Teacher Certification Exams: Predicting Failure on the TExES History (8-12) Content Exam (A Nonparametric Approach using Classification Trees)

by

Dwight Richard Gard, B.B.A., M.S., M.B.A.

A Dissertation

In

HIGHER EDUCATION

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY

Approved

Douglas J. Simpson
Co-Chairperson of the Committee

John P. Murray
Co-Chairperson of the Committee

Eugene W. Wang

Pamela E. Tipton

Accepted

Peggy Gordon Miller
Dean of the Graduate School

May 2011

ACKNOWLEDGMENTS

I give special thanks to my loving wife, Ruth, and our three sons, David, Andrew, and Jonathan for their love and encouragement during this long academic endeavor.

To Dr. Doug Simpson, I offer my sincere appreciation for saying "Yes" to serving as co-chair of my committee when he probably needed to say "No." His sharp-witted sense of humor always made me smile and his wise counsel was invaluable in helping me successfully navigate the dissertation process.

To Dr. John Murray, I would like to express my deep gratitude for his willingness to continue to serve as co-chair of my committee despite having left Texas Tech several years earlier to take a faculty position at another university. I am indebted to him for extending me that courtesy.

To Dr. Eugene Wang, I am thankful for him nudging me in the direction of an interesting dissertation topic. I very much enjoyed all the technical discussions about statistical matters.

To Dr. Pam Tipton, I am grateful for her in-depth knowledge about many aspects of the topic I was investigating. Her pleasant demeanor, accessibility, and willingness to answer so many of my questions are greatly appreciated.

TABLE OF CONTENTS

ACKNOWLEDGMENTS ii

ABSTRACT v

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

I. INTRODUCTION 1

Purpose of the Study 3

Statement of the Problem 4

Research Questions 13

Assumptions 14

Delimitations 14

Limitations 15

Significance of the Study 16

Definition of Terms 16

II. LITERATURE REVIEW 22

Historical Background 22

Research Antecedents 49

Chapter Summary 57

III. METHOD 59

Restatement of Research Questions 60

Research Design and Rationale 61

Data Source 86

Classifying and Coding 90

Data Collection 92

Sampling Bias 94

Data Analysis 97

Reliability, Validity, Generalizability 106

Chapter Summary 106

IV. RESULTS 108

Descriptive Statistics 108

Correlation Analysis 111

Classification Trees 112

Chapter Summary 132

V. DISCUSSION 134

Predictors of TExES Exam Result 134

Implications for Research and Practice 144

Recommendations for Further Research 146

Conclusion 147

REFERENCES 149

APPENDICES

A. Domain Descriptions for History (8-12) 162

B. Graphical Methods for Assessing Model Performance 164

ABSTRACT

Teacher Certification Exams: Predicting Failure on the

TExES History (8-12) Content Exam

(A Nonparametric Approach using Classification Trees)

Previous research efforts concerning teacher certification in Texas focused primarily on the Pedagogy and Professional Responsibilities exam; an exam that all teacher candidates must pass regardless of their specific content area. Few studies have attempted to explore which variables are useful for predicting the outcome of the TExES content-area certification exams, which represents a major gap in the literature. Because of its high failure rate, this study focused on identifying factors that were influential in predicting failure on the TExES History (8-12) certification exam. A convenience sample was used and only those who had taken the TExES History (8-12) exam from 2002 – 2008 were selected (n = 181).

The study is an exploratory data design using classification trees—a nonparametric statistical technique often associated with data mining. The study was different from previous studies in two important aspects: a) the study included a much wider range of variables, and b) nonparametric, classification tree methodology was used to build predictive models.

Using the proportional chance criterion and Press' Q to assess significance, the models were statistically significant ($p < .05$), indicating that the models were capable of predicting outcomes well beyond what would be expected based on chance.

Because classification trees produce a set of decision rules that can be graphically depicted, a model based on a decision tree paradigm is more intuitive, and more easily interpreted and implemented compared to regression methods.

Although classification trees are not widely used in social science research, the success of the technique in the current study suggests that classification trees can be an effective, nonparametric alternative to the more traditional multiple regression and logistic regression methods and provides researchers a glimpse of the capabilities of classification trees.

LIST OF TABLES

3.1 Advantages and Limitations of Classification Trees 68

3.2 Classification Tree Misclassification Matrix 70

3.3 Evaluating Risk Using a Misclassification Matrix 104

4.1 Evaluation Matrix for Five Models Using Minimum Risk Pruning Criterion 116

4.2 Evaluation Matrix for Five Models Using Standard Error Pruning Criterion 117

LIST OF FIGURES

1.1 TExES History (8-12) Initial Pass Rate (Statewide vs. XTU) 5

1.2 Prediction Interval vs. Confidence Interval..... 12

3.1 Classification tree example..... 69

3.2 Methodology for Supervised Modeling 80

3.3 Optimal level of model complexity..... 82

4.1 Best model (minimum risk pruned)—ACT model..... 121

4.2 Best model (standard error pruned)—SAT model 123

4.3 No Scores model (standard error pruned) 128

4.4 TASP model (standard error pruned) 129

4.5 ACT model (standard error pruned)..... 130

4.6 All Scores model (standard error pruned) 131

B.1 Gains Summary and Gains Chart: No Scores model..... 165

B.2 Response chart: No Scores model 166

B.3 Lift (Index) chart: No Scores model..... 166

B.4 Gains Summary and Gains Chart: TASP Scores model..... 167

B.5 Response chart: TASP Scores model 168

B.6 Lift (Index) chart: TASP Scores model..... 168

B.7 Gains Summary and Gains Chart: ACT Scores model 169

B.8 Response chart: ACT Scores model 170

B.9 Lift (Index) chart: ACT Scores model..... 170

B.10 Gains Summary and Gains Chart: SAT Scores model..... 171

B.11 Response chart: SAT Scores model 172

B.12 Lift (Index) chart: SAT Scores model 172

B.13 Gains Summary and Gains Chart: All Scores model 173

B.14 Response chart: All Scores model 174

B.15 Lift (Index) chart: All Scores model 174

CHAPTER I

INTRODUCTION

Overview

Evaluating the competence of professionals has long been a major concern of educators. Evaluating professional competence is widespread and there are high stakes consequences for those seeking certification. According to McGaghie (1991), the transparent purpose of evaluating the competence of professionals is to provide public assurance that the services of professionally credentialed persons are effective and safe. Policy makers and the public at large have been concerned that the high cost of funding higher education has not provided the proper educational outcomes and, as a result, institutions of higher education have been criticized for the quality of teacher education programs.

In 1981, U.S. Secretary of Education T. H. Bell, expressed concern about the widespread public perception that something was seriously wrong with our system of education, and that the United States was losing its world dominance in commerce, science, and technology. There was a sense that a rising tide of mediocrity had eroded the massive educational and technological gains achieved during the Cold War era. These educational gains were largely a result of the launch of the Soviet satellite, Sputnik; an event perceived as a direct threat to the security of the United States.

On August 26, 1981, the National Commission on Excellence in Education was established to study the problem and make recommendations. The final report, *A Nation at Risk: The Imperative for Educational Reform* (1983), harshly criticized the

condition of our educational system. The committee believed the problem was serious enough that it posed a danger to the security and global prominence of the nation. The impact of this seminal report set in motion a series of events and legislation calling for improvement and accountability, leading to the current system of high-stakes achievement testing and standards-based education reform.

In 1984, the Texas Legislature passed House Bill 72, which initiated standardized testing for current teachers to ensure that they possessed basic literacy and math abilities. In addition, new teacher education candidates were required to pass an entry-level basic skills assessment as well as two exit-level assessments—one on pedagogy and one on their specific academic content area. The Texas legislature, concerned about the quality of education, imposed a system of accountability in order to provide a cadre of "highly qualified" teachers as mandated by the *No Child Left Behind Act of 2001* (NCLB). According to NCLB, teachers are considered highly qualified if they (a) hold a bachelor's degree, (b) possess full state certification or licensure, and (c) can demonstrate competence in each subject they teach. To become certified, a candidate must pass a specific content-area exam (e.g., History 8-12) as well as the Pedagogy and Professional Responsibilities (PPR) exam that assesses knowledge of pedagogical theories.

Previous research efforts concerning teacher certification in Texas focused primarily on the PPR exam, an exam that all teacher candidates must pass regardless of their specific content area. However, little research has focused specifically on the success rates of teacher education candidates on content-specific certification exams.

A review of the literature identified only four studies (Gruver, 2008; Jackson, 2006; McIntosh, 2002; Weiss, 2003) that sought to predict success on content-specific certification exams. Among those, only one study (Gruver, 2008) explicitly addressed the History (8-12) certification exam. Identifying factors that affect the success rate on content-specific exams could help teacher education programs accomplish the mission of producing highly qualified teachers. The lack of prior research on content-specific exam pass rates represents a major gap in the literature.

Purpose of the Study

The Texas State Board for Educator Certification (SBEC) implemented a performance-based accountability system to assure the quality of teacher education programs. Perhaps the most visible and significant measure of the success of a program are the pass rates on teacher certification exams. Such measures are reported in aggregate as well as in disaggregated form by gender and ethnicity. An institution must achieve acceptable pass rates in all categories to prevent any type of state-imposed sanctions on its programs. This study examines the results of the History (8-12) TExES exam for students at a large public university in Texas (XTU) pursuing teacher certification and attempts to identify variables capable of predicting failure on that certification exam. The purpose of this study was to conduct an exploratory data analysis concerned with:

1. Determining the descriptive statistics of first-time TExES History (8-12) examinees who took the exam from 2002 – 2008.

2. Using correlational analysis to determine if statistically significant relationships exist among any of the following variables:

   - TExES result (pass/fail)*
   - Transfer status
   - PostBac/Bac status
   - Gender
   - Ethnicity
   - Age at time of exam
   - Lower division GPA
   - Upper division GPA
   - Total GPA
   - History GPA
   - Total number of history courses
   - Number of upper division history courses
   - TASP scores (Read/Write/Math)
   - ACT scores (Read/Math)
   - SAT scores (Read/Math)

   * dependent variable

3. Developing a nonparametric, statistical model using classification trees for predicting failure on the History (8-12) certification exam.

Statement of the Problem

In Figure 1.1, data from the Texas SBEC show a pattern of declining pass rates on the TExES History (8-12) certification exam. In attempting to reverse this trend, an "early warning system" capable of identifying students who are likely to fail the exam provides the opportunity for intervention so at-risk students have a greater chance of success. Such knowledge might also be useful for program evaluation.

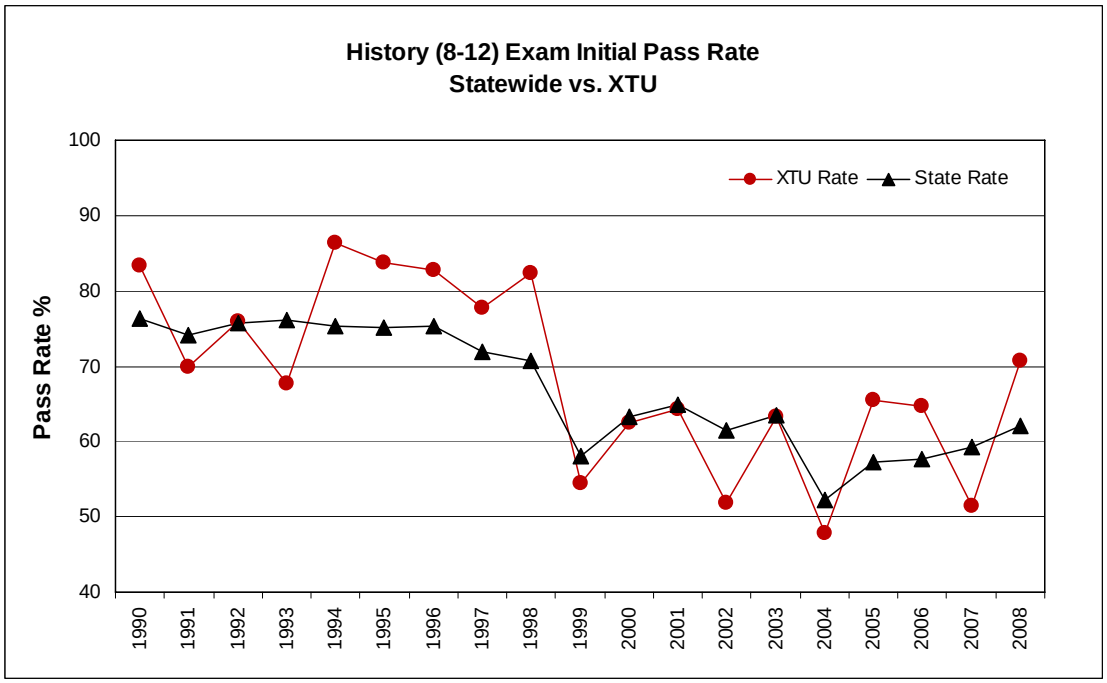

Figure 1.1. TExES History (8-12) Exam Initial Pass Rate. Statewide vs. XTU (r = .833, p < .01).

Low pass rates can create a number of detrimental effects on the institution, the student, and society. Examples of such adverse effects include:

- Low pass rates reflect poorly on the school and its programs.
- Low pass rates signal poor program quality and can affect the number of students seeking admission to a program.
- Low pass rates can affect the accreditation status of a program.
- Students who fail the exam are denied timely entry into their profession and may suffer an opportunity cost which affects them financially and psychologically.
- Students' careers may be placed on hold, or they may simply "throw in the towel" and give up on their aspirations of a teaching career.

- Low pass rates may impact the mission of producing highly qualified teachers.

Pass Rates

Assessment data measure two different types of pass rates. The "first-year pass rate" (also known as the "initial pass rate") reflects performance during the year in which the tests are initially taken. The "cumulative pass rate" (also known as the "final pass rate") is based on performance during the subsequent year and cumulatively measures results over the entire two-year period. If fewer than 30 students are involved, current data are combined with data from the previous one or two years (SBEC, 2001).

To maintain accreditation, a program must achieve a 70% first-year pass rate or an 80% 2-year, cumulative pass rate across all demographic groups (all, male, female, African American, Hispanic, White, Other) (SBEC, 1998).

Individually, however, candidates must achieve a minimum score of 80% on the TExES History (8-12) exam. As shown previously in Figure 1.1, both the XTU and statewide initial pass rates have been below the 70% minimum since 1999, although in 2008 the XTU pass rate was slightly above 70%.

Perspective on Trend

Gruver (2008) suggests that this downward trend is due in part to the increased rigor of the History (8-12) TExES, and the amorphous nature of the history discipline, which is unusually broad in its context. In addition, it was suggested that because of the curricular diversity of the history discipline, two problems were evident: (a) public

school history curriculum differed significantly from the higher education history curriculum, and (b) an incongruent perception by history faculty members of their role in training teachers for public schools. Referencing an American Historical Society report from 1991, Gruver notes that history departments "contained less cohesion of courses—types, eras covered, chronology—than math and science departments and courses studied because math and science remained tied to scope and sequence more than historical studies" (p. 15).

Although Figure 1.1 shows data from 1990 forward, the TExES exam replaced the ExCET exam and was phased in beginning in 2002. Gruver (2008) noted that the History TExES and the History ExCET were similar in the content covered, but the History TExES was more rigorous and more aligned with the state's public education curriculum because it contained an additional instructional domain that focused on the teaching of "doing history." Doing history was defined as "using primary and secondary sources to evaluate and interpret historical events, including their causes and effects, and communicating these conclusions in writing, utilizing thesis statements and relevant evidence" (p. 21). With some semblance of disclaimer, Gruver points out that "many History faculty at state universities in Texas have not recognized the History TExES as a valid measure of a candidate's understanding of or ability to do history" (p. 4). Faculty supported it nonetheless, fearing potential sanctions from the SBEC if history majors did not pass the test. Because of the differences between the ExCET and the TExES exams, which can confound the analysis, only data from the TExES exam was used in the study.

Regardless of the form of the certification exam, the data still show a downward trend in pass rates. The graph shows a conspicuous drop in pass rates in 1999. Because this occurred several years before the phase-in of the TExES exam in 2002, this drop in pass rates must be attributable to factors other than the change in the form of the exam. Although a definitive explanation for this drop in 1999 was not found, Mike Ramsey (research specialist with the Texas Education Agency) posits that at the end of fiscal year 1999, lifetime certificates were discontinued and standard certificates (which have to be renewed approximately every five years) were offered instead. Ramsey suggests that teachers may have taken the History (8-12) test with less preparation as they rushed to get the last of the lifetime certificates. The chart also shows marked declines in 2002 and 2004. Ramsey explains that the first decline coincides with the time when the TExES was introduced, and the second decline occurred when the ExCET was discontinued (personal communication, September 21, 2009).

Performance-based Accountability

The SBEC implemented a performance-based accountability system designated the Accountability System for Educator Preparation (ASEP). The SBEC issues annual accreditation ratings based on the combined performance of all candidates. The data are also disaggregated into groups based on gender and ethnicity. Ratings are based on how well candidates perform on the assessments required for certification. Certification exams are usually taken near the end of a candidate's

preparation program; however, candidates must receive approval from program administrators before they can register for the exams.

Problems with Stepwise Multiple Regression Methods

Researchers studying similar phenomena in a variety of educational environments and situations have overwhelmingly preferred the use of stepwise multiple linear regression in an attempt to predict exam scores. Logistic regression was used to a lesser extent. There are, however, a number of problems with these approaches.

In a number of the studies identified in the review of the literature, researchers either neglected to test that underlying regression model assumptions were met, or neglected to report it. Failure to meet the underlying assumptions means that the results may not be valid, leading to inaccurate estimates of significance, effect size, and statistical power. These inaccuracies can manifest as incorrect measures of significance of the regression coefficients (e.g., indicating significance when it is not) and biased and inaccurate predictions of the dependent variable (Hair, Black, Babin, Anderson, & Tatham, 2006). Without testing the model assumptions, the validity of the results, conclusions, and assertions may be suspect. The use of stepwise regression is not recommended because this approach (based on the order in which variables are entered into the model) results in inflated risks of Type I error and does not include higher-order or interaction terms (McClave, Benson, & Sincich, 2008; Warner, 2008).

Stepwise regression should be used only when necessary, and then only as a variable screening tool for identifying potential variables to be used in the model-building process. Stepwise regression should not be used as the final model for predicting the outcome variable (Kutner, Nachtsheim, Neter, & Li, 2005; McClave, et al., 2008). Warner (2008) summarizes the argument against stepwise methods by stating that the use of stepwise regression models "often yield analyses that are not useful for theory evaluation (or even for prediction of individual scores in different samples)" (p. 551). Berk (2004) is critical of stepwise regression because the "procedures used to select the correct model can be very misleading" (p. 133). In addition, Berk notes that (a) the selected model may not be appropriate for the situation, (b) stepwise methods tend to capitalize on chance, and (c) the "best" model may be substantive nonsense. He summarizes his argument by stating, "there is no necessary correspondence between a selection criterion and a scientific criterion" (p. 133).

Confidence Intervals versus Prediction Intervals

Prior research utilizing stepwise multiple regression analyses generally resulted in low R^2 values, and therefore, low predictability. Low R^2 values resulted in large standard errors of the estimate (approximately seven to eight points). A 95% confidence interval would have a width of approximately +/- 15 points, which is of little practical value for predicting scores. For example, a predicted score of 80 would have a 95% confidence interval of 65 – 95. With an interval this large, we would not know whether an intervention was appropriate except in extreme cases, which would

then be obvious even without using a predictive model. To intervene when it is not necessary would be a waste of time and resources, while not intervening when it is necessary would increase the risk of a student failing the exam.

However, in a few studies (Jackson, 2006; Pisani, Pisani, & Anderson, 2002; Poelzer, Liang; & Simonsson, 2007; Zeng, Simonsson, & Poelzer, 2002), the dichotomous, categorical dependent variable was defined as pass/fail and logistic regression was used to predict the probability of passing, rather than predicting a particular score. Most of the studies involved the PPR certification exam, although four studies were identified that focused specifically on content-area exams (Gruver, 2008; Jackson, 2006; McIntosh, 2002; Weiss, 2003). Only one study, by Gruver (2008), focused explicitly on the History (8-12) certification exam.

Because we are interested in identifying specific individuals for intervention (not the "average" individual), using the standard error of the estimate is not appropriate because doing so predicts an average score for all students possessing a specific combination of values for the independent variables. It seems unlikely that there would be a sufficient number of students possessing exactly the same values of the independent variables to make the process useful, especially if many variables are involved.

The procedure for predicting a score for a specific individual involves the standard error of the prediction. It is always the case that the prediction interval is wider than the confidence interval (Groebner, Shannon, Fry, & Smith, 2008). Using

the prediction interval instead of the confidence interval produces results that are more uncertain and therefore less useful for identifying specific at-risk candidates.

The nature of the relationship between the confidence interval and the prediction interval is illustration in Figure 1.2. The pair of inner curved bands represents the confidence interval. The pair of outer curved bands represents the prediction interval. Although the exact shape of the curves depends on the data, a prediction interval is always wider than a confidence interval. The wider the interval, the less useful it is for predicting an outcome. The further the predicted score is from the mean, the larger is the prediction interval.

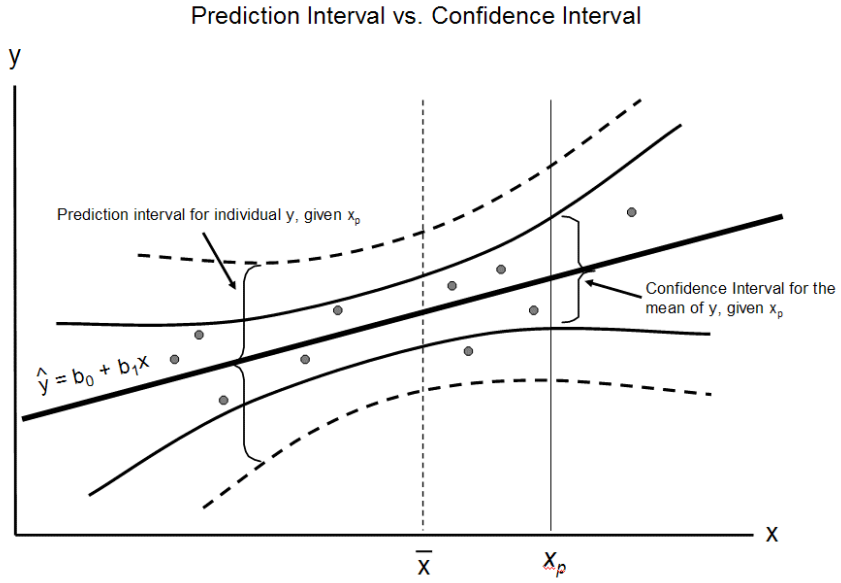

Figure 1.2. Prediction Interval vs. Confidence Interval

Research Questions

Because of the limited amount of previous research involving TExES content area exams in general and the History (8-12) exam in particular, and the importance of high stakes testing in Texas, the study addresses the following research questions:

1. What are the descriptive statistics for XTU candidates taking the History (8-12) TExES certification exam for the first time for the years 2002 – 2008?
2. Does a statistically significant relationship exist among any of the following variables?

   - TExES result (pass/fail)*
   - Transfer status
   - PostBac/Bac status
   - Gender
   - Ethnicity
   - Age at time of exam
   - Lower division GPA
   - Upper division GPA
   - Total GPA
   - History GPA
   - Total number of history courses
   - Number of upper division history courses
   - TASP scores (Read/Write/Math)
   - ACT scores (Read/Math)
   - SAT scores (Read/Math)

   * dependent variable

   (H_0: no statistically significant relationship exists)
3. Does a nonparametric, classification-tree methodology produce a better predictive model for correctly classifying group membership (pass/fail on History exam) at a statistically significant level compared to chance as measured by Press's Q test statistic ($p < .05$), and the proportional chance criterion?

   (H_0: classification tree model is no better than chance)

Assumptions

This study was conducted under the following assumptions:

1. Data provided by the State of Texas and teacher education programs were reliable concerning the History TExES and TASP/THEA exam scores, student transcript information, and student attributes such as gender, ethnicity, and age.
2. The TASP/THEA assessment is a valid, reliable measure of a student's reading, writing, and mathematical ability.
3. The TExES History (8-12) exam is a valid, reliable instrument that accurately measures knowledge of history.
4. A student's grades in general education and teaching field courses accurately reflect the student's level of knowledge, understanding, and achievement.

Delimitations

The study is delimited as follows:

1. Because the newer TExES certification exam is different from the ExCET exam that it replaced, this study includes only students who took the TExES History (8-12) exam during the years 2002 through 2008.
2. Data from XTU were analyzed separately and subsequently merged with data from the Gruver (2008) study for an aggregate analysis. Results of the analysis should only be interpreted in the context of XTU and the three regional state universities involved in the Gruver (2008) study.

3. Although a student may take the TExES History (8-12) exam multiple times, only the score from the first attempt was used.
4. Teaching field GPA was based on all history courses completed, whether taken in residence or from a transfer institution. Because a student might have taken the same course more than once, the highest grade was always used in calculating GPA.

Limitations

The limitations of the current study are as follows:

1. A convenience sample was used in the analysis of XTU students who took the TExES History (8-12) exam from 2002 - 2008. Therefore, the results are not generalizable beyond that context.
2. Because the Gruver (2008) study involved data from three regional state universities in Texas, the results of the aggregate analysis using data from XTU and the other three universities are not generalizable beyond that context.
3. Because each teacher education program is allowed to determine its own requirements (within bounds) for teacher certification, individual history departments may have differing requirements for students to complete a degree and earn a teaching certificate. This variability in requirements introduces an unavoidable confounding effect, which complicates the interpretation of the results and may weaken the validity and reliability of the findings.

Significance of the Study

Because of the pattern of low pass rates on the History (8-12) certification exam and the limitations of prior studies, this study attempts to develop a predictive model using classification trees that will be useful for identifying individuals at risk of failing the certification exam. Because classification trees produce a set of decision rules that can be graphically depicted, a model based on a decision tree paradigm would likely be more intuitive, and more easily applied and interpreted by those seeking to identify at-risk candidates.

Although classification trees are not widely used in the social sciences, the methodology should be amenable to many of the problems in that realm. If the use of classification trees proves to be robust and pragmatic in the context of this study, then any profession seeking to predict the success/failure of examinees on certification/licensure exams could readily adopt the technique. Because of the gap in the research literature for understanding the factors that contribute to the success of teacher education candidates on secondary-level content-area certification exams, this study adds to that body of knowledge.

Definition of Terms

- ASEP (Accountability System for Educator Preparation): An accountability system for monitoring educator preparation programs based on performance standards set by the ASEP Advisory Committee. The ASEP rates educator preparation programs based on their candidates' pass rates on the assessments required for certification. ASEP rules and policies

were developed with input from educators and citizens across Texas (TEA, 2009).

- Certification Field: The content area that a teacher is certified to teach as a result of passing the appropriate certification exam.
- Classification Matrix: a means of assessing the predictive accuracy of a classification tree analysis. A 2-by-2 matrix created by cross-tabulating actual group membership with predicted group membership. Numbers on the diagonal represent correct classifications and off-diagonal numbers represent misclassifications (Hair et al., 2006).
- Classification Tree: A rule-based, statistical algorithm for predicting the class of an object from the values of its predictor variables. One of a number of specific techniques subsumed under the broader topic of data mining.
- Combined data: If the number of First-year or Cumulative test takers in an ethnic or gender group was less than 30, the performance is combined with the performance for that group from the previous one or two years. The "Combined" pass rates are used for accreditation purposes.
- Data Mining: An analytic technique designed to explore data in search of consistent patterns or systematic relationships between variables. The ultimate goal of data mining is prediction.
- Doing history: The use of primary and secondary sources to critically analyze, evaluate, and interpret historical events, including cause and

effect, and communicating conclusions in writing, by arguing thesis statements supported by appropriate evidence.

