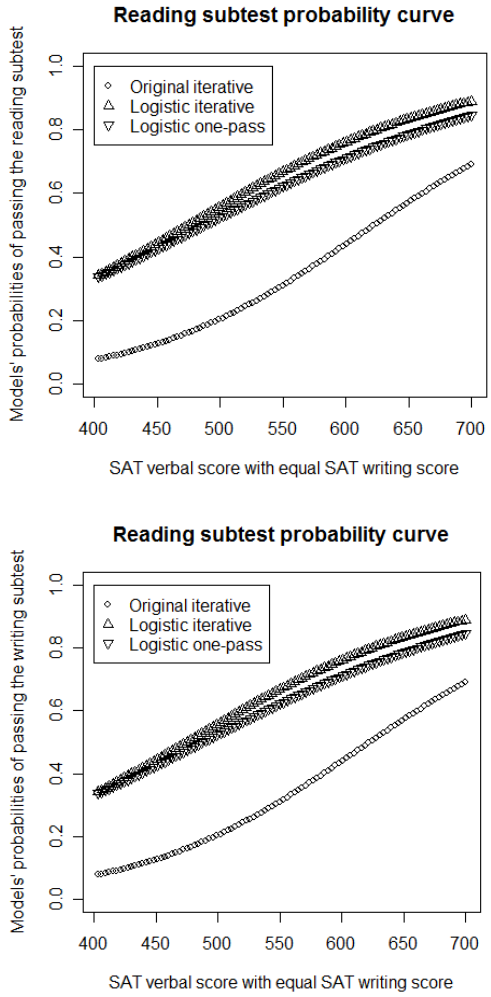

Figure 5. Probability Curves for Original and Logistic Regression Models

It remains an interesting trade-off. The results of the original procedure explain what is happening with the teacher candidates in the data set, but give a biased, pessimistic view of future candidates' chances of passing. The results of the logistic regression provide a more accurate model for predicting future success, but do not do as good of a job showing the effects of the different preparatory sessions.

Table 6. Regression Output for Model 3R

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-5.153766	2.551453	-2.020	0.0434	*
SATW	0.009224	0.004359	2.116	0.0343	*
S4	-1.873888	0.889126	-2.108	0.0351	*
S6	2.087282	0.853973	2.444	0.0145	*
Attempt	0.780352	0.466727	1.672	0.0945	.

Signif. codes: . $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 7. Regression Output for Model 3W

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-14.47324	3.655077	-3.960	7.5e-05	***
SATW	0.014859	0.007507	1.979	0.0478	*
SATV	0.013016	0.007303	1.782	0.0747	.
S1	1.042058	0.621473	1.677	0.0936	.

Signif. codes: . $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 8. Regression Output for Model 6R

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-3.906234	2.160624	-1.808	0.0706	.
SATW	0.008008	0.003954	2.025	0.0429	*
S6	0.960070	0.541339	1.774	0.0761	.

Signif. codes: . $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 9. Regression Output for Model 6W

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-10.16137	2.742013	-3.706	0.00021	***
SATW	0.020315	0.005329	3.812	0.00014	***

Signif. codes: . $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Correspondence: sofronki@emmanuel.edu
mattatom@msn.com
jlederma@wellesley.edu