- Domain: Specific, testable subject matter organized into broad content areas (e.g., World History, U. S. History, etc.)
- ExCET (Examination for the Certification of Educators in Texas): standardized teacher certification tests introduced by the State of Texas in 1986. Examinees had to pass a pedagogy exam and a content area exam. It was replaced by the TExES exam in 2002.
- Final pass rate: The percent of tests passed by candidates of a teacher education program through the second December 31 following the academic year of completion. Known also as cumulative pass rate, it is based on performance over the two-year period. The pass rate is based only on tests required to obtain certification in the field(s) in which the person completed a program during that academic year. The rate reflects a candidate's success on the last attempt made on the test by the second December 31 following the year of completion. If a program's pass rate reflects the performance of fewer than 30 students, current data are combined with data from the previous one or two years (SBEC, 2004).
- General education: Common lower division courses that each student in a public university in Texas must complete to earn a degree.
- Highly Qualified Teacher: A term introduced by the *No Child Left Behind Act of 2001* specifying minimum content preparation requirements for

teachers. According to the legislation, a highly qualified teacher holds a minimum of a bachelor's degree and has passed a state academic subject test in the content area to be taught (Public Law 107-110, 2002).

- Initial pass rate: The percent of tests passed by candidates of a teacher education program through December 31 following the academic year of completion. The pass rate is based only on the tests required to obtain certification in the field(s) in which the person completed a program during the academic year. The rate reflects a candidate's success on the last attempt made on the test by December 31 following the year of completion. If a program's pass rate reflects the performance of fewer than 30 students, current data are combined with data from the previous one or two years (SBEC, 2004).
- Maximum Chance Criterion: Measure of predictive accuracy in the classification matrix. Calculated as the percentage of respondents in the largest group. In the situation of having to make an uninformed classification, the best option is to classify an observation as belonging to the largest group (Hair et al., 2006).
- PPR (Pedagogy and Professional Responsibilities) exam: Exam that assesses a student's knowledge of pedagogical and learning theories.
- Press's Q statistic: A measure of the classification power of the classification analysis compared with the results from a chance model. The calculated value is compared to a critical value based on the chi-square

distribution with one degree of freedom. If the calculated test statistic exceeds the critical value, then the results are significantly better than would be expected by a chance model (Hair et al., 2006).

- Proportional Chance Criterion: Method of assessing the hit ratio. Average probability of classification is calculated considering group sizes (Hair et al., 2006).
- SBEC (State Board for Educator Certification): Entity created by the Texas legislature in 1995. Because of increased accountability, entity oversees the state's teacher education programs. Became part of the Texas Education Agency in 2005.
- SBOE (State Board of Education): The agency of elected officials providing oversight of all aspects of public education, including teacher preparation. In 1995, the State Board for Educator Certification (SBEC) was created by the Texas legislature to govern teacher preparation programs.
- Subject matter exam: Exam for assessing a teacher education candidate's knowledge of the specific content field that will be taught. Also known as teaching field or content area exam.
- TASP (Texas Academic Skills Program): An assessment program introduced by the Texas legislature in 1989. Used for assessing the reading, writing, and mathematics knowledge of students entering public

colleges and universities. The TASP replaced the less rigorous PPST. In 2003, it was renamed the Texas Higher Education Assessment (THEA).

- TExES (Texas Examination of Educator Standards): A state-required testing program for candidates seeking certification as educators in Texas public schools. The TExES program is the result of a collaborative development process by the State of Texas—represented by the State Board for Educator Certification (SBEC) and the Texas Education Agency (TEA)—and its contractor, Educational Testing Service (ETS). The purpose of the exam is to ensure that educators have the required content and professional knowledge for an entry-level position in Texas public schools. The TExES replaced the Examination for the Certification of Educators in Texas (ExCET) in 2002 (SBEC, 2008).
- THEA (Texas Higher Education Assessment): Formerly named the Texas Academic Skills Program (TASP).

CHAPTER II

LITERATURE REVIEW

Historical Background

According to higher education historian Christopher Lucas (Lucas, 1994), "criticizing colleges has always been something of a national pastime, dating back practically to the colonial era" (p. xi). However, during the latter part of the twentieth century, the intensity of the criticism had risen to historically unprecedented levels. Claims of declining educational quality dogged the entire hierarchy of higher education—from the most prestigious to those of more modest reputations. Lucas suggests that quite possibly, the most fundamental tenet of higher education called into question is the issue of academic quality.

In the nineteenth century, teacher preparation was characterized by two different approaches—one used mostly in rural areas, and one espoused by cities and towns (Angus, 2001). Wise and Leibbrand (2000) note that in the 1800s, "teachers knew little more than their students" (p. 613). During the mid-1900s, most policy makers viewed teaching as an activity that could be easily acquired on the job, albeit with some supervision. Consequently, lax preparation and entry requirements were common (Wise & Leibbrand, 2000). Although many of the professional educators of the time were convinced that teachers should be prepared through formal training programs, there were skeptics. Rural communities were inclined to believe that good teachers were born, not made. In contrast to doctors or lawyers, professional educators were not perceived as possessing important, arcane knowledge.

McGaghie (1991) defined a profession by two criteria: (a) the acquisition of an unusually esoteric and complex body of knowledge and skill, and (b) an orientation toward serving the public, with an emphasis on ethical or altruistic practices. Notwithstanding what little "professional knowledge" educators were deemed to possess, many perceived it as little more than common sense. The perception was that many adults could perform the task of schooling, even without specialized training, provided they had the requisite knowledge of the subject they were to teach. That perspective gave rise to the notion that educators did not "possess a body of important, arcane knowledge, and nineteenth century Americans generally resisted the idea that experts should decide key questions about schooling" (Angus, 2001, p. iv). Nonetheless, in the first three decades of the twentieth century, local communities lost their influence over education, and professional educators gained control over the licensing of teachers. The debate over teacher quality has changed little since the nineteenth century. The dominant theme in such debates swirls around "the education profession's relentless efforts to gain control over the licensing of teachers" (Angus, 2001, p. iv).

Ironically, America's desire for the democratization of higher education may have been a significant factor in the decline of academic quality. Unlike countries holding elitist views and insisting on selectivity in higher education admissions, America's egalitarian view held that virtually anyone desiring access to higher education should be afforded the opportunity. It was this belief that nurtured a trend toward less selective admissions (Lucas, 1994).

Before about 1950, teacher's colleges were responsible for preparing teachers, and the strength of those institutions was the ability to integrate subject matter and pedagogy (Noddings, 1998). Before the late 1960s, most states certified teachers by virtue of having completed a teacher education program. The apparent lack of controversy seemed to imply that the educator preparation system was functioning well. However, during the decade of the 1970s teacher education programs came under increasing attacks by groups concerned about the quality of teaching in the public schools (Rubinstein, McDonough, & Allan, 1986).

Seeking to bolster the prestige of the education profession, teachers' colleges recognized the premium placed on doing research, the attribute most associated with large universities. In an attempt to become more like the large, prestigious universities, teachers' colleges began to shift their values and priorities, usually by de-emphasizing the teacher education function. The goal was to become more like a comprehensive, research-oriented university (Lucas, 1994). As a result, teachers' colleges have all but disappeared. Although by 1990 most states had eliminated the bachelor degrees in education, teacher education programs continued to comprise up to 25% of degree requirements for those pursuing secondary teacher certification (Gruver, 2008).

Deterioration of the Education System

Lucas (1994) noted that some conservative critics placed the blame for the deterioration of the education system on the social upheaval during the 1960s and 1970s. In a time when respect for authority and all things traditional were being

eroded, higher education administrators were being pressured by students to allow themselves to have more autonomy in structuring their education. During this time of rebelliousness and sometimes outright hostile confrontation, administrators eager to placate angry students granted them the prerogative to have significant input regarding the curricular requirements for their degrees. Other educational policies and standards were also relaxed in the hope of achieving some measure of educational détente between students and administrators. It is not difficult to understand why many students chose to substitute easier, less demanding courses for the more difficult ones. According to Lucas (1994), this lowering of expectations and standards of excellence would usher in an evolving culture of educational mediocrity.

The Imperative for Educational Reform

Wise and Leibbrand (2000) point out that during the Industrial Age, states originally set low standards to ensure a ready supply of teachers. However, the issue of the decline in the quality of education rose to national prominence in the years following the launch of the Soviet satellite, Sputnik. Perceived as a threat to national security, Sputnik provided a national security rationale for government investment in higher education. With the passage of the National Defense Education Act of 1958, for the first time the federal government began subsidizing higher education directly, rather than through contracts for specific research (Menand, 2010). A well-educated citizenry was now considered a strategic resource.

The 1950s was a decade of intense criticism of the American education system, and colleges of education were often blamed. In describing the state of the American education system in the 1950s, Angus (2001) states that:

> In the late 1950s, the system of teacher preparation came under attack for its low standards of entry and exit, its Mickey Mouse courses, overemphasis on pedagogy rather than subject mastery, the lack of a coherent professional knowledge base, and the absence of reliable evidence that teacher training has a relationship to effective classroom teaching. By the mid-70s, the ideas of the education establishment were rejected by much of the public and even by many within the profession. This skepticism runs broad and deep today. (p. v)

As early as 1981, U.S. Secretary of Education T. H. Bell expressed concern about a widespread public perception that something was seriously wrong with our system of education, and that the United States was losing its dominant position in the world in commerce, science, and technology. There was an uneasy feeling that drifting away from a standard of academic excellence had eroded the massive educational and technological gains previously achieved as a direct result of the launch of Sputnik, which was perceived as a direct threat to the national security of the United States.

On August 26, 1981, the National Commission on Excellence in Education was established to study the problem and make recommendations. So influential was the report that its publication is considered a landmark event in modern American educational history. The final report, *A Nation at Risk: The Imperative for*

Educational Reform (1983), harshly criticized the condition of the American educational system. Among statistics cited by the report, almost 13% of all 17-year-olds in the U.S. were functionally illiterate and 40% of minority youth were functionally illiterate. Nearly 40% of 17-year-olds could not draw inferences from written material, and only 20% could write a persuasive essay.

Authors of the report claimed that too many teachers were being drawn from the lowest quartile of college students, and teacher preparation programs were unduly weighted with pedagogical courses at the expense of specific content knowledge. According to the report, half of the newly minted math, science, and English teachers were not qualified to teach those subjects, and the curriculum had become diluted and devoid of a central purpose. Jorgensen and Hoffman (2003) critically suggested that not only were there insufficient numbers of academically able students being attracted to the teaching profession, but that the quality of teacher preparation programs was lacking. The committee believed the problem was serious enough that it posed a danger to the security of the nation. The analysis of test scores seemed to confirm that America was losing its global prominence, which brought the issue of educational reform to the forefront in U.S. politics (Wakefield, 2003).

Upon publication of the report, "the media opened up against public education with a barrage of newspaper, radio, and, particularly, television coverage. No less a person than the President himself [Ronald Reagan] joined in the public criticism" (Lutz, 1986, p. 81). Although the public at large usually cared little about national debates on educational policy, the general topic itself had begun to capture their

attention. Interestingly, public pressure helped focus attention on certification testing programs because on various occasions parents had received notes or reports from teachers peppered with grammatical and other types of errors (Gorth & Chernoff, 1986). The impact of this seminal report set in motion a series of events and legislation calling for improvement and accountability of the educational system, leading up to the current system of high-stakes achievement testing and standards-based education reform. Roth (1996) believes the impact of this accountability movement created "an entire historical era in the profession, equal in significance to other major periods in education history such as the development of normal schools" (p. 242).

Accountability and the Politicization of Education

After harsh criticism of the education system, policy makers were inclined to implement a system of accountability based upon high-stakes assessment as a necessary component of the push toward standards-based educational reform (Jorgensen & Hoffman, 2003). Education had become politicized. Moral and financial support for schools was contingent upon how well schools were performing, which was usually assessed by standardized testing. Because the tests were deemed to be objective, the public perceived the tests as the most effective way to assess school performance. Although the public seeks some external assurance of the quality, Airasian (1987) notes that "wisely, or not, the public is unable or unwilling to accept testimonials about the status of education from teachers and school administrators, whom they perceive to have a vested interest in that status" (p. 406).

The politicization of education caused increased controversy because decisions beneficial for some groups were detrimental to others. In addition, because of special interest groups, policy makers had to understand the arguments in the educational debate, as well what group was putting forth those arguments (Airasian, 1987). Airasian further argues that, in essence, "test scores become a medium of exchange to be bartered for educational, social, and economic benefits or rewards" (p. 405). The educational benefits of testing were often subordinate to the perceived political benefits. Gratz (2000) argues that politicians look for villains to fight, and incompetent teachers were the villain *du jour* who were destroying the fabric of society.

On January 8, 2002, President George W. Bush signed into law the *No Child Left Behind Act of 2001* (NCLB), marking the beginning of a new era where "accountability, local control, parental involvement, and funding what works became the cornerstones of the nation's education system" (Jorgensen & Hoffman, 2003, p. 6). A new term, *highly qualified teacher*, entered the vocabulary of the education reform movement. It specified minimum content preparation requirements for teachers. According to the language of the legislation, a highly qualified teacher holds a minimum of a bachelor's degree and has passed a state academic subject test in the area to be taught. With the passage of NCLB, the federal government established specific criteria for teachers.

Policy makers were convinced that prospective teachers should be held to an objective, content-oriented standard that specified the minimum level of skill and

knowledge required to be a successful teacher. Because of the wide disparity in the quality of teacher education programs, certification testing was designed to use criterion-referenced exams rather than norm-referenced exams. Advances in measurement theory and statistical techniques provided impetus for the use of criterion-referenced exams.

Texas Education Reform

In 1983, Governor Mark White commissioned a Select Committee on Public Education (SCOPE), chaired by Texas billionaire H. Ross Perot. The committee was to assess the condition of the Texas educational system and to make recommendations for improvement. It was Perot's contention that education policy makers (state administrators, elected officials, and teacher education faculty) were primarily responsibility for the lack of academic achievement. Perot reported that students in colleges of education represented the bottom quartile of SAT scores among students in college at the time. In addition, Perot was convinced that competency testing would improve student and teacher performance, and help to identify and weed out incompetent teachers (Perot, 1984; Shepard & Kreitzer, 1987).

In 1984, based in part on the findings and recommendations of the SCOPE report, the State of Texas enacted House Bill 72. Lutz (1986) remarked that the bill "represents probably the most massive change in the history of Texas public education. This stands as *the* symbol of the Texas education reform" (p. 70). A major component of the reform legislation was a call for teacher competency testing, which was strongly opposed by teachers' unions in Texas. Standardized testing for current

teachers was initiated to ensure that they possessed basic literacy and math abilities. In order to become certified, prospective teachers would now be required to pass an entry-level basic skills assessment as well as two exit-level assessments—one on pedagogy and one on content. Texas embraced the concept that standardized tests, along with other measures, could be used to help evaluate teachers' knowledge as well as teacher education programs. Airasian (1987) contends that standardized tests are perceived to be "scientific" because a precise numerical score is produced. The tests are perceived as "fair" because everyone taking the test must meet the same criteria for passing, and perceived as "objective" because decisions based on the outcomes of the tests are not influenced by the bias of the education profession. Roth (1996) suggests that an uneasy tension exists between the education profession (teacher education programs in particular) and the state legislature. Those involved in teacher education viewed prescriptive governmental mandates as an intrusion into the affairs of the profession, despite teachers performing their duties in the public domain. Roth (1996) believes:

> The basis for this appears to lie in a lack of credibility of teacher education, which motivates lawmakers to enter the scene and set matters straight. Many believe that the issue of credibility is the most critical problem facing the profession today, and it manifests itself in the standard-setting process. (p. 257)

Ultimately, Roth (1996) contends that the move for accountability revolves around issues of trust, or more properly, mistrust. In addition, dissatisfaction with public education contributes to what is essentially a vote of "no confidence" in the education

profession. Evidence of this lack of confidence was demonstrated by the legislature's willingness to set the standards without meaningful input from those in the profession (Roth, 1996).

During the 1987 Texas legislative session, the passage of Senate Bill 994 prohibited undergraduate majors in education. Warner (1990) notes that one obstacle to the initial proposal of Senate Bill 994 was the lack of an agreed upon body of knowledge upon which to specify a teacher education program. Professional educators could not agree on a specialized body of teaching knowledge that was not possessed by the public. With passage of the bill, teacher education was to be developed within a liberal arts curriculum, and education courses were limited to 18 credit hours, including student teaching (Lutz, 1986). What was needed were "teachers who are broadly and deeply educated, not people who mostly studied education" (Angus, 2001, p. v).

Rift between Liberal Arts and Education Programs

After the launch of Sputnik, a national debate ensued about the quality of the American education system, and those involved with teacher education programs bore the brunt of the blame. Angus (2001) notes that suddenly "their sense of being marginalized received a public hearing and the fundamental clash of values between professional schools of education and liberal arts departments came to the forefront" (p. 25). According to Angus (2001), the standard criticisms of schools of education by liberal arts departments were characterized by:

> low standards of admission to and exit from teacher education programs, too many "mickey mouse" courses, overemphasis on professional education courses in relation to academic or liberal arts courses, "educationist" control of state departments of education and the certification function, the field of education lacking a distinct "disciplinary" base, the weakness of the doctorate in education and the commensurate intellectual weakness of the education faculty, and perhaps most important of all, the absence of reliable scientific evidence that any component of the teacher education programs has a predictable relationship to effective classroom teaching. (p. 27)

Ultimately, the responsibility for teacher education became an endeavor divided between liberal arts departments and the programs in education. Not surprisingly, this division of labor became divisive. Many faculty members simply did not consider teacher education to be a discipline worthy of university status and, as such, were not inclined to accept professional educators as academic equals.

The question of what elements should comprise teacher education and certification has long been a source of conflict. Angus (2001) notes that conflicts existed between:

> classroom teachers and other elements of the profession, between professors of education and professors in the liberal arts and sciences, between state department of education officials and those teaching in universities, between so-called "research universities" and former teachers colleges—this issue has

> been one of the most contentious in the long development of teacher certification. (p. 2)

In 1961, James Conant, former president of Harvard University, was asked by a group of education leaders to conduct the first large-scale study of teacher preparation. The study culminated in a seminal work, which did much to arouse a vigorous national debate concerning the education of teachers. According to Olwell (2005), Conant sought "to end the monopoly of colleges of education and locate more teacher preparation coursework in the subject matter departments, such as history, in colleges of arts and sciences" (p. 33). In the preface to his landmark 1963 book, *The Education of American Teachers*, Conant commented on the tension between liberal arts departments and education programs. He wrote:

> Early in my career as a professor of chemistry, I became aware of the hostility of the members of my profession to schools or faculties of education. I shared the views of the majority of my colleagues on the faculty of arts and sciences that there was no excuse for the existence of people who sought to teach others how to teach . . . . When any issues involving benefits to the graduate school of education came before the faculty of arts and sciences, I automatically voted with those who look with contempt on the school of education. (pp. 1-2)

Conant's beliefs about professional education softened over the years, but such beliefs persisted. Conant believed that colleges of education, if left to their own devices, were not likely to integrate more subject matter courses into teacher education (Olwell, 2005). For this, greater involvement by subject matter specialists would be

required. Although personally dismissive of many of the criticisms leveled by liberal arts faculty, Conant chided liberal arts academic faculty for failing to understand the importance of their role in training teachers and declared that teacher preparation should be a university-wide collaboration. Angus (2001) notes, however, "educators were outraged by his conclusion that the only portion of professional education that was clearly necessary was a high quality student teaching experience" (p. 27). Despite the clarity of Conant's observations and the prescriptive nature of his recommendations, not all education professionals were eager to give credence to his suggestions. If for no other reason, perhaps it is human nature to reject ideas originating from outside the profession.

In an attempt to understand this rift, researchers found that several factors contributed to this negativism. Teacher educators had more modest backgrounds than most other faculty members in higher education, worked in lower schools for many years, and seldom ventured far from home to attend college or graduate school (Lanier & Little, 1986).

Tinsley and Hardy (2003) suggested that while faculty members in other disciplines directly pursued their undergraduate and graduate programs uninterrupted, teacher educators often progressed on non-traditional timetables, sporadically taking classes as their schedules permitted. In addition, teacher educators "…were to be forever contaminated by their exposure to public schools. Both academically and socially, teacher educators were, from the outset, destined to be set apart from the rest of academe" (Tinsley & Hardy, 2003, p. 3). In considering the esteem of education

professionals, Reynolds (1995) suggests that teacher education is an unwelcome and unworthy intruder among the "real" fields of study in the liberal arts. Such an attitude is commonplace in the academic culture of many institutions of higher learning in the United States.

Despite teacher educators being held in low esteem by others, research findings by Tinsley and Hardy (2003) indicated that "teacher educators in Texas are able to maintain levels of professional self-esteem that are significantly higher than the levels of professional esteem they perceive from their academic colleagues in other departments" (p. 4). Conant (1963) suggested that in order for teachers to be regarded as true professionals, they must possess some esoteric body of knowledge that sets them apart from others whose general education is at least equivalent to their own.

Harboring a negative view of colleges of education, faculty in liberal arts departments wanted control over their curricular offerings. Even when capitulating to the needs of education programs, liberal arts faculty were reluctant to offer "watered down" versions of courses for the benefit of prospective teachers. Noddings (1998) notes that "instead of facing up to the problem squarely, education professors often insisted that pedagogical knowledge is more important than subject matter knowledge" (p. 87). Darling-Hammond (1997) obliquely asserts that too much subject matter knowledge can interfere with effective teaching by insinuating that Albert Einstein would likely have been ineffectual in organizing the work of 3rd graders. Alluding to the findings of three decades of research, Darling-Hammond (1997) further contends that only a "…threshold level of subject matter knowledge is

important and that knowledge of how to teach is even more important" (p. 308). Littleton (2000) argues that because poor certification exam performance is concentrated primarily in the social sciences, mathematics, and English, "the arts and sciences are not fulfilling their responsibilities for preparing teachers and are holding the university hostage by placing the university accreditation in jeopardy" (p. 8).

This divisiveness is likely rooted in the philosophical debate as to which is more important for success in the classroom—content knowledge or pedagogical knowledge. Although it seems reasonable to presume that content knowledge and pedagogical knowledge are equally important, proponents on each side of the issue present arguments for why more curricular content is needed from their respective domains. The issue of pedagogy is a contentious area and some observers believe too much emphasis is placed on pedagogy (Boyd, Goldhaber, Lankford, & Wyckoff, 2007). Tinsley and Hardy (2003) contend that as education programs were assimilated into larger institutions, narrow-minded academic colleagues believed that teachers are born, not made, and that pedagogy is easily learned. Subject matter expertise was considered the *sine qua non* of professional competence. Some believed that teacher education was not even a legitimate subject. Undoubtedly, educators will claim that the teacher preparation curriculum—including student teaching and pedagogical coursework—is at least as important as content knowledge (Darling-Hammon & Youngs, 2002).

Debating the Efficacy of Certification Exams

Educator interest groups challenge the notion that a single, high-stakes certification exam is sufficiently valid to declare someone fit to teach. Some opponents of the certification exams contend that, as a construct, the nature of teaching is far too complex to be properly judged by the results of a single exam. Use of a single criterion cannot adequately measure whether a teacher possesses the complete set of competencies for success as a teacher (Weitman, 1985). Instead, opponents of certification testing argue that teachers should be judged by what they do in the classroom (Ferguson & Brown, 2000; Harrell, 2009).

Because of the differential passing rates for minority groups taking the exam, discussions regarding test validity and bias also entered the public debate (Gratz, 2000; Justice & Hardy, 2001; Madaus & Pullin, 1987; Rubinstein et al., 1986). One of the major validity problems with certification tests is that "they are not job relevant, because they do not represent what teachers do on the job" (D'Costa, 1993, p. 109). The validation of a passing score does not prove that the score is true or correct. Rather, validation supports the plausibility, credibility, defensibility, and appropriateness of the passing score (Kane, Crooks, & Cohen, 1997).

D'Costa (1993) notes that a job-relevant teacher certification exam must meet three criteria:

- Include the major teaching functions and responsibilities.
- Represent the context of the teaching practice.
- Be based on a body of knowledge sufficient to be effective.

Because the teaching role is complex and difficult to define and translate into required skills, a single test is unlikely to capture all the nuances needed to assess competency. In addition, the teaching profession is not in agreement regarding the criteria of success for teaching. D'Costa further argues that if the validity of certification exams cannot be guaranteed, then assessment is pointless.

Walsh (2001a) argues against the efficacy of certification. Citing numerous research studies, Walsh notes that there exists "a scientifically sound body of research conducted primarily by economists and social scientists, revealing the attributes of an effective teacher, defined as a teacher who has a positive impact on student achievement" (p. iv). This body of research contends that certified teachers were no more effective than those who were not certified (Kirkpatrick, 1992; Walsh, 2001a). Boyd et al. (2007) note that overall, "research suggests that requiring certification exams does not result in a higher proportion of good teachers being selected but does reduce overall participation in teacher preparation" (p. 59).

Verbal ability is the one attribute found to be consistently related to teacher effectiveness. Most researchers consider verbal ability to be a measure of general cognitive ability, with selectivity of college most likely a suitable proxy for this attribute. Weitman (1985) contends that, because many of the skills for effective teaching are not measurable, items used for measuring teacher competency simply cannot represent a fair and universal sample of that skill set. Arguing against the use of a single, high-stakes test for determining competency, Ward and Wells (2006) suggest that it is reasonable to allow teacher educators to make professional judgments

and certification decisions regarding those who have demonstrated their ability to teach effectively.

Certification testing declares as substandard all those who fail the exam regardless of other attributes that research has shown correlates with effective teaching (Walsh, 2001a). Walsh concludes that certification testing is not capable of distinguishing between justifiable and unjustifiable reasons for denying access to the teaching profession. However, "educators, policymakers, the media, and the public mistakenly equate teacher quality with teacher certification" (p. 1).

As proponents of certification, Darling-Hammond and Youngs (2002) strenuously refute the arguments against certification presented by Walsh (2001a). The interpretation of the scientifically based research cited by Walsh (2001a) in support of the position against certification was refuted point-by-point as being flawed due to interpretation errors and misrepresentation of findings and recommendations from a large number of sources. In addition, Darling-Hammond and Youngs contend that unsupported statements, as well as a selective choice of research findings, resulted in specious conclusions. Walsh (2001b) subsequently issued a rejoinder, systematically refuting the refutation. With such vigorous debate on both sides of the issue, the matter of the efficacy of certification exams is far from settled.

If teachers' scores on certification exams are valid predictors of success in the classroom and non-white candidates are disproportionately impacted, then the interest of the low-scoring candidates is pitted against the right of children to receive a quality education (Ferguson & Brown, 2000). Some researchers have found that language

skills are a factor contributing to the disproportional impact on non-whites, and that certification exams are as much a reading test as a subject matter test (Harrell, 2009; Holmes, 1986; Ivie, 1982; Tellez, 2003; Ward & Wells, 2006).

Test Bias

In studying the difference in test performance among various population subgroups, Scheuneman and Slaughter (1991) describe five types of bias that could render test results suspect: historical, cultural, biological, educational, and psychometric (problems with test construction). According to Tellez (2003), the question regarding tests of pedagogical or content knowledge for preservice teachers "is not if they are biased, but rather, who is disadvantaged or privileged by the test and is the level of bias acceptable?" (p. 14).

Wakefield (2003) suggests that there may be unintended consequences of the bias in high-stakes testing. Teacher education professionals would no longer be entrusted with evaluating special cases. Instead, "testing replaces a personal review of strengths and weaknesses and two years of face-to-face encounters" (p. 386). Wakefield further argues that, because of false rejections, the public would be deprived of otherwise capable teachers and "some passionate and gifted teachers will fail to find places of service in public schools" (p. 386). Entry exams often serve to block the admission of many minority and low-income candidates into teacher education programs, while certification exams blocked entry into the teaching profession (Wakefield, 2003).

Millman (1989) contends that a false positive is a more serious error than a false negative because there is more potential harm by allowing an incompetent teacher to teach than unjustly denying a great teacher access to the profession. Because a student could take a certification exam an unlimited number of times, Millman (1989) suggests that a candidate "whose true level of functioning is below the passing standard may nevertheless pass the test on one of their attempts" (p. 5). Millman further argues that the more times a candidate takes the exam, the more likely it is that the competence level will be overstated, thereby increasing the likelihood of a false positive. By allowing multiple opportunities to test, measurement errors will increase the ratio of false positives to false negatives. According to Harrell (2009), "allowing a teacher candidate to test and retest for an unlimited number of attempts does present a challenge to the preservation of test validity" (p. 76).

Passing scores for certification exams should achieve the goals of the credentialing program while avoiding the detrimental effects of false positives. The passing standard should be high enough to provide adequate protection for the public, but not so high that the supply of qualified practitioners is restricted or that competent candidates are excluded from practicing (Kane et al., 1997).

This disproportional impact means that minorities and other disadvantaged groups are not proportionally represented in the classroom. In addition, the careers of minority candidates are often derailed, despite having a university degree (McIntosh & Norwood, 2004).

According to McGaghie (1991), five reasons describe the widespread dissatisfaction with professional competence evaluation:

- Exams tend to focus on a narrow range of practice situations, in contrast to what happens in reality.
- Exams tend to be biased toward assessing acquired knowledge.
- Exams pay little attention to the direct assessment of practical skills, despite claims to contrary.
- Exams are inadequate for assessing personal qualities such as honesty, judgment, maturity, stability, adaptability, etc.
- Almost all exams suffer from a variety of measurement issues, such as validity.

Although certification exams serve as the gateway for entry to the profession, the exams are also intended to protect the public from incompetent practitioners of a profession (Gorth & Chernoff, 1986; Millman, 1989; Roth, 1996). However, Angus (2001) notes that professional educators assert "an equally important purpose of certification is to protect the members of the profession from unfair competition from untrained people…" (p. 22).

Conant (1963) cites four reasons why professional educators staunchly support certification. First, leaders in the field of education believe that there exists a specific body of knowledge that can be taught, and that being trained in the art of teaching makes one a better teacher. Second, requiring specialized training and certification serves a gatekeeping function to control the quality of those seeking to enter the

profession. Without this, anyone with a college education could teach—no experience or training required. Under such conditions, the market for teachers would be flooded and wages deflated. Third, professional certification would serve as a "badge of identity" uniting all members of the profession, while excluding those outside of the profession. Lastly, it is believed that specialized teacher training could serve to insulate a teacher from contentious situations with parents who may have a comparable educational background. Having the certification would allow a teacher to defend, with credibility, decisions made in the classroom (pp. 27-28). If nothing else, certification testing might at least improve the public image of the profession (Madaus & Pullin, 1987).

Hess (2001) states that three assumptions undergird the case for certification:

- Training received during the certification process is so important that those without the training could not perform competently.
- Certification screens out those unfit for the profession.
- Certification elevates the prestige of the teaching profession.

Skeptical of the efficacy of certification, Hess notes that there is considerable debate as to whether certification effectively screens out unacceptable teachers and whether the professionalism of teachers is enhanced. In addition, Hess contends that within the teaching profession there is no well-established, research-based canon of essential knowledge and that the profession has "refused to establish a specific, measurable body of skills or knowledge that teachers must master" (p. 9). With such an inchoate

professional framework, one might speculate whether education professionals really do understand what constitutes mastery.

Because there is widespread agreement that interested observers such as colleagues, supervisors, and families can generally gauge the effectiveness of individual teachers within a given context of students, Hess is not convinced that standardized licensing helps to safeguard teacher quality. If the professional canon is subjective and lacking in specificity, Hess is skeptical of the efficacy of the certification process and posits that certification is a "potentially pernicious way to control quality" (p. 9).

The ExCET and TExES Certification Exams

In 1986, Texas began using the Examination for the Certification of Educators in Texas (ExCET) testing program to assess knowledge of pedagogy and professional responsibilities as well as subject knowledge. Amid the controversy of a deteriorating education system, "the state increased the rigor of the ExCET testing program and began utilizing ExCET results as the primary tool with which to evaluate teacher preparation programs" (Gruver, 2008, p. 4). These tests required applicants to move beyond simple recall of information, requiring "knowledge of patterns and relationships within and between inquiry skills and elements of content within a discipline and to draw relationships between the discipline and real-world situations" (Tackett, 1997, p. 65). Under the new accountability procedures, there have been instances of institutions across the state having been placed under SBEC review

primarily because of the poor performance of their African American students on the ExCET (McIntosh & Norwood, 2004).

In 2002, the Texas Examinations of Educator Standards (TExES) replaced the ExCET. Although functionally similar, the TExES was more rigorous than the ExCET, more closely aligned with the state's public education curriculum, and measured knowledge in pedagogy and subject matter. In terms of content, the History TExES was essentially the same as the History ExCET, with one exception—the TExES version contained an additional instructional domain focused on the teaching of "doing history" (Gruver, 2008).

In order to become certified, candidates were still required to pass the History TExES as well as an exam assessing their pedagogical knowledge and learning theories. According to Gruver (2008), many history faculty members at state universities in Texas did not believe that the History TExES was a valid measure of a candidate's understanding of, or ability to do, history. Facing the possibility of sanctions by state agencies if their history majors do not pass the test, many departments of history acquiesced and aligned their curricula with the domains and competencies of the exam.

Because of the low scores of students taking the TExES certification exam in the History (8-12) content area, it is important to identify students who are at risk of failing. An "early warning system" capable of identifying students who are likely to fail the exam allows for intervention so that at-risk students have a greater chance of success.

Low pass rates can create a number of detrimental effects on the institution, the student, and society. Examples of such adverse effects include:

- Low pass rates reflect poorly on the school and its programs.
- Low pass rates signal poor program quality and can affect the number of students seeking admission to a program.
- Low pass rates can affect the accreditation status of a program.
- Students who fail the exam are denied timely entry into their profession and may suffer an opportunity cost, which affects them financially and psychologically.
- Students' careers may be placed on hold, or they may simply "throw in the towel" and give up on their aspirations of a teaching career.
- Low pass rates may impact the mission of producing highly qualified teachers.

Pass Rates

The Texas State Board for Educator Certification implemented a performance-based accountability system to assure the quality of teacher education programs. Perhaps the most visible and significant measure of the success of a program are the pass rates on teacher certification exams. The simplest, most widely reported, and most useful measure is the initial pass rate (Pitter, Lanham, & McGalliard, 1997). Such measures are reported in aggregate as well as being disaggregated by gender and ethnicity. An institution must achieve acceptable pass rates in all categories to prevent any type of state-imposed sanctions on its programs. A pass rate that falls

significantly below the national average is a cause for concern. However, a high pass rate does not necessarily imply a high quality program. As noted by Pitter et al. (1997), "it is possible that two programs of similarly high pass rates have vastly different levels of quality, where one program simply prepares the students to pass the examination, while the other goes much further" (p. 5).

The SBEC issues annual accreditation ratings based on the combined performance of all candidates. The data are also disaggregated into groups based on gender and ethnicity. Ratings are based on how well a program's candidates perform on the assessments required for certification.

Certification exams are usually taken near the end of a candidate's preparation program; however, candidates must receive approval from program administrators before they can register for the exams. Assessment data measure two different types of pass rates. The *initial pass rate* includes tests a candidate has taken between September 1, yyyy through December 31, yyyy + 1 during which the candidate completed the program, even if the test is taken multiple times. The *final pass rate* is based on performance during the subsequent year and cumulatively measures results over the entire two-year period. If fewer than 30 students are involved, current data are combined with data from the previous one or two years (SBEC, 2001).

According to Kane et al. (1997), there is no single, correct performance standard. The setting of a standard is not a problem of estimation; rather, it is a policy decision. The regulating agency decides what constitutes the minimum acceptable performance level. The decision hinges on a number of factors including the risk

posed by marginal candidates, current standards of practice, and the political environment. The ideal standard should protect the public from incompetent practitioners while at the same time being fair to the candidates so the supply of practitioners is not unduly restricted (Kane et al., 1997).

To maintain accreditation, a program must achieve a 70% first-year pass rate or an 80% two-year, cumulative pass rate across all demographic groups (all, male, female, African American, Hispanic, White, Other) (SBEC, 1998). Individually, however, candidates must achieve a minimum score of 80% on the TExES History (8-12) exam. As shown previously in Figure 1.1, both the XTU and statewide initial pass rates have been below the 70% minimum since 1999, although in 2008 the XTU pass rate was slightly above 70%.

Research Antecedents

Because of the nature of high stakes testing and the consequences, there is good reason to attempt to predict which students are at risk of failing the exam. Although there have been a number of studies attempting to predict success on the PPR portion of the TExES, few studies have focused on predicting the outcome on content-area exams. Because of this, little is known about the variables that affect student success on the History (8-12) teacher certification exam.

Because there is little available research to aid in understanding the factors that contribute to the success of teacher education candidates on secondary-level content-area certification exams, education programs may fall short of their mission of

producing highly qualified teachers. The current study focuses specifically on predicting failure on the History (8-12) content area exam.

Researchers studying similar phenomena in a variety of educational environments and situations have generally preferred the use of stepwise multiple linear regression for predicting exam scores. Regression analysis is one of the most widely used techniques for prediction and the methodology is used in a wide range of situations when there is a need to predict the success or failure of individuals based a number of independent variables. There are, however, a number of problems with this approach.

In reviewing the literature, numerous studies in a variety of contexts other than education were examined, e.g., health sciences (Harwell, 1989; Henderson & Orr, 1989; Larsen, 2002; Lee, 1980), business (Barilla & Jackson, 2008), computer science (Garcia, 1987), and psychology (Yu et al., 1997). In the context of education, most researchers seeking to produce statistical models for predicting outcomes on the exams used stepwise multiple regression analysis, although logistic regression was also used in some studies (Byrd & Williams, 2006; Jackson, 2006; Pisani, Pisani, & Anderson, 2002; Poelzer, Zeng, & Simonsson, 2007; Zeng, Simonsson, & Poelzer, 2002; Weiss, 2003). In using logistic regression, the dichotomous, categorical dependent variable was defined as pass/fail and the model was used to predict the probability of passing, rather than predicting a raw score.

Regardless of the context, the limitations of stepwise regression were often the same—lack of statistical support for validating regression model assumptions, the

focus on a numerical score rather than a pass/fail criterion, failure to address the possibility of interaction among independent variables, failure to address the issue of multicollinearity, use of confidence intervals rather than prediction intervals in judging the aptness of the prediction model, and low R^2 values. The current study was undertaken to address some of these methodological limitations

Within the context of teacher certification, most of the studies involved the PPR certification exam (e.g., Alexander, 1990; Chambers, Munday, Sienty, & Justice, 1999; Hernandez, 1999; Kinnison & Nolan, 2001; Laird, 1998; McDonald, 2000; Nance & Kinnison, 1988; Poelzer, Zeng, & Simonsson, 2000; Poelzer, Zeng, & Simonsson, 2007; Simonsson, Poelzer, & Zeng, 2000; Ward & Wells, 2006; Weiss, 2003; White & Burke, 1994; Zeng, Simonsson, & Poelzer, 2002). Only five studies were identified that focused on predicting success on a TExES content-area exam (Gruver, 2008; Jackson, 2006; McIntosh, 2002; Pisani, Pisani, & Anderson, 2002; Weiss, 2003). Gruver (2008) focused explicitly on the History (8-12) certification exam. Because content-area studies are the most relevant to the current study, only the five studies involving TExES content-area exams are specifically discussed below.

Pisani, Pisani, and Anderson (2002) sought to identify the factors for success on the ExCET exam by studying the variables affecting the pass/fail rate on the social studies content exam. A predominantly Hispanic group of students (n = 367) from a predominantly Hispanic university in South Texas was studied using logistic regression to predict the probability of pass/fail of the ExCET exam (dependent variable) score using the independent variables of gender, age, ethnicity, TASP

reading score, coursework preparation, and GPA. The predicted pass/fail rates were compared to the actual results using the proportional chance criteria. The proportional chance criteria compares a model's predictive accuracy relative to chance. A good discriminating model will predict twenty-five percent better than chance. If the model exceeds this threshold, then the model is considered statistically significant.

The researchers evaluated three domains: (a) Bilingual Elementary Comprehensive (Hispanic), (b) Bilingual Elementary Comprehensive—Social Studies Component (Female), and (c) Bilingual Elementary Comprehensive—Social Studies Component (Hispanic). All three models predicted "pass" more accurately than predicting "fail." The findings, in general, revealed a strong link between TASP reading ability score, GPA, and age with favorable ExCET exam performance. The study also revealed that non-Hispanics scored better on the ExCET exam than Hispanics, although the researchers did not investigate the reasons for this. The authors explained that because very few non-Hispanic students were included in the sample, it may have influenced the results in comparing Hispanics to non-Hispanics.

Using correlation and stepwise linear regression, McIntosh (2002) studied the relationship between the performance on subject area specialization, TASP scores (reading, mathematics, and writing), and the results on the Texas ExCET Elementary Comprehensive certification exam. A convenience sample (n = 280) was selected from the College of Education Teacher Education Program at the University of Houston.

One-tailed tests were used to determine the statistical significance of the five research hypotheses. One-tailed tests are less rigorous than two-tailed tests, making it easier to reject the null hypotheses and to declare the results statistically significant.

The correlation coefficient for subject area specialization courses was not statistically significant ($r = +.15$, $p = .146$). The correlation for three other independent variables were statistically significant—reading achievement ($r = +.49$, $p < .001$), mathematics achievement ($r = +.39$, $p < .001$), and writing achievement ($r = +.31$, $p < .001$). The researcher noted that despite statistically significant correlational results for three of the independent variables, the variables were weak predictors that accounted for little of the variance in the criterion variable.

The stepwise linear regression model was statistically significant, $R^2 = .32$, $p < .001$). McIntosh concluded that the model is a weak predictor of performance on the ExCET exam. Although criterion variable confidence intervals were not provided, the standard error of the estimate was 6.54. A 95% confidence interval would be approximately +/- 15 points. A confidence interval this large makes it difficult to determine whether someone was likely to pass or fail the certification exam.

Weiss (2003) conducted a study seeking to understand the extent to which certain factors might predict performance on the Texas ExCET certification exam (dependent variable) by elementary education teachers. Independent variables included cumulative GPA, English GPA, and TASP scores (reading, writing, and mathematics). The criterion variables were defined as components of the ExCET certification exam concerned with Professional Development and Elementary

Comprehension. Because the ExCET exam may be attempted multiple times, only initial test scores were used in the study. Participants in the study (n = 198) were selected from undergraduate students enrolled in a teacher education program in the fall semester of 1998 and the spring semester of 1999 at a large urban university in Texas.

Results of the correlational analysis revealed that all of the independent variables were highly correlated with the criterion variable ($p < .05$). All independent variables were entered into the regression model. For Professional Development, the overall model was statistically significant ($R^2 = .235$, $p < .001$). For Elementary Comprehension, the overall model was statistically significant ($R^2 = .308$, $p < .001$). Although criterion variable confidence intervals were not provided for either model, the standard error of the estimate for Elementary Comprehension was 7.63. A 95% confidence interval would be approximately +/- 16 points. A confidence interval this large makes it difficult to determine whether someone was likely to pass or fail the certification exam. Weiss concluded that although GPA was included in the overall model, it is not a strong predictor of performance on the ExCET exam. However, the TASP reading score was found to be the strongest predictor of success.

Jackson (2006) investigated the relationship between content knowledge and scores on the TExES certification exam (criterion variable). Independent variables included: (a) the number of upper-level content area courses, (b) the upper-level content area grade point average, (c) the number of months between the last upper-

level content area course was completed in the certification field and the month the student initially attempted the TExES exam.

Multiple content area exams were used as criterion variables. These included: English Language Arts and Reading (8-12), History (8-12), Life Science (8-12), Mathematics (8-12), and Social Studies (8-12). A convenience sample was used to select 144 students who took the TExES exam and were seeking secondary teacher certification through the Secondary Online Post-Baccalaureate Teacher Certification Program at the University of North Texas during the 2002-2003, 2003-2004, and 2004-2005 academic school years. The sample included all students in the Online Post-Baccalaureate Secondary Teacher Certification Program who were seeking initial certification in one of the following certification fields: grades 8-12 English Language Arts and Reading, grades 8-12 History, grades 8-12 Life Science, grades 8-12 Mathematics, and grades 8-12 Social Studies. Thirty-six students took the History (8-12) certification exam.

The analysis of variance revealed significant differences among the five test groups. Only the three-predictor model for the History (8-12) test group was statistically significant, $F(3,32) = 3.753$, $p = .02$, $R^2 = .249$. The correlation between the TExES exam and the number of upper-level content area courses completed in the certification field was not statistically significant ($r = -.057$, $p = .741$). The correlation between the TExES exam and upper-level content area GPA in the certification field was statistically significant ($r = .348$, $p = .038$). There was a statistically significant correlation between the TExES exam and the number of months between when last

upper-level content area course was completed in the certification field and the month the student initially attempted the TExES exam (r = .353, p = .035). A logistic regression analysis for the three-predictor model for the History (8-12) test group revealed that upper-level content area GPA was the only variable that was statistically significant (p = .017).

Upper-level grade point average accounted for approximately 12% of the variance among scores within the History (8-12) test group. Months of time elapsed between last content area course work and the initial state content examination accounted for approximately 13% of the variance among scores.

Gruver (2008) sought to determine the relationship among selected variables and the TExES History (8-12) exam score of first-time test takers at three regional state universities. Using a convenience sample, data were collected from 119 first-time test takers who attended the three regional state universities and took the History (8-12) TExES from October 2002 to August 2007. Data were based on results tabulated in reports prepared in conjunction with the Texas Accountability System for Educator Preparation. Of the 119 subjects, only 81 completed cases were used in the regression analysis. The TExES History (8-12) exam score was regressed on five predictor variables, which included general education GPA, history (teaching field) GPA, Texas Academic Skills Program (TASP)/Texas Higher Education Assessment (THEA) Reading score, age at the time of the History TExES, and gender. The research hypothesis was tested using standard multiple linear regression, using an alpha level of .05 to determine statistical significance.

The results indicated a statistically significant relationship between History (8-12) TExES score and the predictor variables, $F(5, 75) = 6.300$, $p < .001$. The TASP/THEA Reading score was the only statistically significant factor ($r = .532$, $p < .001$) and accounted for nearly all of the variance in predicting scores on the History TExES. A large effect size was observed with approximately 30% of the variance accounted for by the model ($R^2 = .296$).

Chapter Summary

The literature review traced the historical path leading to America's current system of high-stakes testing and performance-based accountability. Criticisms of higher education are nothing new and ample evidence supports the notion that the American system of higher education has been in decline. Disenchanted with the state of affairs, policy makers and the public in general have embraced a system of performance-based measures of accountability.

The need for high stakes testing and accountability creates significant consequences for the individual as well as institutions of higher education. Criterion-referenced certification exams were implemented to measure teacher competence. Because certification exams have had a differential effect on minorities, some have alleged that the exams are biased against minorities. There has been considerable debate as to the efficacy of the certification process in general.

Because the consequences of failing certification exams are serious, the need for predicting outcomes is important. However, previous research demonstrated how ineffective the predictive models have been. Even if the models are statistically

significant, the matter of practical significance, practical application, and suitability to task are still issues of concern.

Because little existing research has focused on content area exams for teacher certification, there is a significant gap in the literature. This study seeks to broaden the base of knowledge in this area by employing classification trees as an alternative methodology for predicting outcomes on the TExES History (8-12) certification exam.

CHAPTER III

METHOD

This study investigated the predictive ability of selected variables of teacher certification candidates who took the History (8-12) Texas Examinations of Educator Standards (TExES) test between 2002 and 2008. There are a number of pathways to certification, e.g., undergraduate field-based program, post baccalaureate field-based program, or post baccalaureate alternative certification program. Because the various pathways to certification are considered equally legitimate and effective, a candidate's pathway to teacher certification was not considered in the analysis. This study focused only on the performance and selected predictor variables of first-time History (8-12) TExES test takers. The dependent variable was the History (8-12) TExES exam result (pass/fail). The independent variables included various academic, testing, and demographic factors. The purpose of this chapter is to describe the research rationale and general procedures used to conduct the study.

In Chapter I, Figure 1.1 showed a pattern of low pass rates on the TExES History (8-12) certification exam. Because of limited prior research, there is a gap in the literature for identifying the factors associated with the performance of teacher certification candidates on TExES secondary-level content area exams in general, and the History (8-12) exam in particular. Methodological limitations identified in prior studies can be ameliorated by employing a different methodology.

Restatement of Research Questions

Because of limited prior research, methodological limitations, and the importance of high stakes testing in Texas, this study addresses the following research questions:

1. What are the descriptive statistics for XTU candidates taking the History (8-12) TExES certification exam for the first time for the years 2002 – 2008?
2. Does a statistically significant relationship exist among any of the following variables?

   - TExES result (pass/fail)*
   - Transfer status
   - PostBac/Bac status
   - Gender
   - Ethnicity
   - Age at time of exam
   - Lower division GPA
   - Upper division GPA
   - Total GPA
   - History GPA
   - Total number of history courses
   - Number of upper division history courses
   - TASP scores (Read/Write/Math)
   - ACT scores (Read/Math)
   - SAT scores (Read/Math)

   * dependent variable

   (H_0: no statistically significant relationship exists)
3. Does a nonparametric, classification tree methodology produce a better predictive model for correctly classifying group membership (pass/fail on History exam) at a statistically significant level compared to chance as measured by Press's Q test statistic ($p < .05$), and the proportional chance criterion?

   (H_0: classification tree model is no better than chance)

Research Design and Rationale

The current study is an exploratory data analysis using classification trees—a nonparametric statistical technique often associated with data mining. Exploratory data analysis is used to identify systematic relationships among variables when there are few, or no, a priori expectations as to the nature of those relations.

The current study examined the results of the History (8-12) TExES certification exam to identify variables useful for predicting failure on that exam. With the ability to predict academic outcomes, institutions can initiate interventions before the student is even aware of the risk.

Correlational analysis (Pearson product-moment correlation) was used to analyze the strength and direction of the relationship among variables. Descriptive statistics are also presented.

Problems with Stepwise Multiple Regression

Researchers studying similar phenomena in a variety of educational environments and situations have overwhelmingly preferred the use of stepwise multiple linear regression for predicting the outcome of exams. Logistic regression was also used, but to a lesser extent. There are, however, a number of problems with these approaches.

In a number of the studies identified in the literature (discussed in Chapter II), researchers either neglected to test that underlying regression model assumptions were met, or failed to report it. Failure to meet the underlying assumptions means the results may not be valid, leading to inaccurate estimates of significance, effect size,

statistical power, and erroneous conclusions. These inaccuracies can manifest as incorrect measures of significance of the regression coefficients (e.g., indicating significance when it is not) and biased and inaccurate predictions of the dependent variable (Hair, Black, Babin, Anderson, & Tatham, 2006). Without testing the model assumptions, the validity of the results, conclusions, and assertions may be suspect. The use of stepwise regression is not recommended because this approach (based on the order in which variables are entered into the model) results in inflated risks of Type I error and does not include higher-order or interaction terms (McClave, Benson, & Sincich, 2008; Warner, 2008).

Stepwise regression should be used only when necessary, and then only as a variable screening tool for identifying potential variables to be used in the model-building process. Stepwise regression should not be used as the final model for predicting the outcome variable (Kutner, Nachtsheim, Neter, & Li, 2005; McClave, et al., 2008). Warner (2008) summarizes the argument against stepwise methods by stating that the use of stepwise regression models "often yield analyses that are not useful for theory evaluation (or even for prediction of individual scores in different samples)" (p. 551).

Berk (2004) is critical of stepwise regression because the "procedures used to select the correct model can be very misleading" (p. 133). In addition, Berk notes that (a) the selected model may not be appropriate for the situation, (b) stepwise methods tend to capitalize on chance, and (c) the "best" model may be substantive nonsense.

He summarizes his argument by stating, "there is no necessary correspondence between a selection criterion and a scientific criterion" (p. 133).

Confidence Intervals versus Prediction Intervals

Even if the regression model assumptions are met, the problem of large standard errors, and therefore large confidence intervals, cause the models to be of little use for the intended purposes. By assigning specific values to the predictor variables, multiple regression produces a predicted value for the *average* outcome for all cases having that specific combination of values for the predictor variables. Because predictor variables often represent continuous data, the probability is exceedingly small that a reasonable number of subjects would have the identical combination of continuous and categorical values for the predictor variables. For example, each candidate would likely have a different combination of TASP scores for reading, writing, and math.

Prior research utilizing stepwise multiple regression analyses generally resulted in low R^2 values, and therefore, low predictability. R^2 measures how well the regression model "fits" the data, and the standard error of the estimate measures the accuracy of the estimates produced by the regression model (Larose, 2006). Low R^2 values resulted in large standard errors of the estimate (approximately seven to eight points). A 95% confidence interval would therefore have a width of approximately +/- 15 points, which is of little practical value for predicting scores. For example, a predicted score of 80 would have a 95% confidence interval of 65 – 95. With an interval this large, we would not know whether an intervention was appropriate except

in extreme cases, which would then be obvious even without using a predictive model. To intervene when it is not necessary would be a waste of resources, while not intervening when it is necessary increases the risk of a student failing the exam.

Because of this situation, it makes sense to predict the outcome for a specific individual, which requires using the *standard error of the prediction* instead of the *standard error of the estimate* (the mean). This leads to the use of the *prediction interval* instead of the *confidence interval*. Because it is easier to predict the mean of a group attribute than to predict the value of that attribute for a specific member of that group, it is always the case that the prediction interval is wider than the confidence interval (Groebner, Shannon, Fry, & Smith, 2008). Using the prediction interval instead of the confidence interval produces results that are more uncertain and therefore less useful for identifying specific at-risk candidates.

Using a Categorical Criterion Variable

The criterion variable in multiple linear regression is a continuous variable. However, in the context of certification exams it is not the exact test score that is of particular interest; we are only interested in knowing if a candidate passed or not. In performance reporting, the SBEC focuses on pass rates, not specific scores. Because of this, it makes sense to employ a technique capable of handling a dichotomous, categorical dependent variable (pass/fail). Logistic regression can handle categorical dependent variables, but the interpretation and pragmatic application of the resulting model can be difficult for those unfamiliar with sophisticated statistical models.

Use of Nonparametric Methods

If violations of the regression assumptions are serious, then nonparametric techniques are recommended. Nonparametric statistical methods are useful because the techniques make no assumptions about normality, linearity, and homoscedasticity. If the assumption of normality is tenable, then parametric methods are likely to have more statistical power. However, if the distribution of the data is not normal, or not known, then nonparametric methods may provide better results and more statistical power (Conover, 1999; Higgins, 2004). Power is the probability that a researcher will be able to "prove" what he/she intended to prove. Conover (1999) points out that "nonparametric methods of analysis often make more efficient use of the data than parametric methods, when the parametric methods are inappropriately applied" (p. 3).

Data Mining and Classification Trees

Data mining is an analytic technique designed to explore data (usually very large datasets) in search of consistent patterns or systematic relationships among variables. These trends and patterns form the basis of predictive models, enabling a researcher to produce new observations from existing data. The researcher attempts to validate these patterns (or systematic relationships) by applying the detected patterns to new data. The challenge of data mining is finding an interpretable model that fits reasonably well, without overfitting. Overfitting occurs when the model fits the training data so well (perhaps precisely) that the model generalizes poorly with respect to future data. The most common data mining tasks are (a) description, (b) estimation, (c) prediction, (d) classification, (e) clustering, and (f) association (Dunham, 2003;

Larose, 2005). The methodology is well suited to *exploratory* data analysis, but not for *explanatory* data analysis (StatSoft, 2010).

Classification is one of the oldest data mining applications, dating back to the 1700s or earlier (Dunham, 2003). Classification tree algorithms develop a set of decision rules for predicting (classifying) the membership of future cases into the classes of a categorical dependent variable based on measurements on one or more predictor variables (AnswerTree 3.1, 2002; StatSoft, 2010). Classification trees (also called decision trees) are particularly well suited to situations with little a priori knowledge or coherent theories regarding which variables are related and how. Tree methods can often reveal simple relationships among just a few variables that could have gone unnoticed using other analytic techniques (StatSoft, 2010).

Classification trees can handle categorical predictors, continuous predictors, or any mix of the two types (Neville, 1999). Attractive as they are as an analytic tool, classification trees are not necessarily preferred over more traditional parametric methods. As an exploratory technique, or as a technique of last resort when traditional methods fail, classification trees are well suited as an analytical technique (StatSoft, 2010). In addition to rendering the issues of normality, linearity, and homoscedasticity irrelevant to the analysis, classification trees produce an intuitive graphical output making the interpretation and application of the model easier than a model with only a strict numerical interpretation (StatSoft, 2010).

Neville (1999) notes that classification trees also have their shortcomings. If the relationship is complex, a simple tree may be too simplistic. In addition, "a tree

gives the impression that certain inputs uniquely explain the variations in the target. A completely different set of inputs may give a different explanation that is just as good" (p. 3).

Generally, results summarized in a classification tree are easily interpreted, making the technique useful for quickly classifying new observations (e.g., the risk of failure for a particular student). Because of this, classification trees are well suited for applications where model comprehensibility is important (Saar-Tsechansky & Provost, 2007). Simple tree models are appealing because they clearly depict how observations are classified and target groups determined (Neville, 1999). For example, higher education institutions can use classification trees for analyzing student characteristics or predicting the likelihood of a variety of outcomes, such as persistence, retention, course success, and failure on high stakes certification exams.

Although classification trees are not widely used in social science research, the methodology is suitable for analyzing certain classes of problems in that realm. Any profession seeking to predict the success/failure of examinees on certification/licensure exams could easily implement the technique. A brief summary of the advantages and limitations of classification trees is presented in Table 3.1.

Table 3.1

Advantages and Limitations of Classification Trees

Advantages	Limitations
Simple to understand and interpret—model is easily understood after a brief explanation.	Algorithms not guaranteed to produce an optimal decision tree.
Easier data preparation—other techniques often require data normalization, creation of dummy variables, and procedures for dealing with missing data.	May create overly complex trees that do not generalize well—called overfitting. Splitting rules and pruning techniques are necessary to minimize this problem.
Ability to handle numerical and categorical data. Many other techniques require specific data types.	Over-sensitivity to the training set. A minor change in a split close to the root can dramatically change the subtree structure, resulting in tree instability.
A nonparametric method that makes no assumptions about data distribution.	Handling of missing data—may require significant computing resources in large data sets.
Can validate a model using statistical tests. Makes it possible to assess model reliability.	
Robust—performs well even for modest violations of model assumptions.	
Scalability—performs well even as data sets become increasingly large.	

Note. (Rokach & Maimon, 2008). Also, "Decision tree advantages" retrieved November 10, 2010, from http://en.wikipedia.org/wiki/Decision_tree_learning.

Figure 3.1 shows the graphical output of a classification tree analysis involving credit ranking. Table 3.2 shows a misclassification matrix (also known as a confusion matrix), showing the degree to which cases were improperly classified.

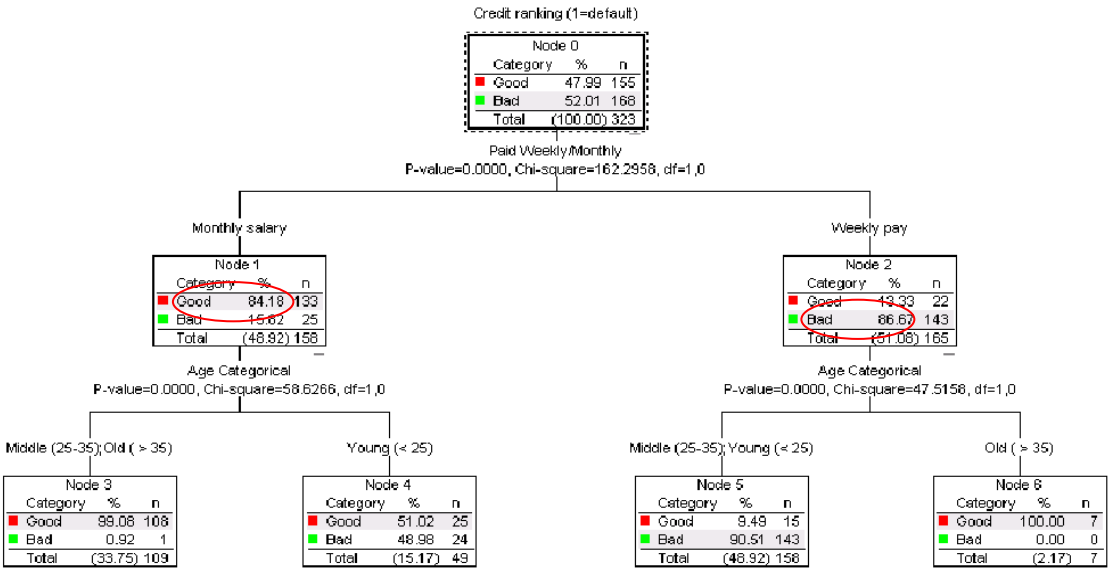

Figure 3.1. Classification tree example adapted from AnswerTree 3.1 User's Guide (2002).

As an illustration of the technique, Figure 3.1 shows that at the first split, predicting credit worthiness can be significantly improved relative to chance (the top level of the tree), by categorizing the cases as either monthly salary or weekly wage. Approximately 85% of the cases were correctly classified just by knowing whether a person is paid a monthly salary or a weekly wage, as opposed to 52% being properly classified based only on chance. Based only on chance, in the absence of any better information, a case would be classified according to which category had the highest proportion in the data set. For the current illustration, because 52% of the data were classified as Bad, an arbitrarily chosen case would therefore be classified as Bad, which produces the highest probability of being correctly classified. Classification tree methodology seeks to significantly improve the probability of correctly classifying cases relative to chance.

From Table 3.2, the risk of incorrectly classifying cases is ((25 + 22) / 323) = 0.145511, which means that the probability of correctly classifying the cases is approximately 85% (1 - 0.145511). The matrix shows, for example, that of the 158 cases predicted as "Good", 133 were correctly classified and only 25 were incorrectly classified. This illustration demonstrates the ease of interpreting the result of a classification tree analysis. A measure of effect size and statistical significance can be determined by evaluating the performance of the classification tree against Press's Q ($p < .05$), and the proportional chance criterion (Hair et al., 2006).

Table 3.2

Classification Tree Misclassification Matrix

Misclassification Matrix				
		Actual Category		
		Bad	Good	Total
Predicted Category	Bad	143	22	165
	Good	25	133	158
	Total	168	155	323
		Risk Statistics		
Risk Estimate		0.145511		
SE of Risk Estimate		0.01962		

Note. Misclassification matrix adapted from AnswerTree 3.1 User's Guide (2002).

Splitting algorithms.

Data mining is a machine learning system that derives decision rules from existing data. These decision rules are used to create the *splits* that form the classification tree. Splits on the predictor variables are used to predict membership in

the classes of the dependent variables for the cases or objects in the analysis. Because classification trees are hierarchical in nature, splits are determined one at a time, starting with the *root node* (the highest level) and continuing with splits resulting in *child nodes* until, based on a set of *stopping rules*, splitting stops. When a child node can be split no further, it is designated a *terminal node*. Comprehensibility typically decreases with an increase in tree size and complexity. With respect to the principle of parsimony, if two trees using the same kind of tests have statistically similar prediction accuracy, then the one with fewer leaves is usually preferred (Lim, Loh, & Shih, 2000).

Simple models are preferred for practical and philosophical reasons. Simple models are easier to test in replication and cross-validation studies, and are less costly to apply. Philosophically, simpler models are easier to understand, and therefore have a "beauty" that their more complicated counterparts often lack (StatSoft, 2010).

All classification algorithms perform essentially the same function of classifying cases into categories. Algorithms differ on their performance characteristics and features (AnswerTree 3.1, 2002). In a recent study, Lim, Loh, and Shih (2000) analyzed 33 different classification algorithms based on accuracy and performance. The results showed that the differences in mean error rates of many algorithms were statistically insignificant, which means the differences among algorithms were also probably insignificant in practical terms. In view of the findings of insignificance, the researchers suggest that the selection of an algorithm should

probably be based on other criteria, such as training time or interpretability of the model.

Lim, Loh, and Shih (2000) note that, unlike error rates, there were very large differences in training times among the algorithms. The algorithm with the lowest mean error rate, took about fifty times as long to train as the next most accurate algorithm. The magnitude of the time difference was in hours and was consistent over a wide range of sample sizes. For very large applications (perhaps on the scale of 10^4 – 10^6 cases) where time is a factor, using one of the quicker algorithms may be advantageous.

Of the many splitting algorithms available, five algorithms are the most common. AnswerTree 3.1 User's Guide (2002) and Clementine 11.1 Node Reference (2007) provide descriptions of the various algorithms and are discussed below.

Classification and regression tree (C&RT).

Classification and Regression Tree (C&RT or CART) is a binary tree-growing algorithm that partitions data into two subsets such that cases within each child node are more homogeneous than its parent node. The procedure can use any combination of continuous or discrete variables. Because it is a recursive process, it repeats itself until the homogeneity criterion (purity) is reached, or until some pre-established stopping criterion has been satisfied. A node is considered "pure" if 100% of cases in the node fall into a specific category of the target field. The same predictor variable may be used multiple times at different levels in the tree. It is biased toward choosing predictor variables with more levels for splits (StatSoft, 2010).

Because binary algorithms tend to grow trees that have many levels, the resulting tree may be difficult to interpret, especially if the same variable was used to split a number of successive levels. The procedure can be effective as a data reduction tool (identifying relevant variables) and for detecting interactions among variables. Because C&RT is complex, computation can take a long time with large data sets. The technique of *surrogate splitting* is used to handle missing values (AnswerTree 3.1, 2002).

Chi-squared automatic interaction detector (CHAID).

Chi-squared Automatic Interaction Detector (CHAID) is one of the most frequently used methods and it works for all types of variables. A highly efficient statistical technique, it uses a statistical test of significance for evaluating potential predictor variables. Values considered statistically homogeneous (similar) with respect to the target variable are merged and values that are heterogeneous (dissimilar) are handled separately. If multiple predictor variables are statistically significant, CHAID selects the predictor with the smallest p-value. Bonferroni adjusted p-values can be used to control Type I errors. The choice of which statistical test to use depends upon the measurement level of the target variable. If the target variable is continuous, an F test is used. If the target variable is categorical, a chi-squared test of independence is used. This category-merging process stops when all remaining categories differ at the specified testing level.

Unlike C&RT, CHAID is not constrained to binary splits. Therefore, it can produce more than two categories at any particular level in the tree, which tends to

create wider trees with less depth than strictly binary methods. Missing values are handled algorithmically by aggregating them into a single valid category (AnswerTree 3.1, 2002).

Exhaustive chi-squared automatic interaction detector.

Exhaustive CHAID, a variation of CHAID, was developed to address some of the weaknesses of the CHAID method. Occasionally, CHAID may not find the optimal split for a variable because it stops merging categories when it determines that all remaining categories are statistically different. Exhaustive CHAID mitigates this problem by finding the set of categories producing the strongest association with the target variable and computes an adjusted p-value for that association. Exhaustive CHAID finds the best split for each predictor, and then chooses which predictor to split by comparing the adjusted p-values.

Exhaustive CHAID and CHAID use the same statistical tests and handle missing data similarly. Because it is more thorough than CHAID, it takes longer to compute and sometimes finds splits that are more useful. However, depending on the data, there may be no difference between Exhaustive CHAID and CHAID results (AnswerTree 3.1, 2002).

Quick, unbiased, efficient statistical tree (QUEST).

Quick-Unbiased-Efficient-Statistical-Tree (QUEST) is a relatively new binary algorithm created for computational efficiency, but like C&RT, can produce trees of unwieldy size. QUEST performs approximately unbiased variable selection when determining splits. That is, if all predictor variables are equally informative with

respect to the criterion variable, predictor variables are selected with equal probability. Compared to C&RT, this lack of bias in variable selection for splits is an advantage when some predictor variables have few levels and other predictor variables have many levels. Predictors with many levels are more likely to produce "fluke theories," which fit the data well but have low predictive accuracy when generalized to future data (StatSoft, 2010). Unlike CHAID and C&RT, QUEST separates the issues of predictor selection and splitting, applying different criteria to each (AnswerTree 3.1, 2002). Like C&RT, QUEST uses surrogate splitting to handle missing values.

C5.0 algorithm.

The C5.0 algorithm involves the concept of *entropy* to measure the amount of uncertainty or randomness in a set of data. The less the entropy the more the information gain (Dunham, 2003). C5.0 splits on the field that provides the maximum information gain at each level. The target field must be categorical. Unlike binary algorithms, splits into more than two subgroups are allowed.

C5.0 can produce two kinds of models—decision trees and rule sets. A decision tree is a straightforward description of the splits found by the algorithm. Exactly one prediction is possible for any particular case presented to a decision tree. A rule set is a set of decision rules that attempts to predict the correct classification of individual cases. Rule sets are derived from decision trees, represent a simplified version of the information of the decision tree, and can produce less complex models.

C5.0 models are reasonably robust to problems arising from missing data and large numbers of input fields. The algorithm is very efficient in terms of computation

time. In addition, C5.0 models tend to be easier to understand than some other model types because rules derived from the model have a straightforward interpretation. C5.0 can employ the powerful technique of *boosting* to increase accuracy of classification (Clementine 11.1, 2007). With boosting, the general idea is to compute a sequence of very simple trees, and to derive weights to combine the predictions from those models into a single prediction (StatSoft, 2010).

Specifying the criteria for predictive accuracy.

The goal of classification tree analysis is to produce accurate predictions. However, an unequivocal operational definition of "accurate prediction" can be elusive. Often, accurate prediction is specified in terms of minimizing costs. The general idea is that the best prediction has the lowest misclassification rate (StatSoft, 2010). Sometimes, however, the cost consequences of misclassifications are not symmetrical. The cost of a false negative may be more (or less) costly than a false positive, both of which are misclassification errors.

Cost is not necessarily denominated in financial terms. In medical research, for example, the consequence of a misclassification (e.g., a false negative) may result in death, which is a consequence far worse that what might be associated with a false positive. In whatever denomination cost is specified, classification trees seek to minimize that cost. If costs of misclassification are symmetrical, then classification trees simply seek to minimize classification errors.

Classification trees allow misclassifications to be weighted based on the relative cost of one type of error compared to the other. For example, if it is

determined that a false negative is three times more consequential than a false positive, costs would be assigned in a 3:1 ratio. Assigning asymmetrical cost is often a subjective assessment, requiring substantial judgment on the part of the researcher. The choice of misclassification costs should be defensible. In the absence of defensible asymmetrical costs, the default choice should be equal weighting.

In the context of the current study, the classification outcomes are either pass or fail. The two types of errors that can arise are to (1) predict an individual to pass when they ultimately will fail, or (2) predict an individual to fail when they ultimately would have passed. The cost of misclassifying an individual as "fail" is to expend resources for intervention when none was required. In the case of misclassifying an individual as "pass," no intervention is provided, but the student ultimately fails the exam. In the latter case, the cost is the psychological consequences to the individual, reputational cost to the institution, financial cost to the individual who must pay to retake the exam, and the opportunity cost of delayed entry into the profession.

For the current study, the degree to which one type of error is more costly than the other is uncertain. Although an asymmetrical cost ratio is not constrained to the use of discrete numbers, a ratio of 1:1.25, for example, conveys a level of precision that is most likely beyond the researcher's ability to actually quantify such differences. Therefore, the current study will assign equal weights to misclassification errors. It should be noted that the costs assigned to the misclassification errors could dramatically affect the structure of the classification tree that is produced. If desired, sensitivity analysis could be performed by varying the cost weighting to evaluate the

affect on the structure of the classification tree. For the current study, sensitivity analysis was not performed.

Determining when to stop splitting.

If there is no limit on the number of splits, eventually "pure" classification will be the result and each terminal node would contain only one class of cases or objects. The algorithm would have ultimately extracted all information from the data, including random variation or "noise." The general approach in addressing this issue is to stop generating splits when subsequent splits result in little marginal improvement of the prediction. Pure classification is usually unrealistic and often contributes to producing a classification tree that has been overfitted.

Rokach and Maimon (2008) state that, typically the goal is to find an optimal decision tree that minimizes generalization error. Other goals might also be specified, such as, minimizing the number of nodes or minimizing the depth of the tree.

Smaller trees are more comprehensible. If a classification tree becomes too complicated (i.e., has too many nodes), then the normally easily interpreted graphical representation becomes useless (Rokach & Maimon, 2008). Tree complexity can be controlled using stopping criteria and *pruning* methods. Pruning produces a simpler tree, with approximately the same accuracy in predicting or classifying "new" observations. Pruning is employed after the tree has been constructed.

Splitting proceeds until some specified stopping criteria is triggered. The following are common stopping rules (Rokach & Maimon, 2008, p. 19; StatSoft, 2010):

- All instances in the training set belong to a single value of *y*.
- The maximum specified tree depth has been reached.
- The number of cases in the terminal node is less than the minimum number of cases for parent nodes.
- If the node were split, the number of cases in one or more child nodes would be less than the minimum number of cases for child nodes.
- The best splitting criterion is not greater than a certain threshold.
- All terminal nodes are pure or contain no more cases than a specified minimum fraction (percentage) of the cases.

Validating the model.

Inductive learning.

Classification models are based on *inductive learning* (also known as *training*), whereby a model is constructed by generalizing from a set of data that is presumed to be similar to future unseen data. Inductive learning can be characterized as *supervised* or *unsupervised*. Unsupervised learning proceeds without specifying the dependent variable. With supervised learning, the dependent variable is specified and the algorithm attempts to discover the relationship among the dependent and independent variables (Rokach & Maimon, 2008). In essence, the algorithm discovers the rules implicit in the existing data and "learns" how to classify data. These rules are then applied to future data. Because the current study specifies the categorical dependent variable (pass/fail), the learning process is considered supervised.

During the inductive learning process, training data is required. Typically, the entire data set is divided into a training subset, a test subset, and a validation subset. The algorithm learns the classification rules from the training subset. The test subset and validation subset simulate future data. The provisional model is run against the test subset and adjusted to minimize the classification error rate. The adjusted model is then run against the validation subset where it is adjusted again to minimize the error rate (Rokach & Maimon, 2008). To be effective, a reasonably large sample size is required. A schematic for supervised modeling is shown in Figure 3.2.

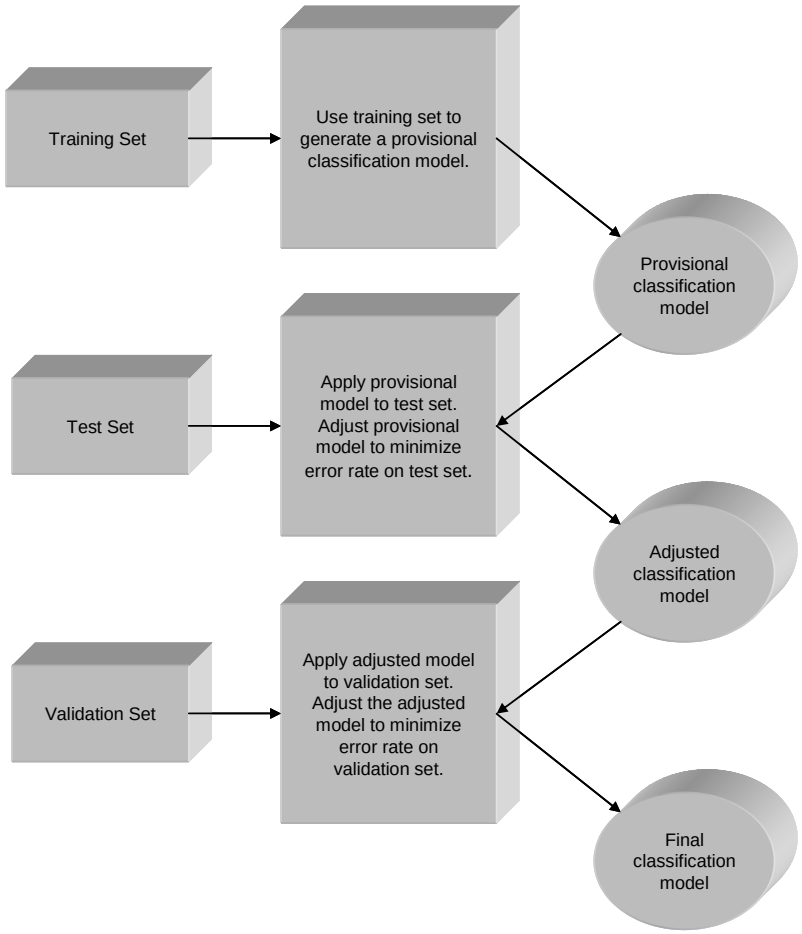

Figure 3.2. Methodology for Supervised Modeling (Rokach & Maimon, 2008, p. 92)

Model complexity, error rates, and overfitting.

Rokach and Maimon (2008) note that "typically, the accuracy of the provisional model is not as high on the test or validation sets as it is on the training set, often because the provisional model is *overfitting* on the training set" (p. 92). Overfitting occurs when the algorithm captures too much information, incorporating random "noise" and idiosyncratic features of the training set, creating a model with poor generalizability. Overfitting can be avoided by using a stopping criterion based on a statistical test of the significance of the best test, or by winnowing the structure of the decision tree after it has been produced. Most authors prefer the latter because it allows potential interactions among attributes to be explored before deciding whether the result is worth keeping (Kohavi & Quinlan, 1999).

Figure 3.3 illustrates the relationship between subset error rates and the location of a theoretically optimal model. Note that as model complexity increases, the error rate on the training set decreases monotonically. However, for the validation set, the error rate decreases to a point and then begins to increase. The optimal level of model complexity is the point where the error rate of the validation set is minimized. Complexity greater than this is considered overfitting; complexity less than this is considered underfitting (Rokach & Maimon, 2008).

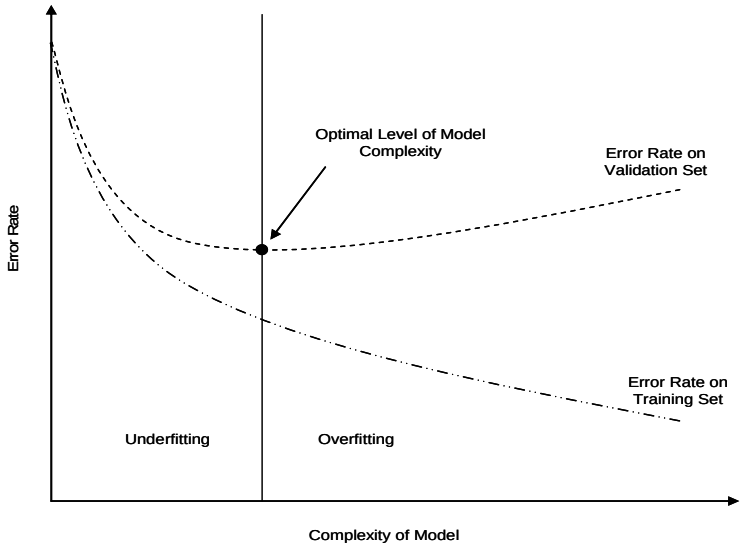

Figure 3.3. Optimal level of model complexity. Located at the minimum error rate on the validation set curve (Rokach & Maimon, 2008, p. 93)

In discussing the trade-offs between model complexity and accuracy, Rokach and Maimon (2008) note that:

> There is an eternal tension in model building between model complexity (resulting in high accuracy in the training set) and generalizability to the test and validation sets. Increasing the complexity of the model in order to increase the accuracy on the training set eventually and inevitably leads to a degradation in the generalizability of the provisional model to the test and validation sets (p. 92).

Creating validation subsets.

Test sample cross-validation.

The most preferred type of cross-validation is test sample cross-validation. The classification tree is computed from the learning sample, and its predictive accuracy is evaluated by applying it to the test sample. The test and learning samples can be formed by collecting two independent data sets. Alternatively, if a large enough learning sample is available a selected proportion (i.e., a third or half) of the cases can be set aside for use as the test sample. If the misclassification rates (or costs) for the test sample are greater than for the learning sample, poor cross-validation is indicated. A different sized tree might produce better results (StatSoft, 2010).

V-fold cross-validation.

V-fold cross-validation is appropriate when no test sample is available or the learning sample is too small to have the test sample taken from it (Breiman, Friedman, Olshen, & Stone, 1993). A specified number (V) of random subsamples (as equal in size as possible) are formed from the learning sample. The classification tree of a specified size is computed V times, each time leaving out one of the subsamples from the computations to be used as a test sample for cross-validation. Each subsample is used V-1 times in the learning sample and just once as the test sample. The cost (or error rate) is computed for each of the V test samples. The results are then averaged to give the V-fold estimate of the cost or error rate for the model (StatSoft, 2010).

Assessing model significance.

Maximum chance criterion.

A well-functioning model should be able to predict with greater accuracy than what would be expected by chance. In the absence of any better information, any particular case would be classified according to which category comprised the highest proportion in the data set (Hair et al., 2006). Known as the Maximum Chance Criterion (MCC), it is calculated as follows:

MCC = (nL / NL) * (100)

where

nL = number of subjects in the larger of the two groups

NL = total number of subjects in the combined groups

For example, if the data contained 60 males and 40 females and one case was selected at random, based on chance the prediction is that it would most likely be a male. If a large number of cases were selected, each would be predicted to be a male. Therefore, the classification success rate would be close to 60%. To be useful, the model should be able to achieve a successful classification rate of better than 60%. It could be concluded that unless a model can achieve prediction accuracy greater than the MCC, it should be regarded as useless. The question then becomes—how much better must it be before it is considered statistically significantly better.

Proportional chance criterion.

When group sizes are unequal and the goal is to correctly classify all groups, not just the largest group, the Proportional Chance Criterion (PCC) is considered by many to be the most appropriate measure (Hair et al., 2006).

The PCC is calculated as follows:

$PCC = p^2 + (1 - p)^2$

where

p = proportion of subjects in one group

(1 - p) = proportion of cases in the other group

Using the previous example of a data set containing 60 males and 40 females, the proportional chance criteria calculation would be:

$PCC = (.6)^2 + (1 - .6)^2$

$= .36 + .16$

$= .52$

In calculating PCC, the question arises as to which data set to use to determine group sizes—the overall sample or the training (or validation) subset. Hair et al. (2006) suggest that if the training sample and validation sample are sufficiently large (i.e., size > 100), derive separate standards for each sample. If the separate samples are not sufficiently large, use the group sizes from the overall sample. Hair et al. (2006) suggest that the classification accuracy should be at least 25% greater than what is expected based on chance (PCC).

Press's Q statistic.

Press's Q is a statistical measure of classification accuracy relative to chance. According to Hair et al. (2006), this test statistic compares the number of correct classifications with the total sample size and the number of groups. The calculated test statistic is compared to the chi-square critical value with 1 degree of freedom at the desired confidence level (p. 303). Hair et al. (2006) caution that the test is sensitive to sample size. Like traditional hypothesis tests, large samples are more likely to show significance than small samples. In addition, researchers should be careful in drawing conclusions based solely on this statistic. With large sample sizes, lower classification rates may still be considered significant.

Press's Q statistic is calculated as follows:

$$\text{Press's Q statistic} = \frac{[N - (nK)]^2}{N (K - 1)}$$

where

N = total sample size

n = number of observations correctly classified

K = number of groups

Data Source

The current study was based upon secondary data provided by the College of Education at XTU, and secondary data used in a similar study by Gruver (2008). For both data sets, the target population was first-time test-takers of the History (8-12) TExES certification exam. The XTU data spanned from 2002 to 2008, while the

Gruver (2008) data spanned from 2002 to 2007. Data were extracted from transcript data and TASP/THEA data provided by XTU College of Education. Not every candidate had scores for all three of the TASP/THEA exams. Because classification trees algorithmically handle missing data, cases with missing data were not eliminated. This study will use the term TASP when referring to either the TASP or THEA exam.

Test scores on the History (8-12) TExES exam were obtained from the SBEC through the Accountability System for Educator Preparation. Results of certification exams are available to the public from the SBEC, but only in summary form—individual exam data is not accessible. Proprietary exam data for specific candidates is only accessible by an authorized institutional representative. The certification officer from XTU College of Education obtained the History (8-12) TExES exam data on behalf of the researcher.

The TExES History (8-12) Preparation Manual (2006) explains that the test framework for the History (8-12) exam is organized into three broad content areas called domains. Each domain covers one or more of the educator standards for this field. Within each domain, the content is further defined by a set of competencies. Appendix A contains the description of the History (8-12) domains and competencies.

Each competency is composed of two major parts (p. 3):

- Competency statement—broadly defines what an entry-level educator in this field in Texas public schools should know and be able to do.
- Descriptive statements—describes the knowledge and skills eligible for testing.

The History (8-12) exam is composed of the following three domains (p. 6):

- Domain I World History (approx. 37% of the test)
- Domain II U.S. History (approx. 42% of the test)
- Domain III Foundations, Skills, Research, and Instruction (approx. 21% of the test)

For Domains I and II, the following knowledge standards are assessed (p. 6):

- History
- Geography
- Economics
- Government
- Citizenship
- Culture
- Science, Technology, and Society

Within Domain III, the following teaching standards are assessed (p. 7):

- Has comprehensive knowledge of the social sciences and recognizes the value of the social sciences.
- Effectively integrates the various social science disciplines.
- Uses knowledge and skills of social studies, as defined by the Texas Essential Knowledge and Skills (TEKS), to plan and implement effective curriculum, instruction, assessment, and evaluation.

Among other data items, the History (8-12) TExES exam data includes an overall exam score and separate scores for each of the three domains tested. Because

domain scores are not useful for prediction (because they occur after the fact), and because it is the *overall* score that determines the pass/fail result, domain specific scores were not used as variables in this study. Because a candidate may have taken the exam multiple times before passing, in some cases more than one set of exam scores was available. In such situations, only the result of the first attempt was used in this study.

In addition to the overall exam score (subsequently coded as pass/fail), the following items were extracted from the History (8-12) TExES exam data and serve as predictor variables:

- Test date
- Birthdate (used to derive a candidate's age at the time exam was taken)
- Gender
- Ethnicity

Various combinations of birthdates, names, and social security numbers were used to link the candidates' exam data, transcript data, and TASP/THEA scores. After the transcript data and exam data were linked, personal information was removed so that confidentiality was preserved.

Although data from the Gruver (2008) study contained fewer predictor variables than the XTU data, all variables in the Gruver (2008) data (with the exception of general education GPA) were a subset of the XTU variables. Predictor variables in the Gruver (2008) data set include:

- general education grade point average (GPA)
- history GPA
- TASP/THEA Reading score
- age at the time of the examination
- gender

Ethnicity was not included among the variables because, as Gruver (2008) notes:

> The study originally intended to examine ethnicity as a predictor variable, but the data set showed that nonwhite participants made up less than 10% of the sample, making any potential findings based on ethnicity unreliable and un-generalizable. Despite the evolution and postmodernist subfield expansion of the discipline of history, white males continue to seek and attain history teacher certification more than other demographic groups (p. 114).

Classifying and Coding

TExES exam score (continuous variable) is the raw score a candidate received on the exam. To be designated as Pass, a minimum score of 240 was required. Although results of the TExES History (8-12) exam were available in raw score form, the score was recast into the categorical Pass/Fail form. The Pass/Fail form of the results of the TExES exam was designated the dichotomous dependent variable for this study.

All dichotomous variables (i.e., TExES result, Transfer status, PostBac/Bac status, and Gender) were coded as 0/1 to permit the use of point biserial correlation

analysis. Point biserial correlation permits dichotomous categorical variables to be correlated with continuous variables. Tabachnick and Fidell (2007) note that continuous by continuous, dichotomous by continuous, and dichotomous by dichotomous correlations can all be analyzed using the Pearson product-moment correlation if dichotomous variables are coded 0/1.

Transfer status indicates whether a student transferred from another institution. PostBac/Bac status indicates if a candidate entered the teacher education program having previously been awarded a baccalaureate degree.

Lower division GPA was calculated using courses with a 1xxx or 2xxx designation. Upper division GPA was calculated using courses with a 3xxx or 4xxx designation. The various GPAs were calculated using all courses taken by a student, using all available credit hours and quality points, whether they took the courses in residence or transferred the courses from another institution. The presumption was that if a course transferred successfully, then it was deemed equivalent to a similar course taught at XTU, and therefore entered into the various GPA calculations. For purposes of GPA calculations, upper division history courses did not include graduate level courses that may have been taken prior to taking the TExES exam. However, graduate level history courses were included in the number of upper division history courses taken. It should be noted that candidates with graduate-level history courses was a rare occurrence. The premise of this framework is that the outcome of the TExES exam is predicated on the totality of a candidate's knowledge at the time the exam is taken.

Ethnicities were classified as White, Hispanic, African-American, and Other. The Other category included those of other ethnicities or unspecified ethnicity. Because the African-American and Other categories contained too few responses to be useful, only White and Hispanic categories were used and coded 1/0 to permit point biserial correlations.

Several forms of standardized achievement tests satisfy admission requirements for entry into the teacher education program. A candidate may provide TASP, ACT, or SAT (as well as the GRE) as evidence of their academic ability. The TASP assessment includes scores for reading, math, and writing. However, because these TASP component exams can be taken separately, in some cases a candidate presenting the TASP scores may not have scores in all three areas at the time the TExES exam is taken. The ACT and SAT assessments include scores for reading and math. Some candidates may have scores from more than one type of assessment test. At the extremes, some candidates could have scores from all three assessments, while others may have no scores available at the time of the TExES exam.

Data Collection

Protection of Human Subjects

Prior to data collection, a proposal was submitted to the Institutional Review Board (IRB) for approval, requesting exemption from rules governing the use of human subjects. The proposal specified that the study would rely entirely upon secondary data provided by the XTU College of Education, the SBEC, and secondary data used in the study by Gruver (2008). Because the study involved the analysis of

secondary data, no human subjects were directly involved in the study; therefore, no informed consent was required. The IRB approved the proposal and exempt status was granted. Personal data by which individuals could be identified were removed to preserve the anonymity of the individuals involved. Gruver (2008) agreed to provide access to the data from his study upon confirmation of IRB approval.

The certification officer in the XTU College of Education was personally contacted and a verbal request was made for transcript data and TASP/THEA scores for students who took the History (8-12) TExES certification exam between January 1, 2002 and December 31, 2008. After access to the data was granted, the certification officer downloaded exam score data from the SBEC through the Accountability System for Educator Preparation and provided the researcher access to that data as well. All data were compiled into Microsoft Excel and individually identifiable data items were coded in such a way that the identities of the subjects were protected. Gruver (2008) collected data similarly, which involved three other Texas universities.

Sampling Method

A convenience sample was used and the target population was first-time test-takers of the History (8-12) TExES certification exam. The sampling frame was the population of XTU students seeking teacher certification in the History (8-12) content area from January 1, 2002 – December 31, 2008. Within this group, only those who had taken the History (8-12) TExES exam since 2002 were selected.

In the Gruver (2008) study, the researcher employed a convenience sample with the same target population, where the sampling frame was the population of

students from three Texas universities seeking teacher certification in the History (8-12) content area from 2002 – 2007. Exam score data was obtained from the SBEC in a manner similar to the current study.

Missing Data

With the XTU data, because not every candidate had scores for all the TASP/THEA exams (reading, writing, and math), some missing data was expected. In addition, data from the Gruver (2008) only provides the TASP/THEA reading score. In merging the two data sets, missing data was inevitable. However, classification tree methodology is particularly adept at algorithmically handling missing data. With the XTU data, it was anticipated that the final sample size would be close to the initial sample size.

Merging Data Sets

After combining the data from transcripts, SBEC data, and TASP/THEA exam scores, XTU data were merged with data from the Gruver (2008) study. All data were compiled into an Excel spreadsheet where the data were cleaned and subsequently exported into SPSS and AnswerTree for analysis.

Sampling Bias

Sampling involves selecting a small group from a larger group (population) and studying the sample to make inferences about the population. The ability to make valid inferences depends on how well the sample represents the population. The more representative the sample, the greater will be the *external validity*, which is the degree

to which the sample generalizes to the population. An unrepresentative sample provides no useful insight into the characteristics of the population (Vogt, 2007).

If a sample is unbiased, differences between the sample and the population they represent should only be the result of chance. If the differences are not due solely to chance, then sampling bias exists. Bias is a systematic error that can prejudice the results of a study.

Selection bias, a type of sampling bias, exists if there is a tendency to overrepresent or underrepresent a part of the population (Alreck & Settle, 2004; Hildebrand & Ott, 1998). The greater the bias, the lower is the validity. Selection bias is likely whenever researchers adopt sampling strategies based on judgment or convenience. Ideally, a true random sample is preferred because each member of the population has an equal probability of being selected, producing an unbiased estimate of a population parameter. Random sampling, also known as *equal probability selection method*, ensures that inferences made from the sample are not distorted by selection bias (Groves et al., 2004). Unfortunately, true random samples are frequently unachievable in the social sciences.

Because of the difficulty in obtaining a true random sample, convenience sampling is probably the most common sampling technique (Alreck & Settle, 2004). The use of a convenience sample severely restricts the ability to generalize to the target population, and any attempt to generalize from a convenience sample should be viewed with a great deal of skepticism (Vogt, 2007).

According to Alreck and Settle (2004), there are many potential sources of sampling bias. Among the many sources of sampling bias, the following were relevant to the current study:

- Accessibility bias—some participants are more readily selected than others, leading to overrepresentation.
- Nonresponse bias—certain respondents are more prone to refuse to participate, resulting in underrepresentation.
- Self-selection bias—certain types of respondents volunteer to participate more often than other types, resulting in overrepresentation.

Because the current study involved secondary data collected on all candidates taking the History (8-12) TExES exam, no participants were more likely to be selected than another; therefore, no group would tend to be overrepresented. Because all candidates seeking certification were required to take the exam, there was no possibility of nonresponse bias; therefore, no group would tend to be underrepresented. Although each candidate self-selected to take the exam, those not taking the exam would not be part of the target population. Therefore, self-selection bias was not an issue. Because the secondary data already existed and was not created specifically for the current study, subjective judgment on the part of the researcher was not a source of bias. The major source of bias is due to the use of a convenience sample. Because of this, one should not attempt to generalize the results beyond the context of the schools involved in the analysis.

Data Analysis

Data from XTU and the Gruver (2008) were analyzed separately. An aggregate analysis was conducted by merging data from the Gruver (2008) study with the XTU data. Both sets of data were statistically analyzed using SPSS and AnswerTree (a product of SPSS).

Data were first analyzed to provide descriptive statistics for the XTU data. Descriptive statistics for the Gruver (2008) data were summarized from the results of that study.

Correlational Analysis

Correlational analysis (Pearson product-moment correlation) was used to assess the significance of the relationships among variables. The significance of the correlations among variables was tested using an alpha level of .05. Because the study is exploratory, no a priori assumptions or hypotheses are specified in the correlational analysis.

To reduce the size and complexity of examining correlations among the dependent variable and the numerous independent variables, three sets of correlation matrices are presented. The first group of correlations examined the relationships among the TExES result and the various forms of GPA, i.e., lower division GPA, upper division GPA, total GPA, and history GPA. The second group of correlations examined the relationships among the TExES result and scores on the TASP, ACT, and SAT exams. The third group of correlations examined the relationships among

the TExES result, number of history courses taken, age, and the categorical variables composed of transfer status, PostBac/Bac status, gender, and ethnicity.

Analysis of Group Differences

To the extent that group differences among categorical variables were explored, cross-tabulations and chi-square measures were used to evaluate the statistical significance of the hypothesis of independence among the categorical variables. Odds ratios were used to recast the statistical results of cross-tabulations into a form that may provide a more intuitive understanding of the results. T-tests for independent groups were used to evaluate group differences among nominal-scaled variables and various interval/ratio-scaled variables. All tests of statistical significance were evaluated against an alpha level of .05.

Classification Trees

The third, and most important part of the study, was developing classification tree models for predicting failure on the TExES History (8-12) exam. Choosing the best overall model required evaluating the models based on the combination of model complexity, classification accuracy, and generalizability of the model. A number of different classification tree algorithms were explored to determine which algorithm produces the best fitting model, without overfitting. Overfitting occurs when a classification tree incorporates too much detail during the training data phase. Too much specificity during the training data phase means that the classification tree probably will not perform well when applied to another data set.

Handling Missing Data

Because of the amount of missing data, it is important to consider how the algorithms handle missing data. In CHAID models, cases with missing data can be merged with another node, or a separate "missing data" node can be created for cases with missing data. In C&RT and QUEST models, missing data is handled using surrogate variables. Surrogate variables are independent variables with similar prediction qualities as the variable that has missing data. If data is missing, the variable with the strongest association in terms of prediction ability to the variable with the missing data is used as a substitute, enabling the classification process to continue.

Model Construction

Regardless of the manner in which missing data is handled, the goal is to seek the model that produces the best prediction; that is, the model with the least amount of misclassifications. The process of building classification trees involves:

1. developing a modeling strategy;
2. choosing the most appropriate classification tree algorithm;
3. setting the stopping criteria;
4. validating the model;
5. pruning the tree; and
6. evaluating model performance.

Developing a modeling strategy. For a given set of variables, the classification tree algorithm may generate a tree with multiple levels of branching.

For example, the algorithm may choose to make the first split based on the TASP reading score, and the second split based on history GPA. Using the TASP score may generate an effective predictive model in a theoretical sense, but in a practical sense, it may be less than desirable. It is possible that a student will not have a TASP reading score. Although surrogates may be used, they may be less effective than the original variable. Therefore, it is conceivable that the best fitting model may not be the best model for use in practice. There is an inherent tension in the trade-offs required in choosing the best fitting model or the model best suited for use in practice. Because of these kinds of situations, five different types of models were constructed for this study.

Because not all students had scores for the various TASP, ACT, and SAT exams, the first model was constructed using all variables except TASP, ACT, and SAT scores. This was considered the "base" model, containing variables common to all models. This model was referred to as the "No Scores" model. The next model, referred to as the TASP model, used all the base model variables plus the addition of TASP scores. The same procedure was done for ACT and SAT scores. Those models are referred to as the ACT model and the SAT model respectively. The last of the five models used all variables, including all scores for the TASP, ACT, and SAT exams. This model is referred to as the "All Scores" model. Using the All Scores model means there was a considerable amount of missing data, especially among the TASP, ACT, and SAT scores, although the use of surrogates mitigates the problem.

Choosing an algorithm. In seeking the best fitting model, care must taken not to "over fit" the model so that it does not generalize well to another set of data. There

is a constant tension in balancing the competing criteria of model complexity, prediction accuracy, and model generalizability. A large number of models were constructed using each of the algorithms to assess the fit and generalizability of the models. In general, C&RT models dominated QUEST and CHAID models. Therefore, the model building process concentrated on using C&RT algorithms.

Stopping criteria. In building a classification tree, stopping criteria are needed to prevent a tree from becoming too large and unwieldy. Stopping criteria can be specified by the number of levels of the tree (the depth), the specified minimum number of cases in a node has been reached (the size of the node), or the amount of improvement (impurity) in the model is less than a specified amount. The first stopping criterion to be reached terminates the tree-growing process. In general, the AnswerTree software defaults were used. Deviations from default stopping rules were: (a) minimum number of cases for parent node = 10, (b) minimum number of cases for child node = 5, and (c) cross-validation using four v-folds.

Validating the model. Decision rules used in building classification trees are "learned" from a set of data, and then those decision rules are tested against a different set of data to determine how closely the model fits the new data. This is a measure of the ability of the model to generalize beyond the data used to construct the model. If the data set is large enough, the data can be partitioned into a training set and a test set, perhaps splitting the data into two equal parts based on random selection of cases. Other splitting proportions could also be chosen.

If the data set is small, partitioning may create subsets of data that are too small for building an effective set of decision rules. In such cases, cross-validation is used to split the data set into V number of subsets (known as V-folds), with cases randomly selected for inclusion in each of the subsets. For each of the V iterations, one subset of data is excluded. The tree is grown based on all data except the subset that has been set aside. After growing the tree, the model is fitted to the excluded subset. The overall generalizability of the model is based on the average amount of misclassifications across all the subsets. For this study, the data set was considered too small for partitioning, therefore V-fold cross-classification was used.

Pruning the tree. After a tree has been fully grown and validated against the test set(s), the model may be pruned to reduce the complexity of the model. Pruning attempts to reduce model complexity (reducing number of levels and number of nodes) as much as possible without significantly increasing misclassification rates.

Pruning can be accomplished by attempting to minimize the risk of misclassification. However, seeking to minimize misclassification risk tends to "over fit" the model, thereby reducing the generalizability of the model. Overfitting can be ameliorated by using a less aggressive pruning criterion. A tree pruned using the standard error rule chooses the smallest subtree with misclassification risk close to that of the subtree with the minimum risk. Misclassification rates may increase slightly compared to the "minimum risk" criterion, but a less complex model will likely be the result. For this study, both types of pruning criteria were applied to the five different models, generating ten models.

Evaluating model performance.

Misclassification risks. Classification trees are designed to predict an outcome. In general, the better the prediction ability, the better is the model. Although, as discussed previously, the best fitting model may be overly complex, may not generalize well to other data, or it may not be well suited for use in practice. Nonetheless, the ability to correctly classify cases is the best measure for assessing the performance of a model.

Misclassification risk is assessed using a misclassification matrix, also known as a *confusion matrix*. A misclassification matrix is a 2 X 2 table that compares predicted outcomes against actual outcomes. Table 3.3 shows an example of a misclassification matrix. In the row for Prediction Category, when predicting Pass, the model correctly predicted Pass 93 times out of 107 cases. The model misclassified 14 cases. Likewise, when predicting Fail, the model correctly predicted 55 cases, but misclassified 19 cases. Correct classifications lie along the left, downward sloping diagonal. Misclassifications lie along the opposite diagonal. The percent correctly classified is calculated as (93 + 55) / 181 = .8177, or 81.77%. Misclassification must then be one minus the correctly classification percentage (1 - .8177 = .1823). This number is shown as the Risk Estimate in Table 3.3. Because it is an estimate based on a sample, the standard error of the risk estimate is also provided. The effectiveness of a model can be assessed by examining the risk estimate. Models can be compared to each other by comparing the risk estimates of the models. The smaller the risk, the better is the prediction ability.

Table 3.3

Evaluating Risk Using a Misclassification Matrix

Misclassification Matrix				
		Actual Category		
		Pass	Fail	Total
Predicted Category	Pass	93	14	107
	Fail	19	55	74
	Total	112	69	181

	Risk Statistics	Cross-Validation
Risk Estimate	0.18232	Not Available
SE of Risk Estimate	0.0286992	Not Available

Assessing statistical significance. A well-functioning model should be able to predict with greater accuracy than what would be expected by chance. In assessing the appropriateness of the model, the classification rate of the model was evaluated against the proportional chance criterion and the Press's Q statistic. In using the proportional chance criterion, Hair et al. (2006) suggest that the classification accuracy should be at least 25% greater than the prediction based on chance. In using Press' Q statistic, the calculated test statistic was compared to the chi-square critical value with 1 degree of freedom at an alpha level of .05.

Choosing Among Competing Models

Ten classification tree models were developed. No single model was clearly dominant in all situations. A model superior in one aspect, was often dominated by another model on a different aspect. This makes selecting the best model a matter of assessing the performance measures for the model, and assigning weights to those

measures based on the perceived importance. Performance measures used to evaluate the quality of the model were:

- model complexity (number of levels, total nodes, and total terminal nodes)
- model improvement (impurity) at the first and second level splits
- percentage of Pass predictions correctly classified
- percentage of Fail predictions correctly classified
- total percentage of all predictions correctly classified

Because of the number of models under consideration, an evaluation matrix was created for analyzing the numerous trade-offs involved in selecting the best model. Models were grouped based on the pruning method employed and analyzed as a group.

For models pruned using the standard error criterion, one standard error (the default) was chosen as the pruning parameter, although other values could have been chosen. Choosing larger standard error values makes it less likely that a split will occur because the difference threshold is greater, resulting in trees of less complexity. However, misclassification risks will likely increase as complexity decreases. Conversely, a smaller standard error value makes it more likely that a split will occur because the difference threshold is less, resulting in more complexity—which will likely result in smaller misclassification risks. Highlighted cells in each matrix designate the best outcome(s) for an item on a particular row.

Reliability, Validity, Generalizability

The validity and reliability of the data were considered adequate because the instrumentation of the TExES exam and TASP/THEA exams were well vetted before adoption by the State of Texas. Because of the use of convenience sampling and the lack of consistency in the requirements established by the various teacher education programs, external validity should not be presumed tenable. Therefore, the results of the study are not generalizable beyond the immediate context of the study.

Chapter Summary

This chapter presented the rationale and design of the current study. Methodological limitations of previous studies were identified and generally concerned the questionable choice of using stepwise multiple regression, considered by some statistical experts to be an inappropriate methodology for the type of research being conducted in those studies (e.g., Berk, 2004; Kutner, Nachtsheim, Neter, & Li, 2005; McClave, Benson, & Sincich, 2008; Warner, 2008).

The consequences of large standard errors produced by the regression models were addressed, as well as the questionable use of confidence intervals for making predictions. Arguments were present for why the use of prediction intervals may be more appropriate than using confidence intervals, although the use of prediction intervals made the task of prediction even more ambiguous and difficult. Arguments were also presented for why a categorical variable is probably more appropriate than a continuous variable for use as the criterion variable in the context of the current study.

Because of the distributional requirements of parametric methods and the consequences of violations of model assumptions, the use of nonparametric methods was discussed. Classification trees (a nonparametric method) were presented as an alternative methodology to the use of stepwise multiple regression and logistic regression. The advantages and limitations of classification trees were discussed, and arguments were advanced for why the use of classification trees might be superior to stepwise multiple regression in certain situations.

Lastly, data source, data collection procedures, sampling method, and the method of data analysis were articulated. The problems of reliability, validity, and generalizability of the results were also discussed.

CHAPTER IV

RESULTS

This chapter discusses the results of the analyses of (a) variable attributes and distributions using descriptive statistics, (b) direction and strength of relationships among variables using correlation analysis, (c) group differences among categorical variables, and (d) classification tree models for predicting failure on the TExES History (8-12) certification exam.

Multiple models were evaluated in an attempt to balance the complicated interrelationships of model complexity, prediction accuracy, and generalizability. No single model was clearly superior in all respects. Generally, models with the smallest risk of misclassification were models of moderate complexity. Models with the least complexity had the highest risk of misclassification, suggesting that a certain amount of complexity is needed to improve prediction accuracy. However, models with the most complexity tend to overfit the available data, and therefore may not generalize well to new data.

Descriptive Statistics

XTU Data

General descriptive statistics. The XTU data (n = 181) were collected using a convenience sample and consisted of approximately 65% males and 35% females. In terms of ethnicity, 81.2% were White, 14.4% were Hispanic, and African-American/Other accounted for the remaining 4.4%. White males were the largest

segment (n = 96), constituting 53% of the total sample. The small proportion of African-American/Other made it difficult to produce meaningful results for these two ethnic categories. Therefore, analyses involving ethnicity used only the two largest ethnic categories, White and Hispanic, which represented 95.6% of the total sample.

The majority of the candidates (91.7%) were categorized as Baccalaureate, while 8.3% were categorized as Post Baccalaureate. Transfer students made up 75.7% of the sample. Overall, from 2002 through 2008, 61.9% passed the TExES exam, while 38.1% failed the exam.

TASP Reading and Math scores were available for almost 56% of the candidates. Only 42.5% of the candidates had all three TASP scores available, although the TASP Writing score was inconsequential for any of the models that were developed. For the ACT, 52.5% of the candidates had those scores available, while 59.1% of the candidates had SAT scores available. Only 11.6% of the candidates had all three sets of assessment scores available, while 11% of the candidates had none of the assessment scores available.

Gruver (2008) Data

Data from the Gruver (2008) study (GData) were made available for this study. A convenience sample (n = 119), grouped only by gender, was composed of 60% males and 40% females. Due to missing data, the regression model developed in the Gruver study was based on a reduced sample size, n = 81.

The study by Gruver (2008), which involved three regional state universities, resulted in a multiple linear regression model with TExES exam score as the

continuous dependent variable. Independent variables included gender, age at time of test, TASP reading score, General Education GPA, and History GPA. The resulting multiple regression model was statistically significant, $F(5, 75) = 6.300$, $p < .001$. A large effect size was observed ($R^2 = .296$). The TASP/THEA Reading score was the only statistically significant predictor in the model, accounting for approximately 26% of the unique variance.

The current study calculated History GPA differently than Gruver, who calculated History GPA considering only the history courses taken in residence at a particular institution. The current study calculated History GPA using all available history course information, whether courses were transferred or taken in residence. Institutions accept transfer courses because of the presumed equivalence between transferred courses and in-residence courses; therefore, it seemed reasonable to include those courses in GPA calculations for the current study.

Two of the three institutions in GData had pass rates over 88%, indicating the outcomes of the TExES exam were highly homogenous. This high degree of homogeneity makes it difficult to improve pass/fail predictions. Random chance would suggest that, on average, predicting a candidate to pass would be correct 88% of the time. Because of these differences and the possibility that the data could skew the results from the XTU data, the data from the Gruver study was not included in the final analysis.

Although data from the Gruver study was excluded from the final analysis, it should be noted that a classification tree was constructed using only GData. When

pruned for minimum risk, the classification tree had only one branch, splitting on the TASP Reading score. No other variables entered the model. This result is consistent with the multiple regression model developed by Gruver. However, when pruned using the standard error criterion, the tree had no branches; the resulting model could not improve prediction over random chance classification.

Correlation Analysis

All forms of GPA were moderately correlated with TExES result (pass/fail), $p < .001$. Of the four GPAs, History GPA was the most highly correlated with TExES result, $r = +.407$. History GPA was highly correlated with total GPA, $r = +.835$, $p < .001$.

Although the TASP Reading score was not significantly correlated with TExES result (pass/fail), $r = +.183$, $p = .101$, it was moderately correlated with the TExES exam raw score, $r = +.374$, $p < .01$. ACT Reading ($r = +.612$), ACT Math ($r = +.366$), SAT Reading ($r = +.458$) and SAT Math ($r = +.423$) were all significantly correlated with the TExES result, $p < .001$.

The ACT Reading score and SAT Reading score were highly correlated with each other, $r = +.834$, $p < .001$. The ACT Reading and SAT Reading score were moderately correlated with TASP Reading score, $r = +.464$, and $r = +.511$, respectively, $p < .001$. The TASP writing score was not significantly correlated with any other variable, $r = +.034$.

The ACT and SAT reading scores were more highly correlated with each other than with the TASP reading score, suggesting that the TASP exam was measuring the

reading construct in a different manner than either the ACT or SAT, although the nature of this difference was not discernible from the data.

Although the strength of correlation was only moderate, transfer status, PostBac/Bac status, and ethnicity were significantly correlated with TExES result, $p < .01$. Total number of history courses, number of upper division history courses, gender, and age were not significantly correlated with the TExES result. Of the categorical variables, transfer status was the most highly correlated with TExES result, $r = -.286$, $p < .001$. The inverse relationship between the two variables indicates that non-transfer students were more associated with passing the exam than were transfer students.

Classification Trees

The model building process was thoroughly discussed in Chapter III, but is briefly reiterated here. AnswerTree 3.1 (a product of SPSS) was used to develop the classification trees. Unless otherwise noted (see discussion in Chapter III), in choosing model-building parameters, the software defaults were used.

The following variables were used for developing classification tree models:

- TExES result (pass/fail)*
- Transfer status
- PostBac/Bac status
- Gender
- Ethnicity
- Age at time of exam
- Lower division GPA
- Upper division GPA
- Total GPA
- History GPA
- Total number of history courses
- Number of upper division history courses
- TASP scores (Read/Write/Math)
- ACT scores (Read/Math)
- SAT scores (Read/Math)

* dependent variable

When using all variables, the resulting classification tree may involve a particular assessment score, i.e., TASP, ACT, or SAT. In practice, such a model may be problematic because a candidate may not have a score for the assessment upon which the model is based. Because of this possibility, five different types of models were developed for this study.

The first model (the No Scores model) used all variables except TASP, ACT, and SAT scores and was considered the "base" model, containing variables common to all models. The second model (the TASP model) used all the base model variables plus the addition of TASP scores. The same procedure was used for ACT and SAT scores. Those models are referred to as the ACT model and the SAT model respectively. The last of the five models (the All Scores model) used all variables, including all scores for the TASP, ACT, and SAT exams.

After a tree has been fully grown, pruning attempts to reduce model complexity as much as possible without significantly increasing misclassification rates. Two different pruning algorithms were used—minimum risk pruning and standard error pruning. The minimum risk pruning algorithm minimizes the risk of misclassification but tends to "over fit" the model, diminishing the generalizability of the model. The standard error algorithm produces the smallest subtree with misclassification risk close to that of the subtree with the minimum risk. Misclassification rates may increase slightly compared to the minimum risk criterion, but a less complex model and increased generalizability will likely result. For this

study, both types of pruning criteria were applied to the five different models, resulting in ten different models.

No single model was clearly superior across all criteria. A model that was superior under one criterion was often inferior under a different criterion. Selecting the best model was a matter of assessing the performance measures for the model, and assigning weights to those measures based on the perceived importance. Performance measures used to evaluate the quality of the model were:

- model complexity (number of levels, total nodes, and total terminal nodes)
- model improvement (impurity) at the first and second level splits
- percentage of Pass predictions correctly classified
- percentage of Fail predictions correctly classified
- total percentage of all predictions correctly classified

Because of the number of models under consideration and the complexity of evaluating the trade-offs involved in selecting the best model, an evaluation matrix was created for analyzing the models. The Evaluation Matrix shown in Table 4.1 evaluates five different models using the minimum risk pruning criteria. The Evaluation Matrix shown in Table 4.2 evaluates the same five models using the standard error pruning criterion. Highlighted cells in each matrix indicate the best outcome(s) for an item on a particular row.

The default (one standard error) was chosen as the pruning parameter when using the standard error criterion. Choosing larger standard error values results in trees of less complexity. However, misclassification risks would likely increase as

complexity decreases. Conversely, a smaller standard error value results in more complexity—which will likely result in smaller misclassification risks.

The matrix is divided into four sections:

- model complexity;
- the nature of the first and second level splits and the model improvement associated with those splits;
- classification performance; and
- overall model performance based on subjective weights for the four components.

In the top section of the matrix, models were evaluated based on complexity. Complexity is a function of the number of levels, the number of nodes, and the number of terminal nodes. A terminal node is one that cannot be split because a stopping rule has been reached.

There are many ways to define complexity. For this study, complexity was defined as a function of the sum of the number of levels, the number of nodes, and the number of terminal nodes. However, it is generally understood that complexity is not a linear function. As the number of components in a system increases, the permutations of possible interactions increase dramatically.

Table 4.1

Evaluation Matrix for Five Models Using Minimum Risk Pruning Criterion

	Model Summaries and Evaluation Matrix (Prune: Minimum Risk)				
	No Scores[a]	TASP	ACT	SAT	All Scores
Tree Size					
Nodes	15	11	9	11	9
Levels	4	4	3	3	4
Terminal nodes	8	6	5	6	5
Model Complexity					
Complexity Penalty	**1.25**	**1.25**	**1.25**	**1.25**	**1.25**
Relative Complexity	61.5	45.0	34.5	42.3	37.1
First Level Split	Hist GPA	Hist GPA	ACT Read	SAT Read	ACT Read
Improve	0.1060	0.1060	0.1808	0.1111	0.1808
Second Level Split(s)					
Split 2.1	Transfer Flag	TASP Read	Transfer Flag	UpDiv GPA	Transfer Flag
Improve	0.0182	0.0346	0.0606	0.0288	0.0619
Split 2.2	Tot Hist Crs	na	na	UpDiv GPA	na
Improve	0.0078	na	na	0.0228	na
Pass					
Correct	93	97	92	95	97
Predicted	107	115	107	109	116
Pct Correct	86.9%	84.3%	86.0%	87.2%	83.6%
Fail					
Correct	55	51	54	55	50
Predicted	74	66	74	72	65
Pct Correct	74.3%	77.3%	73.0%	76.4%	76.9%
Total Cases	181	181	181	181	181
Total Correct	148	148	146	150	147
Total Pct Correct	81.8%	81.8%	80.7%	82.9%	81.2%
Std Error	2.9%	2.9%	2.9%	2.8%	2.9%
Misclassify Risk	18.2%	18.2%	19.3%	17.1%	18.8%
Model Evaluation	Weight				
Complexity	85				
Pass% Correct	85				
Fail% Correct	100				
Total% Correct	90				
Model Rating	249.4	260.4	267.9	265.3	266.9

Note. All models include the common set of variables: Transfer Flag, Gender, Ethnicity, Age at Test, Lower Division GPA, Upper Division GPA, History GPA, Total GPA, Total number of History Courses, Number of Upper Division History Courses. Shaded cells represent the best outcome for any particular row.
[a]The "No Scores" model includes only the common set of variables. Subsequent models include exam scores for TASP, ACT, SAT, and All Scores respectively

.

Table 4.2

Evaluation Matrix for Five Models Using Standard Error Pruning Criterion

	Model Summaries and Evaluation Matrix (Prune: One Standard Error)				
	No Scores[a]	TASP	ACT	SAT	All Scores
Tree Size					
Nodes	9	7	7	7	7
Levels	4	3	3	3	3
Terminal nodes	5	4	4	4	4
Model Complexity					
Complexity Penalty	**1.25**	**1.25**	**1.25**	**1.25**	**1.25**
Relative Complexity	37.1	27.1	27.1	27.1	27.1
First Level Split	Hist GPA	Hist GPA	ACT Read	SAT Read	ACT Read
Improve	0.1060	0.1060	0.1808	0.1111	0.1808
Second Level Split(s)					
Split 2.1	Transfer Flag	TASP Read	Transfer Flag	UpDiv GPA	Transfer Flag
Improve	0.0182	0.0346	0.0606	0.0288	0.0619
Split 2.2	na	na	na	na	na
Improve	na	na	na	na	na
Pass					
Correct	97	90	95	99	101
Predicted	118	106	114	121	126
Pct Correct	82.2%	84.9%	83.3%	81.8%	80.2%
Fail					
Correct	48	53	50	47	44
Predicted	63	75	67	60	55
Pct Correct	76.2%	70.7%	74.6%	78.3%	80.0%
Total Cases	181	181	181	181	181
Total Correct	145	143	145	146	145
Total Pct Correct	80.1%	79.0%	80.1%	80.7%	80.1%
Std Error	3.0%	3.0%	3.0%	2.9%	3.0%
Misclassify Risk	19.9%	21.0%	19.9%	19.3%	19.9%
Model Evaluation	Weight				
Complexity	85				
Pass% Correct	85				
Fail% Correct	100				
Total% Correct	90				
Model Rating	264.0	276.7	280.3	283.3	283.0

Note. All models include the common set of variables: Transfer Flag, Gender, Ethnicity, Age at Test, Lower Division GPA, Upper Division GPA, History GPA, Total GPA, Total number of History Courses, Number of Upper Division History Courses. Shaded cells represent the best outcome for any particular row.
[a]The "No Scores" model includes only the common set of variables. Subsequent models include exam scores for TASP, ACT, SAT, and All Scores respectively.

Complexity was characterized as an exponential function and the nature of this relationship was captured by first summing the levels, nodes, and terminal nodes, and then raising that quantity to the power of a subjectively chosen value. The resulting value was scaled so that the range of values was between zero and one, similar to the ranges of the other evaluation factors. Complexity penalty values were constrained to be between 1.00 and 1.50. For this study, the chosen value of the complexity penalty was 1.25. The more complex the model, the more it was penalized. Symbolically, the relative complexity of the model is calculated as: $(a + b + c)^{1.25}$.

For models pruned using the minimum risk criterion, a sensitivity analysis revealed that the overall best model depended on the value of the complexity penalty. With a complexity penalty of 1.25, the ACT model was the best overall model using the minimum risk criterion. For models pruned using the standard error criterion, a sensitivity analysis revealed that the choice of best model did not change for any complexity penalty between 1.00 and 1.50. For models pruned using this criterion, the SAT model was the best overall model.

In the section for evaluating splits, the splitting variable and a measure of improvement based on that split are identified. The improvement measure was taken from the classification tree diagram generated by AnswerTree. The larger the measure of improvement, the more useful the splitting variable was for prediction. Although models often had more than two levels, in general, splits more than two levels deep yielded little incremental improvement.

In the section for evaluating the accuracy of classification, values were taken directly from the misclassification matrix generated by AnswerTree. Total percent correct and misclassification risk are two sides of the same coin. Misclassify risk is simply one minus total percent correct.

The model with the best accuracy for predicting Pass was never the best model for predicting Fail. This created a tension in choosing the best model. For this study, the focus was on identifying candidates who were likely to fail the TExES exam. Therefore, in general, models better at predicting Fail were preferred to models that were better at predicting Pass, if the accuracy of predicting Pass (and total prediction accuracy) was not significantly degraded. This preference for accuracy of predicting Fail is reflected in the weights assigned in the Model Evaluation section near the bottom of the matrix. Among the four factors, the ability to predict Fail was assigned the greatest weight.

In the Model Evaluation section, each of the four components of model performance was subjectively assigned a weight based on the perceived importance of that factor in determining the best fitting model. A weight of 100 was assigned to the most important factor, which is the ability to predict Fail. Other factors were weighted relative to 100. Because the researcher was unwilling to accept overly complex models to achieve prediction accuracy, complexity had a weighting of 85.

A larger complexity value indicates more complexity, and therefore, a less desirable model. For the other three factors, a larger number indicates a more desirable model. So that the scales of the four factors are consistent (i.e., larger

indicates a more desirable model), the reciprocal of the complexity value was used in calculating the final model rating.

No model was clearly superior to other model in all respects. The best model for predicting Pass was not necessarily the best model for predicting Fail.

Summary of Model Results

Models pruned with minimum risk rule. For models pruned using the minimum risk criterion, overall model ratings ranged from 249.4 to 267.9, with the ACT model rated highest. The No Scores model ranked lowest. Complexity ranged from 61.5 (No Scores model) to 34.5 (ACT model). Pass prediction rates ranged from 83.6% to 87.2%, and Fail prediction rates ranged from 73.0% to 77.3%. The SAT model correctly predicted Pass 87.2% of the time (the highest of the five models), and correctly predicted Fail 76.4% of the time. The TASP model was the best at predicting Fail, 77.3%. Overall prediction rates (pass and fail combined) ranged from 80.7% (ACT model) to 82.9% (SAT model). Because the SAT model had the highest success rate for overall prediction accuracy, it had the lowest misclassification risk, 17.1%. The most complex model (No Scores model) had a misclassification risk of 18.2%. Despite having the highest misclassification risk (19.3%) and lowest Fail prediction rate (73.0%), the ACT model was less complex and had the best overall model score. Figure 4.1 shows the classification tree for the model with the highest rating—ACT model.

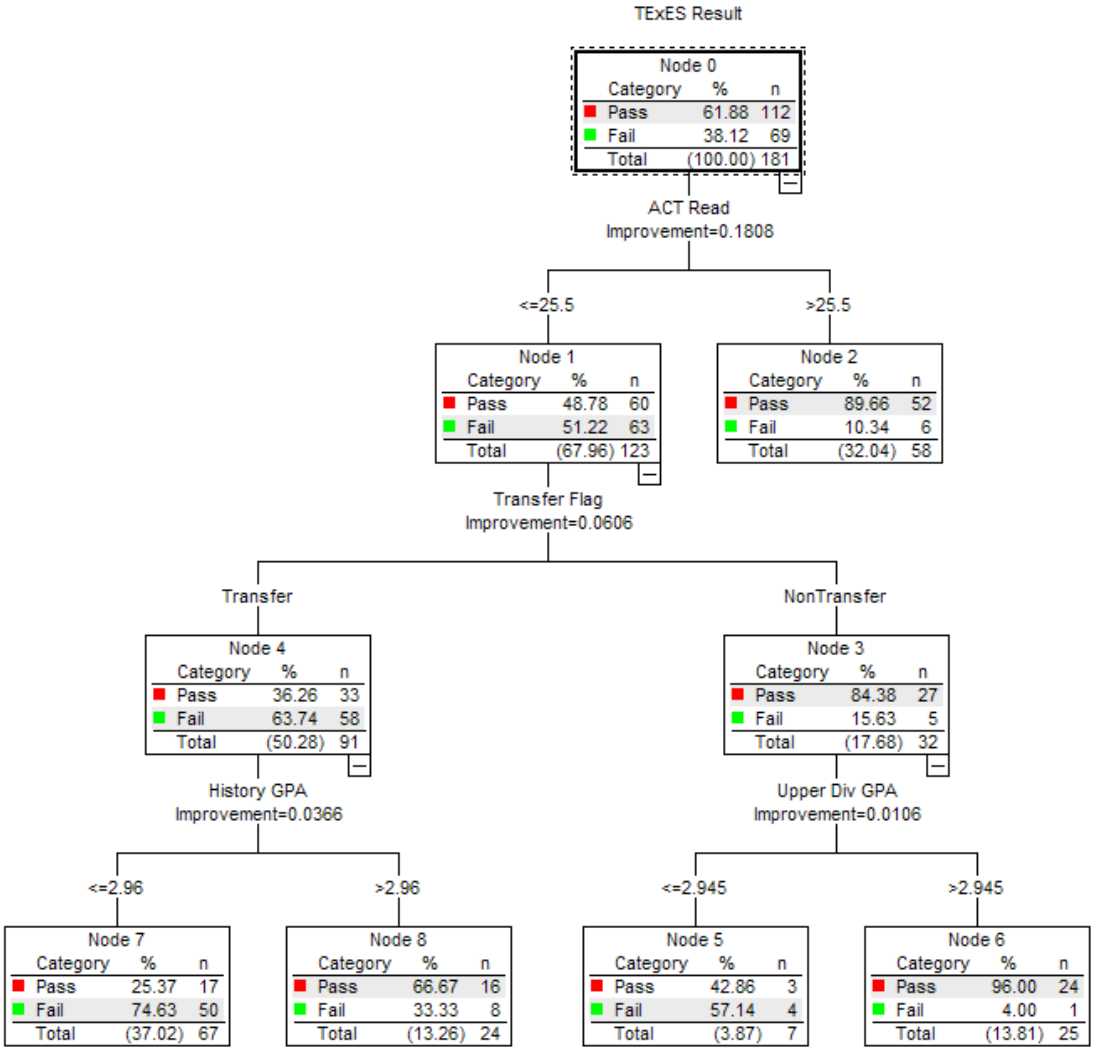

Figure 4.1. Best model (minimum risk pruned)—ACT model**.** Misclassification risk = 19.3%.

For the ACT model, four variables were required to form the tree—ACT Reading score, transfer status, history GPA, and upper division GPA. The first split used ACT Read. If the ACT score is less than or equal to 25.5 (Node 1), the model predicts Fail. For ACT scores greater than 25.5 (Node 2), the model predicts Pass, with a passing rate of almost 90%. In Node 1, the general prediction is Fail, but was correct only 51.2% of the time. However, this is better than the 38.1% prediction rate

for Fail shown in the topmost node (Node 0), which are the results for the overall sample. That is, in the total sample, 69 of 181 (38.1%) cases failed the TExES exam. Progressing deeper into the tree, prediction rates become better because the model makes increasing refinements in its ability to segregate those who passed and those who failed, i.e., the model next used transfer status to further segregate the candidates. Adding additional levels to the model provided diminishing incremental improvements in predictive ability.

Models pruned with the standard error rule. For models pruned using the standard error criterion, overall model ratings ranged from 264.0 (No Scores) to 283.3 (SAT). Complexity ranged from 37.1 (No Scores model) to 27.1 (all other models). The SAT model correctly predicted Pass 81.8% of the time, and correctly predicted Fail 78.3% of the time. The TASP model was the best at predicting Pass (84.9%), while the All Scores model had the lowest Pass prediction rate, 80.2%. The All Scores model was the best at predicting Fail (80.0%), while the TASP model had the lowest Fail prediction rate, 70.7%. The TASP model was the best at predicting Pass and the worst at predicting Fail, while the All Scores model was the best at predicting Fail and the worst at predicting Pass. The All Scores model was very consistent, predicting Pass and Fail equally well, 80.2% and 80.0% respectively. Overall prediction rates ranged from 79.0% (TASP model) to 80.7% (SAT model). The results show that there was more variability in the rates for predicting Pass/Fail than in the overall prediction rates. Because the SAT model had the highest success rate for overall prediction accuracy, it also had the lowest misclassification risk, 19.3%. The

TASP model had the highest misclassification risk, 21.0%. The most complex model (No Scores model) had a misclassification risk of 19.9%. Figure 4.2 shows the classification tree for the model with the highest rating—SAT model.

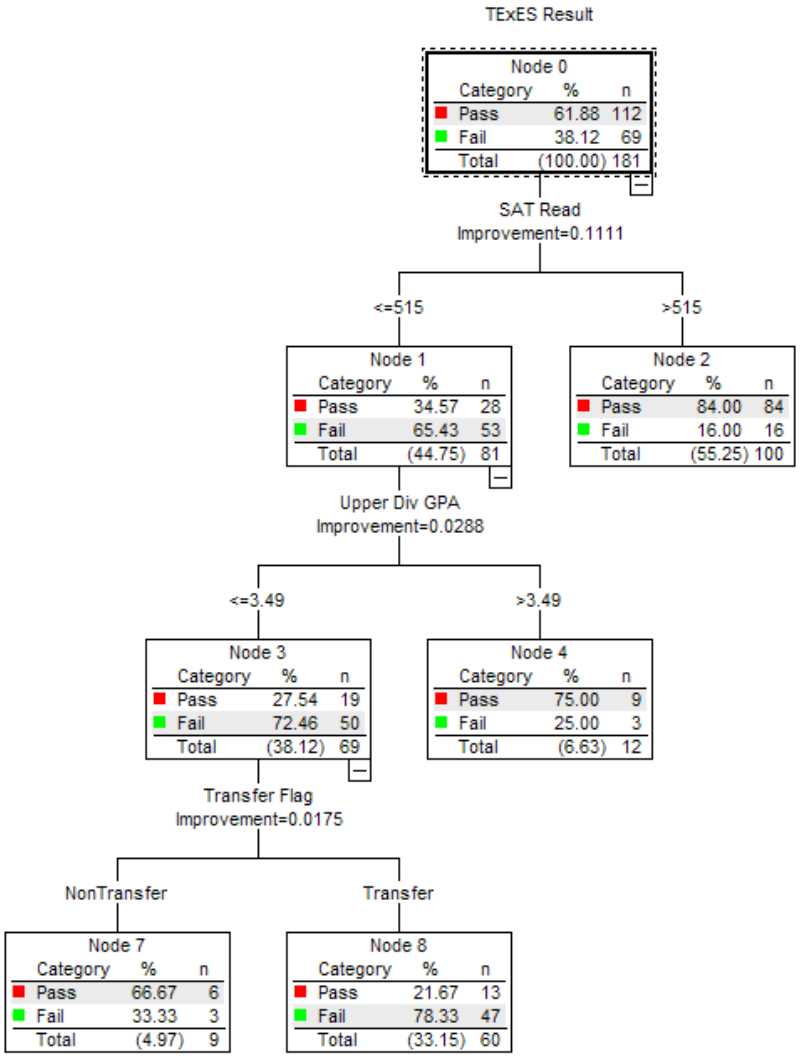

Figure 4.2. Best model (standard error pruned)—SAT model**.** Misclassification risk = 19.3%.

For the SAT model, only three variables were required to form the tree—SAT Reading score, upper division GPA, and transfer status. If the SAT Reading score is

less than or equal to 515 (Node 1), the model predicts Fail, while if the SAT Reading score is greater than 515 (Node 2), the model predicts Pass, with 84% of candidates passing the exam.

In Node 1, the general prediction is Fail, but was correct only 65.4% of the time. However, this is significantly better than the 38.1% prediction rate for Fail shown in the topmost node (Node 0), which are the results for the overall sample. Progressing deeper into the tree, the prediction rates become better because the model has made increasing refinements in its ability to segregate those who passed and those who failed, i.e., the model next used upper division GPA to segregate the candidates further. Upper division GPA improved model impurity by .029. Adding additional levels to the model provides diminishing incremental improvements in predictive ability.

Using the minimum risk pruning criterion, the ACT model was the best, whereas using the standard error pruning criterion, the SAT model was the best. By using the standard error pruning criterion a less complex tree was produced, resulting in two fewer total nodes and one less terminal node. The reduction in complexity was achieved with no increase in misclassification risk (19.3%) and the prediction rate for Fail increased from 73.0% to 78.3%. Overall prediction rate remained the same at 80.7%. The overall model rating for the standard error pruned tree (283.3) was 15.4 points better than the model pruned using the minimum risk criterion (267.9).

Assessing Model Significance

Assessing model performance is accomplished in several ways. The performance of the model can be evaluated using the Maximum Chance Criterion (MCC), where the maximum chance is defined as the group with the largest proportion in the sample. For this study, approximately 62% of students passed and 38% failed. Therefore, chance suggests that, in the absence of any better information, all new cases should be classified as Pass because it represents the outcome with the highest probability. According to Hair et al. (2006), the MCC should be used when the objective is to maximize the percent of correctly classified classes. It is the most conservative standard, making it harder to reject the null hypothesis.

Although less conservative than the MCC, the Proportional Chance Criterion (PCC) is considered by many to be the more appropriate benchmark when group sizes are unequal. As a measure of statistical significance, Hair et al. (2006) suggested that the classification accuracy of the model should be at least 25% greater than PCC or MCC. Because the measure of significance is a heuristic, no specific p-value is associated with the finding of significance.

Press' Q can also be used to evaluate model performance. Unlike MCC or PCC, this measure takes into account the number of groups, not just the model in general. Press' Q test statistic is compared to a critical value—the chi-square value for one degree of freedom at the specified alpha level. Rejecting the null hypothesis indicates that the model performs significantly better than chance.

Graphical methods are also available for assessing the performance characteristics of the model. Three types of charts are commonly used—gains charts, response charts, and lift (index) charts. For the five classification trees pruned using the standard error criterion, these charts and a discussion of how they are interpreted are located in Appendix B.

Models pruned using minimum risk criterion. Among the models pruned using the minimum risk criterion, the ACT model was judged the best overall model. Evaluated against the maximum chance criterion, the model was 1.30 times better than chance. Evaluated against the proportional chance criterion, the model was 1.53 times better than chance. Because performance measures exceeded 1.25 in both cases, the model was statistically significant.

When evaluated against Press' Q, the ACT model value had a Q value of 68.07, which exceeded the critical value of 10.83 ($\chi^2(1)$, $p = .001$). Because the observed value of Q was greater than the critical Q, the null hypothesis was rejected, indicating the model performed significantly better than chance.

Models pruned using standard error criterion. Among the models pruned using the standard error criterion, the SAT model was judged the best overall model. Evaluated against the maximum chance criterion, the model was 1.30 times better than chance. Evaluated against the proportional chance criterion, the model was 1.53 times better than chance. Because performance measures exceeded 1.25 in both cases, the model was statistically significant.

When evaluated against Press' Q, the SAT model value had a Q value of 68.07, which exceeded the critical value of 10.83 ($\chi^2(1)$, $p = .001$). Because the observed value of Q was greater than the critical Q, the null hypothesis was rejected, indicating the model performed significantly better than a chance model.

In what may seem like an unlikely coincidence, both models had exactly the same values used in evaluating significance. Although the two models differed in classification accuracies for Pass and Fail, the overall proportion of correct classifications was the same for both models. The ACT model correctly classified 92 cases for Pass and 54 cases for Fail; the SAT model correctly classified 99 cases for Pass and 47 cases for Fail. Both models had an overall classification rate of 80.7% (146 / 181). Because all three measures of significance use the overall proportion of correct classifications, the results were identical. Statistically, both models were highly significant.

Balancing competing goals

Selecting the best model depends on making compromises among three competing goals—model complexity, prediction accuracy, and generalizability of the model. Seeking to increase prediction accuracy usually requires a model with more complexity, which often results in overfitting the existing data, leading to poor generalizability. Alternatively, seeking less complexity usually leads to better generalizability, but with poorer prediction accuracy. Compared to models prune using the one standard error criterion, models pruned using the minimum risk criterion produced trees of greater complexity and increased risk of overfitting. Therefore,

models pruned using the minimum risk criterion were eliminated from final consideration.

The best model (SAT) was shown in Figure 4.2. For convenience and ease of comparison, the other four classification trees pruned using the standard error criterion are shown below in Figures 4.3 through 4.6.

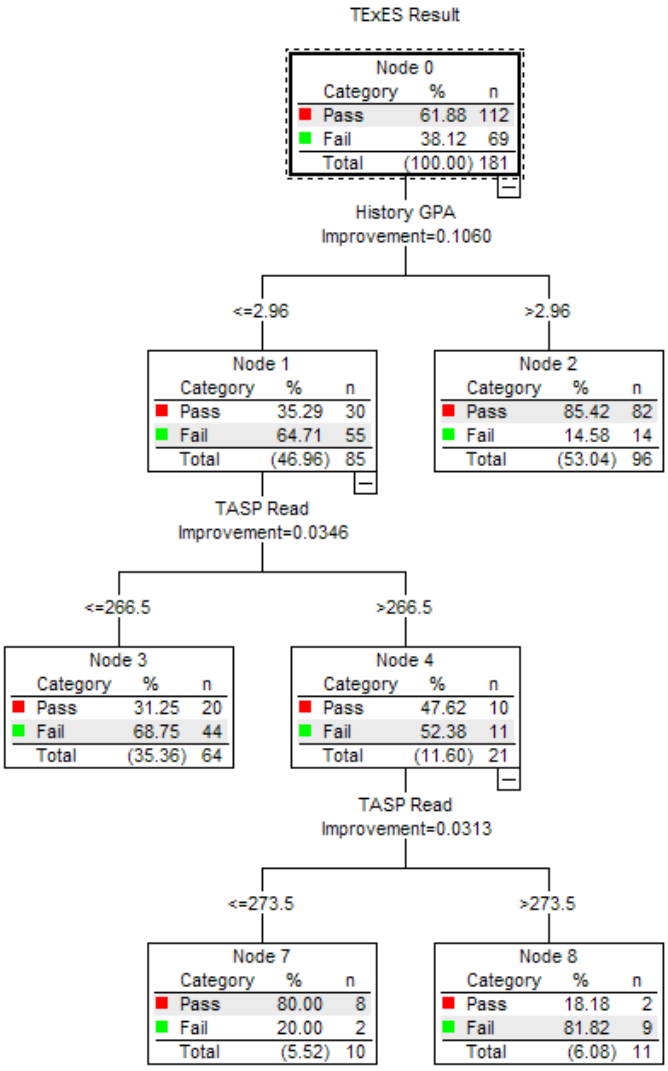

Figure 4.3. TASP model (standard error pruned). Misclassification risk = 21.0%.

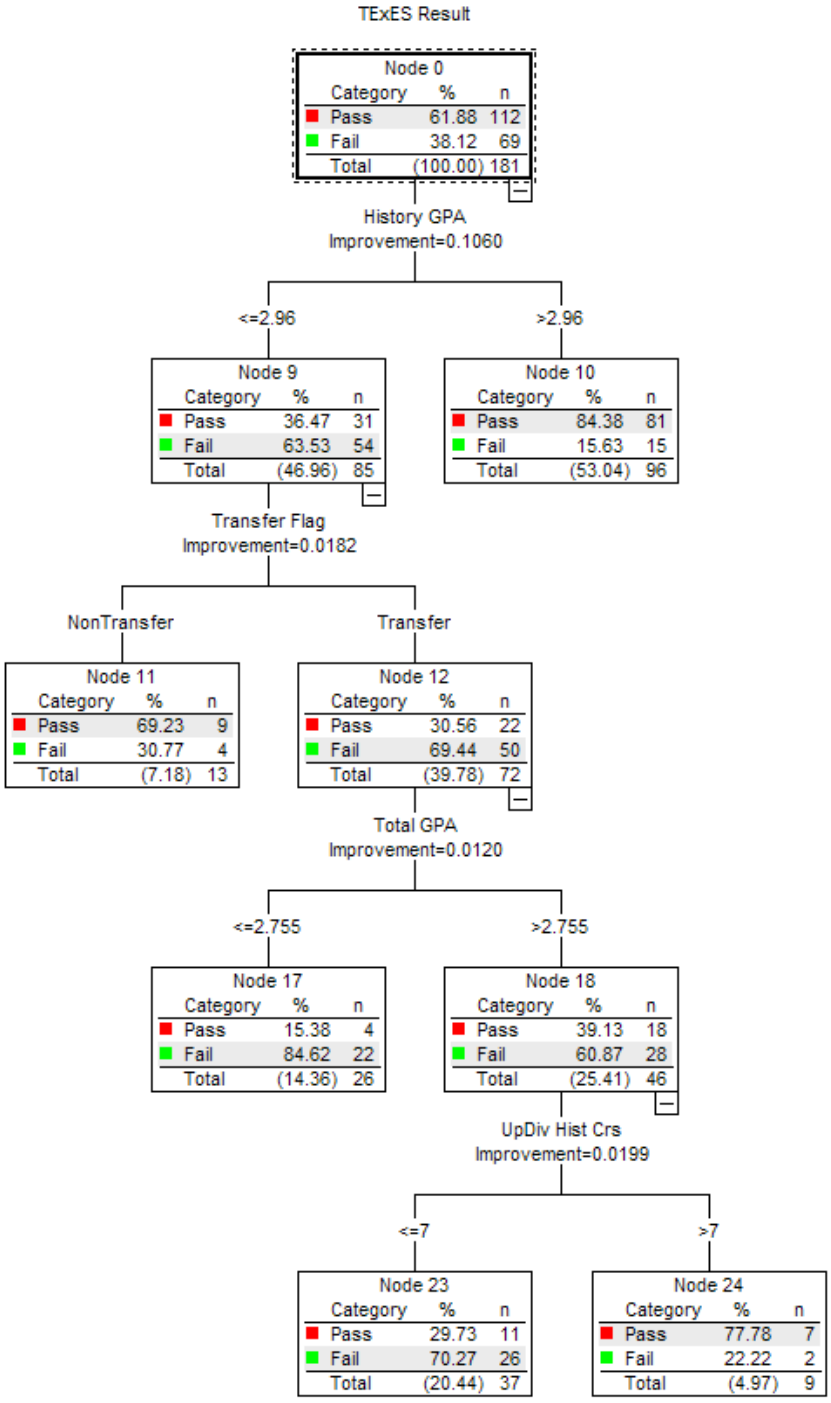

Figure 4.4. No Scores model (standard error pruned). Misclassification risk = 19.9%.

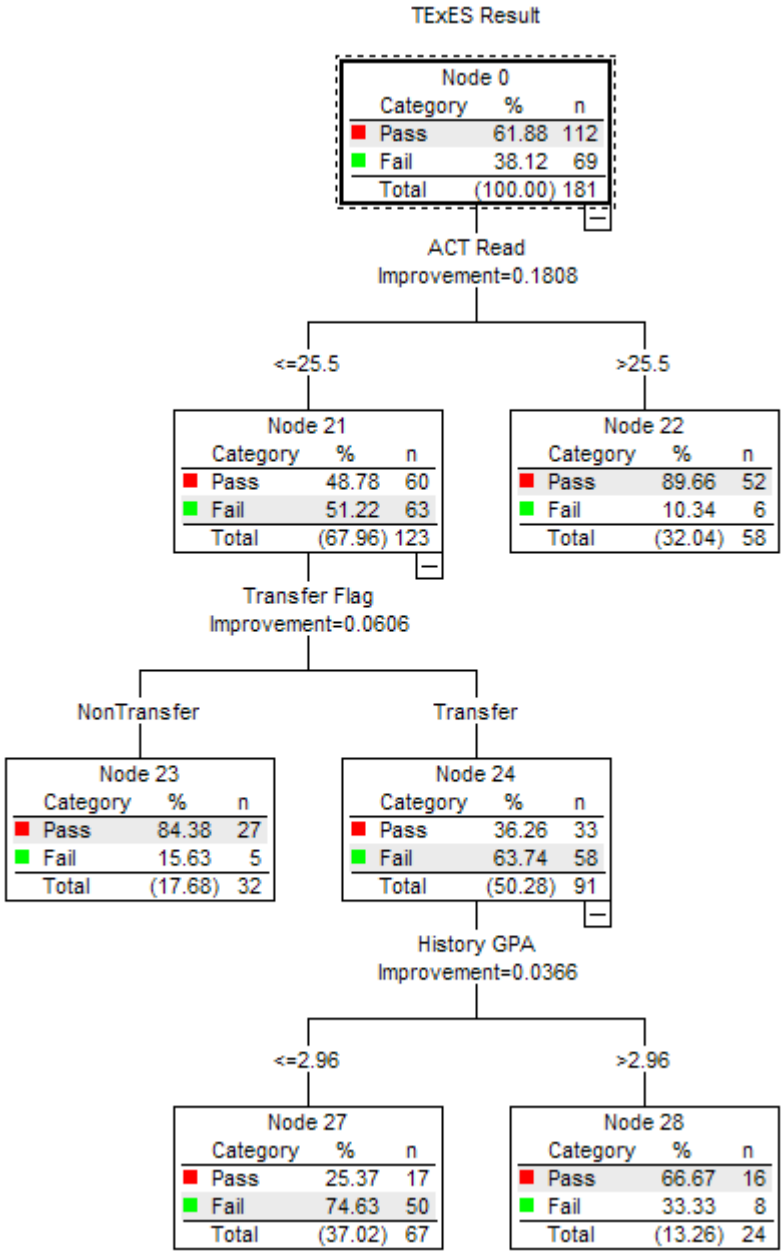

Figure 4.5. ACT model (standard error pruned). Misclassification risk = 19.9%.

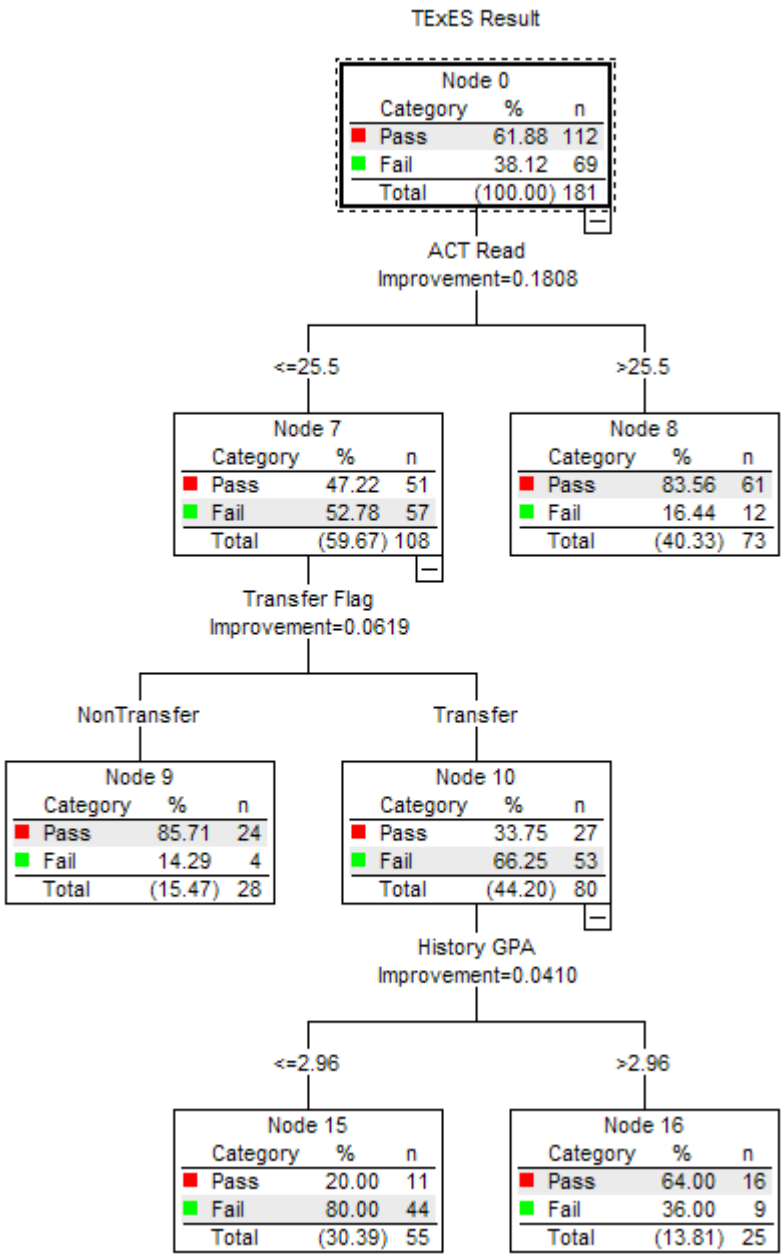

Figure 4.6. All Scores model (standard error pruned). Misclassification risk = 19.9%.

Chapter Summary

This chapter discussed the results of analyses using descriptive statistics, correlation analysis, and the construction of classification tree models for predicting failure on the TExES History (8-12) certification exam.

Although data from the Gruver (2008) study were available, highly homogenous results on the dependent variable and incongruities in how various independent variables were calculated between the current study and the Gruver study made the data unsuitable for inclusion with data from this study. Therefore, the current study focused exclusively on XTU data. Because ethnic groups other than White or Hispanic were too small in number to provide meaningful results, any analysis involving ethnicity involved only two groups, White and Hispanic.

All forms of GPA were significantly correlated with the outcome of the TExES exam. The ACT and SAT scores were also significantly correlated with the TExES result. Interestingly, the TASP reading score was significantly correlated with the TExES exam raw score, but was not significantly correlated with the TExES result (pass/fail). The TASP writing score was uncorrelated with any other variable.

Of the categorical variables, only transfer status, PostBac status, and ethnicity were significantly correlated with the TExES result. Gender, age, and number of history courses had little to do with the result of the TExES exam.

Because not all students had TASP, ACT, and SAT scores, five types of models were constructed and evaluated using two different pruning criteria (minimum risk vs. standard error), resulting in ten different models. The base model (No Scores)

included all variables except for TASP, ACT, and SAT scores. Three other models added to the base model by including the scores for TASP, ACT, and SAT, respectively. The last model (All Scores) included all variables.

Models pruned with the minimum risk criterion performed slightly better, but created models of more complexity and risked overfitting the data. Models overfitted to the existing data may not generalize well to other data. In attempting to reduce model complexity and overfitting, higher misclassification rates may result. Therefore, in choosing a model, one must seek to balance the issues of model complexity, lack of generalizability (overfitting), and misclassification rates. The SAT model (pruned using the standard error criterion) was chosen as the best overall model because it produced a model of less complexity than the ACT model, increased the proportion of cases correctly classified as Fail, and achieved this with no increase in misclassification rate.

After a model is chosen, it must be evaluated to determine if it performs better than a chance model. Three methods were discussed for assessing performance relative to chance—the maximum chance criterion, the proportional chance criterion, and Press' Q. All three methods produced statistically significant results for both the ACT model and the SAT model.

CHAPTER V

DISCUSSION

The purpose of this exploratory study was to address a specific set of research questions in order to develop a model for predicting failure on the TExES History (8-12) exam. To be meaningful and useful, the model should be able to predict the outcome significantly better than chance. The model building process relied upon classification trees—a nonparametric technique borrowed from the realm of data mining. The use of nonparametric procedures removes the burden of the restrictions associated with parametric procedures, most notably the assumptions of normality, linearity, and homoscedasticity. This chapter discusses the findings and significance of the results, implications for research and practice, and recommendations for future research.

Predictors of TExES Exam Result

The current study was different from previous studies in two important aspects: (a) the study included a much wider range of variables, and (b) nonparametric, classification tree methodology was used to build predictive models. Although classification trees have not been widely used in social science research, the success of the technique in the current study suggests that classification trees could be an effective alternative to the more traditional multiple linear regression and logistic regression procedures.

In a number of previous studies (Gruver, 2008; McIntosh, 2002; Pisani, Pisani, & Anderson, 2002; Weiss, 2003), the TASP reading score was a statistically

significant variable in predicting the outcome of the TExES exam. Those findings were confirmed by the current study, but only to a certain extent. The TASP reading score was significantly correlated with the raw score of the TExES exam, but was not significantly correlated with the TExES result (pass/fail). Previous studies used the TASP reading score as an independent variable rather than ACT or SAT scores.

The current study included reading scores from the ACT and the SAT. Each of the three reading scores had statistically significant correlations with the other reading scores. The ACT and SAT were more highly correlated with each other than they were with the TASP reading score, and were more highly correlated with the TExES result (pass/fail). Because the TASP writing score was not significantly correlated with any other variable in the study, the data suggest that TASP writing scores can be safely ignored when predicting the outcome of the History (8-12) TExES exam.

Reading Scores

Because the ACT and SAT were more highly correlated with each other than with the TASP reading score, one might speculate that the TASP reading exam is measuring reading ability in a different, and less useful, manner than the ACT and SAT exams, at least for purposes of predicting performance on the History (8-12) TExES exam. Although other researchers used the TASP reading score as a predictor variable, the data suggest that using the ACT or SAT reading scores provide better predictions.

Although there were statistically significant differences between Whites and Hispanics with respect to GPAs, reading scores, and pass rates on the History (8-12)

TExES exam, ethnicity never entered the models as a predictor. The reason for this is analogous to evaluating the significance of main effects in multiple regression when there is interaction among the variables. When interaction is present, the interpretation of main effects is untenable. In the current study, ethnicity was significant as a main effect in isolation, but when combined with other variables, the main effect was not useful as a predictor.

Transfer Status

Of the four categorical variables used in the study (transfer status, PostBac status, gender, and ethnicity), transfer status was the most highly correlated with the result of the TExES exam. Transfer status was one of the two most frequently used variables among all ten models; the other was history GPA. Approximately three-fourths of the candidates were transfer students. Non-transfer candidates had a pass rate of 86%, while transfer candidates had a pass rate of 54%. Expressed as an odds ratio, transfer students were 5.4 times more likely to fail the TExES exam. This difference in pass rates was highly significant. Gender and ethnicity were not significantly correlated with transfer status, and group differences in TASP, ACT, and SAT reading scores, lower division GPA, upper division GPA, history GPA, and total GPA were also not statistically significant.

If transfer students are significantly more likely to fail the exam, but transfer status is not related to gender or ethnicity, and differences in the various reading scores and GPAs are not statistically significant, then the reason for the difference in pass rates is puzzling. Based on the evidence, all that can be inferred is that being a

transfer student somehow increases the odds of failure. Another question of interest is why the majority of those entering the teacher education program are transfer students.

Post Baccalaureate vs. Baccalaureate

Post baccalaureate candidates were a small proportion of the sample (approximately 8%), and had a pass rate of 93%, compared to baccalaureate candidates who had a 59% pass rate. Baccalaureate candidates were 9.7 times more likely to fail the TExES exam than were post baccalaureate candidates. The data suggest that administrators need not be overly concerned about post baccalaureate candidates. The reasons for the difference in performance were not discernable from the available data.

Model Building

Statistician George E. P. Box is usually associated with the oft-quoted phrase, "Essentially, all models are wrong, but some are useful." Such is the case for the models developed during the course of this study. The objective was to develop a classification tree model based on decision rules that could predict the result (pass/fail) of the TExES exam significantly better than chance. Five different classification tree models were developed, subsequently pruned using two different procedures, producing ten models vying for selection as the best performing model. Unfortunately, no single model consistently outperformed any other model across all criteria. For example, the model that was the best for predicting Fail was not the best model for predicting Pass.

Balancing competing goals. Choosing the best model was based on determining a reasonable compromise among three competing goals—model complexity, prediction accuracy, and generalizability of the model. Seeking to increase prediction accuracy usually requires a model with more complexity, which often results in overfitting the existing data, leading to poor generalizability. Alternatively, seeking less complexity usually leads to better generalizability, but with poorer prediction accuracy. Compared to models pruned using the one standard error criterion, models pruned using the minimum risk criterion produced trees of greater complexity and increased risk of overfitting. Therefore, they were eliminated from consideration.

Theoretical vs. practical. Administrators in teacher education programs should be able to use the model to predict the outcome on the TExES exam (pass/fail) for teacher certification candidates. The best overall model was not deemed the best model for use in a practical sense. The best model was based on branches hierarchically splitting on the SAT reading score, upper division GPA, and transfer status, in that order. However, as previously pointed out, not all candidates will have a SAT reading score available. Therefore, in trying to predict the outcome of the TExES exam for a candidate who does not have a SAT reading score, the model would not be useful. In the absence of a SAT reading score, there would be no way to predict an outcome using this model.

It is interesting that transfer status was one of three key variables in the best overall model. As is sometimes the case, prediction does not necessarily provide

explanation. Transfer status, and its relationship with gender, ethnicity, reading scores, GPAs, and the TExES result, may be the most interesting and enigmatic of all the variables under study.

Across the five models, the average prediction rate for Pass was 82.5%, while the average prediction rate for Fail was 76.0%. The data suggest that it is more difficult to predict Fail than to predict Pass. The best model correctly predicted Pass 81.8% of the time, while correctly predicting Fail 78.3% of the time. The overall prediction accuracy was 80.7%, implying an overall misclassification rate of 19.3%. A chance model would predict Pass, and would be correct approximately 62% of the time.

The best model (SAT) was significantly better at predicting the outcome of the TExES exam than chance. The model had a prediction rate between 1.34 (MCC) and 1.57 (PCC) times better than chance, and was highly significant using the Press' Q criterion.

Although the resulting model may be satisfying in a theoretical sense, it may not be useful in practice, for reasons mentioned previously. The alternative is to choose one of the other four models for use in practice. However, three of them rely on a reading score from a standardized assessment. If the reading score is not available, the model cannot be used. The only remaining model (No Scores) includes none of the standardized scores from TASP, ACT, or SAT. The model creates branches based on history GPA, transfer status, total GPA, and number of upper division history courses. The inclusion of the number of upper division courses as a

predictor is puzzling because as a main effect, it was not significantly correlated with the TExES result. The appeal of this model is that it does not depend on any of the standardized scores from TASP, ACT, or SAT, and could be used for any candidate, provided data are available for the four variables in the model.

One peculiar outcome involved both versions of the TASP model; those with higher TASP reading scores were more likely to be classified as Fail. This counter-intuitive outcome is puzzling. One explanation may be that the TASP reading score was not significantly correlated with the TExES result (pass/fail), although it was significantly correlated with the raw score on the TExES exam. Based on this finding, neither of the TASP models should be considered viable models. This finding reinforces previously mentioned conclusions about using TASP scores for predicting the outcome of the TExES exam. The ACT and SAT reading scores are indicators that are more reliable and therefore, it is recommended that ACT and SAT scores be used as variables in preference to the TASP scores.

Another approach would be to have all models available, choosing the one best suited to the situation based on the information available. However, using multiple models increases the complexity of implementation. The All Scores model had the highest prediction rate for Fail (80.0%), the SAT model had the highest model rating and the lowest misclassification risk (19.3%), and the No Scores model offered the most flexibility. For a small increase in misclassification risk and a small decrease in the Fail classification rate, the No Scores model offers the best compromise and is the model recommended for use in practice.

Based on the weightings of all components, the SAT model produced the best overall model. For theoretical purposes, the All Scores model is recommended because it makes the maximum use of the available data, providing insight into the variables affecting the result of the TExES exam.

The alternative model (No Scores model) had four levels, with the first branch based on history GPA. If the history GPA is less than or equal to 2.96, the model predicts Fail; otherwise, Pass is predicted. The second splitting criterion was based on transfer status. If a candidate is a transfer student, the prediction is Fail; otherwise, the prediction is Pass. The third splitting criterion was based on total GPA. If the total GPA is less than or equal to 2.76, the prediction is Fail; otherwise, the prediction is also Fail. To resolve the ambiguity, another level was required, based on the number of upper division history courses. If the number of upper division history courses is less than or equal to seven, the prediction is Fail; otherwise, the prediction is Pass.

The No Scores model correctly predicted Pass 82.2% of the time, and correctly predicted Fail 76.2% of the time. The overall prediction rate was 80.1%. The prediction rate for Fail was 2.1 percentage points less than the best model (SAT model), but the prediction rate for Pass was 0.40 percentage points better than the best model. The misclassification rate of 19.9% was 0.6 percentage points more than the best model rate of 19.3%.

The trade-off for using the No Score model is increased flexibility but with slightly less efficiency, at least from a quantitative perspective. This seems like a reasonable trade-off to gain the flexibility needed to use the model in practice. The

model had a prediction rate between 1.29 (MCC) and 1.51 (PCC) times better than chance and was highly significant using the Press' Q criterion. Although the No Scores model was slightly less efficient than the best model, it was still a well functioning model, performing significantly better than chance.

Variables omitted from the models. Sometimes, reflecting on what was omitted from a model can be as interesting as reflecting on the variables included in the model. Gender, ethnicity, lower division GPA, age, and ACT/SAT math scores were not used in any of the models. Because there was no statistically significant difference in the age and gender of those who passed compared to those who failed, it is not surprising that these two variables were not included in any of the models. However, there were statistically significant differences in the others variables between those who failed and those who passed.

The ACT and SAT math scores were highly correlated with their respective reading scores, suggesting that the model could discriminate between those who passed and those who failed without using math scores. The ACT and SAT reading scores were more highly correlated with the TExES result (pass/fail) than were the math scores. The only math score to enter any of the models was the TASP, which was a factor only in the All Scores model using the minimum risk pruning criterion.

Figure 5.1 shows the number of models in which a variable was included in a classification tree. The two most frequently used variables were transfer status and history GPA, each occurring in eight different models. The next two most frequently

encountered variables were ACT reading score and upper division GPA, each occurring four times.

Transfer status is one of the two most frequently used variables for predicting the outcome on the TExES exam. The TASP Reading score, the variable often used as an independent variable in previous research, was one of the least useful variables for predicting the result of the TExES exam (pass/fail). The TASP Read variable was used only in the two TASP models, which were discredited as viable models because of the counter-intuitive (perhaps nonsensical) results. Based on the evidence, it seems prudent to avoid using the TASP reading score for predicting the outcome of the TExES exam when more effective variables are available—the ACT and SAT.

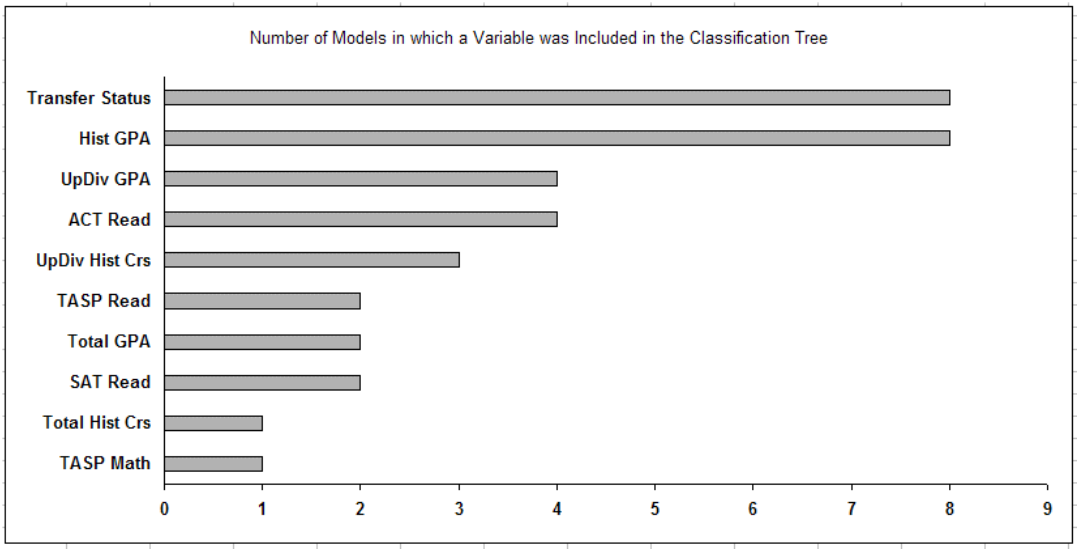

Figure 5.1. Frequency of occurrence of variables in classification trees

Implications for Research and Practice

Research Implications

The problem that provided impetus for this study was the low pass rate on the TExES History (8-12) certification exam. However, few studies have attempted to explore which variables might be helpful in predicting the outcome on the TExES content-area certification exam. In those studies, multiple linear regression was the dominant method used for predicting the raw score of the TExES exam, often with disappointing results.

The use of classification trees in the current study demonstrated that the method is well suited for handling prediction scenarios. The models that were developed were highly statistically significant, were of modest complexity, and easily understood. Although missing data are always an issue in a research study, classification trees are especially well suited to handling missing data because they reduce the amount of information loss due to missing data. Because the results of the current study were encouraging, others researchers should consider the use of classification trees as an alternative to regression.

For researchers who may conduct other studies involving the TExES content-area certification exam, the current study made clear the problem of using TASP scores as independent variables. It is recommended that the ACT and SAT scores be used in lieu of the TASP scores. The prominent influence of transfer status was an unexpected finding, of which other researchers should be aware. The current study also revealed that certain variables were of little use in predicting the outcome of the

TExES exam. Researchers should take note of those variables when conducting similar studies, possibly including them as independent variables to see if similar results are obtained.

A variable that was unavailable for the current study, but might be useful as a predictor variable, is whether an institution requires a candidate to take, and pass, the practice exam before taking the actual TExES exam. It seems plausible that requiring the practice exam would increase pass rates, and knowing this could lead to predictions that are more accurate.

Implications for Practice

Low pass rates for any TExES content-area certification exam are a cause for concern. Because of the serious consequences of failure for the institution and the candidate, the ability to predict which candidates are likely to fail is the first step toward averting this outcome. The current study demonstrated that models developed using classification trees were capable of predicting outcomes much better than chance. Based on relatively few criteria, the best model was able to correctly predict failure 78% of the time. For a TExES content exam with low pass rates, the ability to predict failure this well provides the opportunity for preemptive interventions. Requiring little or no instruction, the model is easy to interpret and implement in practice.

Not all institutions require candidates to take, and pass, the practice TExES exam before being allowed to take the actual TExES exam. However, it seems

reasonable to presume that this might contribute to higher pass rates. For institutions experiencing low pass rates, perhaps this policy should be reviewed.

Recommendations for Further Research

In general, because so few studies focused on the outcomes of TExES content-area exams, further research is needed to replicate the findings of the current study. Studies need to be conducted across a wide variety of institutions to assess the generalizability of the findings. In addition, further research is needed to explore which factors influence the outcome of other TExES content-area exams. The gap in the literature involving outcomes on TExES content-area exams needs to be filled.

The current study found that transfer students were 5.4 times more likely to fail than were non-transfer students. If transfer students are much more likely to fail the exam, but transfer status was not related to gender or ethnicity, and differences in the various reading scores and GPAs were not statistically significant, then the reason for the difference in pass rates is puzzling. Further research is needed to validate these findings, and if corroborated, to explore the reasons behind this finding.

The current study broadly defined transfer student as anyone having one or more transfer credits. However, a more nuanced definition may reveal some interesting differences among transfer students. For example, a student may be more properly classified as a non-transfer student if the only transfer credits were the result of courses taken in high school that were eligible to be counted as college credit. Alternatively, an otherwise non-transfer student may take an occasional course from another institution while home during the summer. Students in situations such as

these may be quite different from those students who attended a community college before transferring to a university.

Rather than treating transfer status as dichotomous, the variable could be treated as a continuous variable based on the number of hours transferred. If treated as a continuous variable, the classification tree algorithm may be able to determine a split based on the number of transfer hours. A more nuanced, layered treatment of transfer status may provide insight into the puzzling results found in the current study involving transfer status. It is recommended that future research take a more nuanced approach to defining transfer status.

Although differential pass rates as a function of ethnicity are a perennial topic of debate, ethnicity was not a significant predictor in the current study. This may be because previous studies focused only on the main effects of ethnicity. However, as the current study showed, ethnicity was not significant in a multivariate framework. Therefore, for researchers who may conduct similar studies, it is recommended that ethnicity be considered in a much broader, multivariate framework so that interaction effects can be investigated. In general, when using multiple linear regression, it is recommended to move beyond main effects and routinely test for interaction effects.

Conclusion

The consequences of failure on TExES exams are not trivial. High failure rates adversely affect the reputation of the institution and could result in academic sanctions. Students who fail the TExES certification exam are denied timely entry

into their profession and may suffer an opportunity cost that can affect them financially and psychologically. It is in everyone's best interest to minimize failures.

The ability to make reasonably accurate predictions of which students are likely to fail is an important tool in helping students achieve success. Few studies have attempted to explore which variables are useful for predicting the outcome of the TExES content-area certification exams. The current study focused on identifying factors that were influential in predicting the outcome on the TExES History (8-12) certification exam. Not only did the study include a broader set of variables, but took a nonparametric approach using classification trees.

The results of the study were encouraging. Statistically, the classification tree models were highly significant and capable of predicting outcomes well beyond what would be expected based on chance. Although the social sciences rarely use classification tree methods in data analysis, the encouraging results of the models developed in the current study provide other researchers a glimpse of the capabilities of classification trees.

Exploratory studies may be able to provide predictions, but generally cannot provide explanation of the phenomena under study. Having identified several important variables that are important in predicting the outcome of the exam, those variables can serve as guideposts for seeking a deeper understanding of the "why" behind the phenomena.

REFERENCES

Airasian, P. W. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education, 95*(3), 393-412.

Alexander, A. D. (1990). *An analysis of factors contributing to the prediction of teacher education candidates performance on the examination for the certification of educators in Texas (ExCET).* Unpublished doctoral dissertation, Texas Southern University, Houston, TX.

Alreck, P. L., & Settle, R. B. (2004). *The survey research handbook* (3rd ed.). New York: McGraw-Hill Irwin.

Angus, D. L. (2001). Professionalism and the public good: A brief history of teacher certification. *Thomas B. Fordham Foundation.* Retrieved March 10, 2010, from http://www.edexcellence.net/doc/angus.pdf

AnswerTree 3.1 user's guide. (2002). Chicago, IL: SPSS.

Barilla, A. G., & Jackson, R. E. (2008). The CPA exam as a postcurriculum accreditation assessment. *Journal of Education for Business, 83*(5), 270-274.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage Publications, Inc.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *The Future of Children, 17*(1), 45.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and regression trees*. New York: Chapman and Hall.

Byrd, J. K., & Williams, C. L. (2006). A value-added approach to program assessment in higher education: Examination of an educational leadership preparation program. *ERIC Document Reproduction Service No. ED493289*

Chambers, S., Munday, R., Sienty, S., & Justice, M. (1999). Predictors of success on the Texas state certification tests for secondary teaching. *College Student Journal, 33*(1), 10.

Clementine 11.1 node reference. (2007). Chicago, IL: SPSS.

Conant, J. B. (1963). *The education of American teachers*. New York: McGraw-Hill.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: John Wiley & Sons, Inc.

D'Costa, A. G. (1993). The impact of courts on teacher competence testing. *Theory into Practice, 32*(2), 104.

Darling-Hammon, L. (1997). *The right to learn*. San Francisco, CA: Jossey-Bass.

Darling-Hammon, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher, 31*(9), 13-25.

Dunham, M. H. (2003). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education.

Ferguson, R. F., & Brown, J. (2000). Certification test scores, teacher quality and student achievement. In D. Grissmer & J. M. Ross (Eds.), *Analytic Issues in the Assessment of Student Achievement*. Washington, DC: National Center for Education Statistics.

Garcia, R. A. (1987). *Identifying the academic factors that predict the success of entering freshmen in a beginning computer science course.* Unpublished doctoral dissertation, Texas Tech University, Lubbock, TX.

Gorth, W. P., & Chernoff, M. L. (1986). Overview of teacher certification. In W. P. Gorth & M. L. Chernoff (Eds.), *Test for teacher certification*: National Evaluation Systems.

Gratz, D. B. (2000). High standards for whom? *Phi Delta Kappan, 81*(9), 681-687.

Groebner, D. F., Shannon, P. W., Fry, P. C., & Smith, K. D. (2008). *Business statistics: A decision-making approach* (7th ed.). Saddle River, NJ: Prentice Hall.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Gruver, E. (2008). *Utilizing the history Texas examinations of educator standards to predict student success at three regional state universities.*, Texas A&M University - Commerce. Retrieved July 14, 2009, from Dissertations & Theses @ Texas A&M System.(Publication No. AAT 3318635).

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate analysis* (6th ed.): Pearson.

Harrell, P. E. (2009). Do state examinations measure teacher quality? *Educational Studies, 35*(1), 65-79.

Harwell, G. A. (1989). *The identification of major predictors of success of medical technology graduates of a new university health sciences center program.* Unpublished doctoral dissertation, Texas Tech University, Lubbock, TX.

Henderson, S. J., & Orr, S. P. (1989). Identifying students at risk for failure on the licensing examination for registered nurses. *ERIC Document Reproduction Service No. ED322197*

Hernandez, B. L. M. (1999). *The effect of cumulative grade point average and Texas academic skills program test scores on ExCET professional development test scores in undergraduate education majors at Texas Wesleyan university (examination for the certification of educators in Texas).* Unpublished doctoral dissertation, Texas Woman's University, Denton, TX.

Hess, F. M. (2001). Tear down this wall: The case for a radical overhaul of teacher certification. *Progressive Policy Institute* Retrieved March 8, 2010, from http://www.ppionline.org/documents/teacher_certification.pdf

Higgins, J. J. (2004). *Introduction to modern nonparametric statistics* (1st ed.). Pacific Grove, CA: Duxbury Press.

Hildebrand, D. H., & Ott, R. L. (1998). *Statistical thinking for managers* (4th ed.). Mason, OH: Thomson South-Western.

Holmes, B. J. (1986). Do not buy the conventional wisdom: Minority teachers can pass the tests. *Journal of Negro Education, 55*(3), 251-271.

Ivie, S. D. (1982). Why black students score poorly on the NTE. *High School Journal, 65*, 171.

Jackson, J. K. (2006). *Teacher certification content area tests: Predictors of teacher knowledge for post-baccalaureate secondary candidates.* Unpublished doctoral dissertation, University of North Texas, Denton, TX.

Jorgensen, M. A., & Hoffman, J. (2003). History of the No Child Left Behind Act of 2001 (NCLB). Retrieved June 3, 2009, from http://pearsonassess.com/NR/rdonlyres/D8E33AAE-BED1-4743-98A1-BDF4D49D7274/0/HistoryofNCLB_Rev2_Final.pdf

Justice, M., & Hardy, J. C. (2001). Minority students and the examination for the certification of educators in Texas (ExCET). *Education, 121*(3), 592-596.

Kane, M., Crooks, T., & Cohen, A. (1997). Justifying the passing scores for licensure and certification tests. *ERIC Document Reproduction Service No. ED411304.*

Kinnison, L., & Nolan, J. (2001). An examination of GPA, TASP, and ExCET scores for undergraduate teacher education candidates. *National Forum of Teacher Education Journal* Retrieved Feb 20, 2009, from http://www.nationalforum.com/Electronic%20Journal%20Volumes/Kinnison,%20An%20Examination%20of%20GPA,%20TASP,%20and%20ExCET%20Scores%20For%20Undergraduate%20Teacher%20Education%20Canidates.pdf

Kirkpatrick, D. W. (1992). Rethinking teacher certification. *ERIC Document Reproduction Service No. ED399254*

Kohavi, R., & Quinlan, R. (1999). Decision tree discovery. Retrieved March 13, 2010, from http://ai.stanford.edu/~ronnyk/treesHB.pdf

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill.

Laird, W. S. (1998). *The predictive effects of factors associated with prospective secondary teachers' performance on the examination for the certification of educators in Texas (ExCET) (prospective teachers).* Unpublished doctoral dissertation, Texas A & M University, Commerce, TX.

Lanier, J., & Little, J. (1986). Research on teacher education. In M. Whittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 527-569). New York: Macmillan.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: John Wiley & Sons.

Larose, D. T. (2006). *Data mining: Methods and models*. Hoboken, NJ: John Wiley & Sons.

Larsen, J. T. (2002). *Predictors of academic difficulty in first- and second-year medical students.* Unpublished doctoral dissertation, Texas Tech University, Lubbock, TX.

Lee, L. A. M. (1980). *Pre-nursing cognitive and non-cognitive variables as predictors of performance on state board test pool examinations for registered nurses.* Unpublished doctoral dissertation, Texas Tech University, Lubbock, TX.

Lim, T., Loh, W., & Shih, Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning, 40*(3), 203-228.

Littleton, M. (2000). *Accountability in teacher education: Systems and trends.* (ERIC Document Reproduction Service No. ED 441 021).

Lucas, C. J. (1994). *American higher education: A history*. New York: St. Martin's Griffin.

Lutz, F. W. (1986). Education politics in Texas. *Peabody Journal of Education, 63*(4), 70-89.

Madaus, G. F., & Pullin, D. (1987). Teacher certification tests: Do they really measure what we need to know? *The Phi Delta Kappan, 69*(1), 31-38.

McClave, J., Benson, P., & Sincich, T. (2008). *Statistics for business and economics* (10th ed.). Upper Saddle River, NJ: Pearson.

McDonald, D. R. (2000). *Prediction models for performance on the elementary and secondary professional development examination for the certification of educators in Texas.* Unpublished doctoral dissertation, Texas A & M University, Commerce, TX.

McGaghie, W. C. (1991). Professional competence evaluation. *Educational Researcher, 20*(1), 3-9.

McIntosh, S., & Norwood, P. (2004). The power of testing: Investigating minority teachers' responses to certification examination questions. *Urban Education, 39*(1), 33-51.

McIntosh, S. E. (2002). *The relationship between performance on the Texas academic skills program test, grade point average, and performance on the examination for certification of educators in Texas.* Unpublished doctoral dissertation, University of Houston, Houston, TX.

Menand, L. (2010). *The marketplace of ideas: Reform and resistance in the American university*. New York: W. W. Norton & Company.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5-9.

Nance, J. L., & Kinnison, L. R. (1988). An examination of ACT, PPST and ExCET performance of teacher education candidates. *Teacher Education and Practice, 5*(1), 25-30.

A Nation at Risk: The imperative for educational reform. (1983). Retrieved March 15, 2010, from http://datacenter.spps.org/sites/2259653e-ffb3-45ba-8fd6-04a024ecf7a4/uploads/SOTW_A_Nation_at_Risk_1983.pdf

Neville, P. G. (1999). Decision trees for predictive modeling. Retrieved March 13, 2010, from http://www.sasenterpriseminer.com/documents/Decision%20Trees%20for%20Predictive%20Modeling.pdf

No Child Left Behind Act of 2001. (2001). Public Law 107-110.

Noddings, N. (1998). Teachers and subject matter knowledge. *Teacher Education Quarterly, 25*(4), 86-89.

Olwell, R. B. (2005). James Conant's uncompleted revolution: Methods faculty and the historical profession, 1978-2004. *The History Teacher, 39*(1), 33-41.

Perot, H. R. (1984). Speech to the Texas state legislature, *House Journal of the 68th Legislature--second special session*. Austin: State of Texas.

Pisani, J. S., Pisani, M. J., & Anderson, R. J. (2002). Predictors of success for the Texas ExCET exam in a predominantly Hispanic university environment. *Teacher Education and Practice, 15*(3), 54-82.

Pitter, G. W., Lanham, C. H., & McGalliard, D. (1997). Licensure examination results as outcome indicators: Issues and challenges. *ERIC Document Reproduction Service No. ED409784*

Poelzer, G. H., Zeng, L., & Simonsson, M. (2007). Teacher certification tests: Using linear and logistic regression models to predict success of secondary pre-service teachers. *College Student Journal, 41*(2), 305-313.

Reynolds, R. J. (1995). The professional self-esteem of teacher educators. *Journal of Teacher Education, 46*(3), 216-227.

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. Hackensack, NJ: World Scientific Publishing Co.

Roth, R. A. (1996). Standards for certification, licensure, and accreditation. In J. Sikula, T. Buttery & E. Guyton (Eds.), *Handbook of research on teacher education* (2nd ed., pp. 242-278). New York: Macmillan.

Rubinstein, S. A., McDonough, M. W., & Allan, R. G. (1986). The changing nature of teacher certification programs. In W. P. Gorth & M. L. Chernoff (Eds.), *Testing for teacher certification*: National Evaluation Systems.

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research, 8*, 1625-1657.

SBEC. (1998). Report of accreditation ratings issued under accountability system for educator preparation (ASEP). Retrieved July 16, 2009, from http://www.sbec.state.tx.us/SBECOnline/brdinfo/agendas/1998_09/7.pdf

SBEC. (2001). ASEP policies and information. Retrieved July 14, 2009, from http://www.sbec.state.tx.us/SBECOnline/edprep/2001policyinfo.pdf

SBEC. (2004). ASEP frequently asked questions (FAQS). Retrieved June 14, 2009, from http://www.sbec.state.tx.us/SBECOnline/ASEP2/faq.asp

SBEC. (2008). TExES faculty manual. Retrieved April 6, 2009, from http://www.texes.ets.org/assets/pdf/testprep_manuals/texes_faculty_manual.pdf

Scheuneman, J., & Slaughter, C. (1991). Issues of test bias, item bias, and group differences and what to do while waiting for the answers. *ERIC Document Reproduction Service No. ED400294*

Shepard, L. A., & Kreitzer, A. E. (1987). The Texas teacher test. *Educational Researcher, 16*, 22-31.

Simonsson, M., Poelzer, H., & Zeng, L. (2000). *Teacher certification tests: Variables that determine success for secondary pre-service teachers.* Paper presented at the American Education Research Association, New Orleans, LA.

StatSoft. (2010). *Electronic Statistics Textbook.* Retrieved March 10, 2010, from http://www.statsoft.com/textbook/data-mining-techniques/?button=1

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Allyn and Bacon.

TEA. (2009). Accountability system for educator preparation (ASEP). Retrieved July 17, 2009, from http://www.tea.state.tx.us/index3.aspx?id=458

Tellez, K. (2003). Three themes on standards in teacher education. *Teacher Education Quarterly, 30*(1), 9-18.

TExES history (8-12) preparation manual. Retrieved February 23, 2009, from http://www.texes.ets.org/assets/pdf/testprep_manuals/133_history8_12_55055_web.pdf

Tinsley, R., & Hardy, J. C. (2003). Faculty pressures and professional self-esteem: Life in Texas teacher education. Retrieved March 24, 2010, from http://www.usca.edu/essays/vol62003/tinsley.pdf

Vogt, W. P. (2007). *Quantitative research methods for professionals*. Boston: Pearson - Allyn and Bacon.

Wakefield, D. (2003). Screening teacher candidates: Problems with high-stakes testing. *Educational Forum, 67*(4), 380-388.

Walsh, K. (2001a). Teacher certification reconsidered: Stumbling for quality. *ERIC Document Reproduction Service No. ED460100*

Walsh, K. (2001b). Teacher certification reconsidered: Stumbling for quality – a rejoinder. *ERIC Document Reproduction Service No. ED481389*

Ward, M. J., & Wells, T. J. (2006). The relationship between preservice teachers' reading ability and their achievement on teacher certification examinations. *Teacher Education and Practice, 19*(1), 14-23.

Warner, A. R. (1990). Legislated limits on certification requirements: Lessons from the Texas experience. *Journal of Teacher Education, 41*(4), 26-33.

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: Sage Publications.

Weiss, P. L. (2003). *Prospective elementary education teachers' performance predictors on the examination for the certification of educators in Texas.* Unpublished doctoral dissertation, University of Houston, Houston, TX.

Weitman, C. J. (1985). Teacher competency: A public farce! *ERIC Document Reproduction Service No. ED267030*

White, W. F., & Burke, C. M. (1994). Can Texas teacher certification be predicted from SAT scores and grade point averages? *Journal of Instructional Psychology, 21*(3), 298-299.

Wise, A. E., & Leibbrand, J. A. (2000). Standards and teacher quality. *Phi Delta Kappan, 81*(8).

Yu, L. M., Rinaldi, S. A., Templer, D. I., Colbert, L. A., Siscoe, K., & Van Patten, K. (1997). Score on the examination for professional practice in psychology as a function of attributes of clinical psychology graduate programs. *Psychological Science, 8*(5), 347-350.

Zeng, L., Simonsson, M., & Poelzer, H. (2002). Teacher certification tests: Variables that predict pass/fail status on elementary professional development examination for preservice teachers. *ERIC Document Reproduction Service No. ED464138*

APPENDIX A

Domain Descriptions for History (8-12)

TEST FRAMEWORK FOR FIELD 133:
HISTORY (8-12)

Domain I **World History (approximately 37% of the test)**
Standards Assessed:

History Standards IV–X:

History: The social studies teacher applies knowledge of significant historical events and developments, as well as of multiple historical interpretations and ideas, in order to facilitate student understanding of relationships between the past, the present, and the future.

Geography: The social studies teacher applies knowledge of people, places, and environments to facilitate students' understanding of geographic relationships in Texas, the United States, and the world.

Economics: The social studies teacher knows how people organize economic systems to produce, distribute, and consume goods and services, and uses this knowledge to enable students to understand economic systems and make informed economic decisions.

Government: The social studies teacher knows how governments and structures of power function, provide order, and allocate resources, and uses this knowledge to facilitate student understanding of how individuals and groups achieve their goals through political systems.

Citizenship: The social studies teacher understands citizenship in the United States and other societies, and uses this knowledge to prepare students to participate in our society through an understanding of democratic principles and citizenship practices.

Culture: The social studies teacher understands cultures and how they develop and adapt, and uses this knowledge to enable students to appreciate and respect cultural diversity in Texas, the United States, and the world.

Science, Technology, and Society: The social studies teacher understands developments in science and technology, and uses this knowledge to facilitate student understanding of the social and environmental consequences of scientific discovery and technological innovation.

Domain II **U.S. History (approximately 42% of the test)**
Standards Assessed:

History Standards IV–X:

History: The social studies teacher applies knowledge of significant historical events and developments, as well as of multiple historical interpretations and ideas, in order to facilitate student understanding of relationships between the past, the present, and the future.

Geography: The social studies teacher applies knowledge of people, places, and environments to facilitate students' understanding of geographic relationships in Texas, the United States, and the world.

Economics: The social studies teacher knows how people organize economic systems to produce, distribute, and consume goods and services, and uses this knowledge to enable students to understand economic systems and make informed economic decisions.

Government: The social studies teacher knows how governments and structures of power function, provide order, and allocate resources, and uses this knowledge to facilitate student understanding of how individuals and groups achieve their goals through political systems.

Citizenship: The social studies teacher understands citizenship in the United States and other societies, and uses this knowledge to prepare students to participate in our society through an understanding of democratic principles and citizenship practices.

Culture: The social studies teacher understands cultures and how they develop and adapt, and uses this knowledge to enable students to appreciate and respect cultural diversity in Texas, the United States, and the world.

Science, Technology, and Society: The social studies teacher understands developments in science and technology, and uses this knowledge to facilitate student understanding of the social and environmental consequences of scientific discovery and technological innovation.

Domain III **Foundations, Skills, Research, and Instruction (approximately 21% of the test)**
Standards Assessed:

History Standards I–III:

The social studies teacher has a comprehensive knowledge of the social sciences and recognizes the value of the social sciences.

The social studies teacher effectively integrates the various social science disciplines.

The social studies teacher uses knowledge and skills of social studies, as defined by the Texas Essential Knowledge and Skills (TEKS), to plan and implement effective curriculum, instruction, assessment, and evaluation.

Source: TExES Preparation Manual—History 8-12

APPENDIX B

Graphical Methods for Assessing Model Performance

Performance characteristics of classification trees can be displayed graphically. Three types of charts are commonly used—gains charts, response charts, and lift (index) charts. Excerpted from AnswerTree 3.1 User's Guide, the following describes how to interpret these charts. The associated charts for each of the five models are shown below.

Gains Chart

Gain is defined as the proportion of total hits that occur in each increment relative to the total number of hits in the tree. A hit is a case that has the correct category of the target variable. Gains are calculated as: (hits in increment / total number of hits) x 100. The cumulative gains are plotted using a gains chart (pp. 119-120).

Response Chart

Response is defined as a percentage of records in the increment that are hits. Response is calculated as: (responses in increment / records in increment) x 100. The cumulative response is plotted using a response chart (p. 121).

Lift (Index) Chart

Lift (Index) is defined as the percentage of records in each increment that are hits compared with the overall percentage of hits in the training data set. The lift chart is a plot of the values in the *Index* (%) column. Index values are calculated using the following equation:

$$\frac{\text{(hits in increment / records in increment)}}{\text{(total number of hits / total number of records)}} \times 100$$

The cumulative lift is plotted using a lift (index) chart (p. 122).

Gain Summary

Target variable: TExES Result Target category: Fail

Statistics

Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Resp: %	Index (%)
17	10	18	15	22.1	84.6	222.0
17;23	20	36	29	42.1	80.6	211.5
23	30	54	42	60.4	77.2	202.5
23;11	40	72	51	73.6	70.5	185.0
11;24;10	50	91	55	79.6	60.4	158.4
10	60	109	58	83.7	53.0	139.0
10	70	127	61	87.8	47.7	125.1
10	80	145	63	91.8	43.7	114.7
10	90	163	66	95.9	40.6	106.5
10	100	181	69	100.0	38.1	100.0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

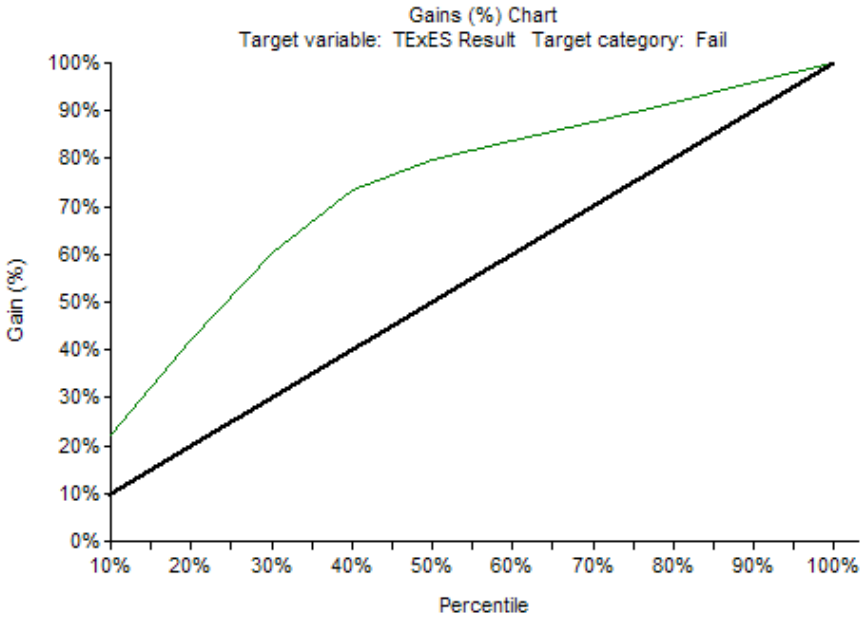

Figure B.1. Gains Summary and Gains Chart: No Scores model. Prune: one standard error. Misclassification risk: 19.9%.

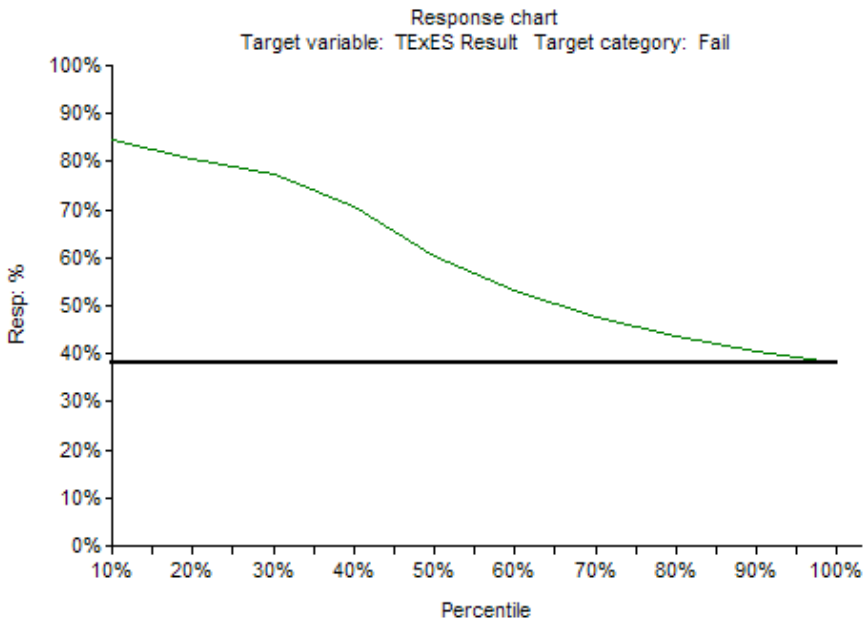

Figure B.2. Response chart: No Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

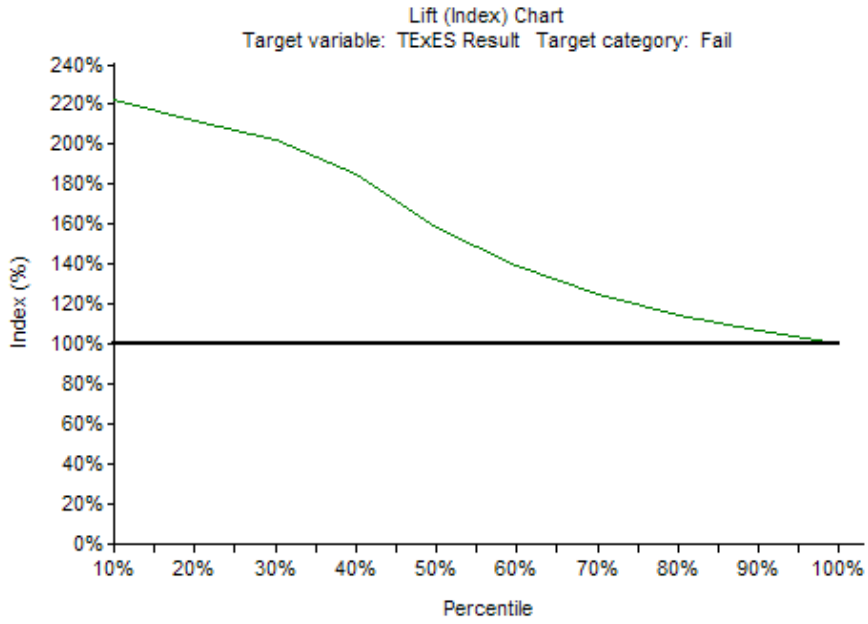

Figure B.3. Lift (Index) chart: No Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

Gain Summary

Target variable: TExES Result Target category: Fail

Statistics

Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Resp: %	Index (%)
8;3	10	18	14	20.0	76.7	201.3
3	20	36	26	38.0	72.7	190.8
3	30	54	39	55.9	71.4	187.3
3	40	72	51	73.8	70.7	185.6
3;7;2	50	91	56	81.0	61.4	161.1
2	60	109	59	84.8	53.7	140.8
2	70	127	61	88.6	48.1	126.3
2	80	145	64	92.4	44.0	115.3
2	90	163	66	96.2	40.7	106.8
2	100	181	69	100.0	38.1	100.0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

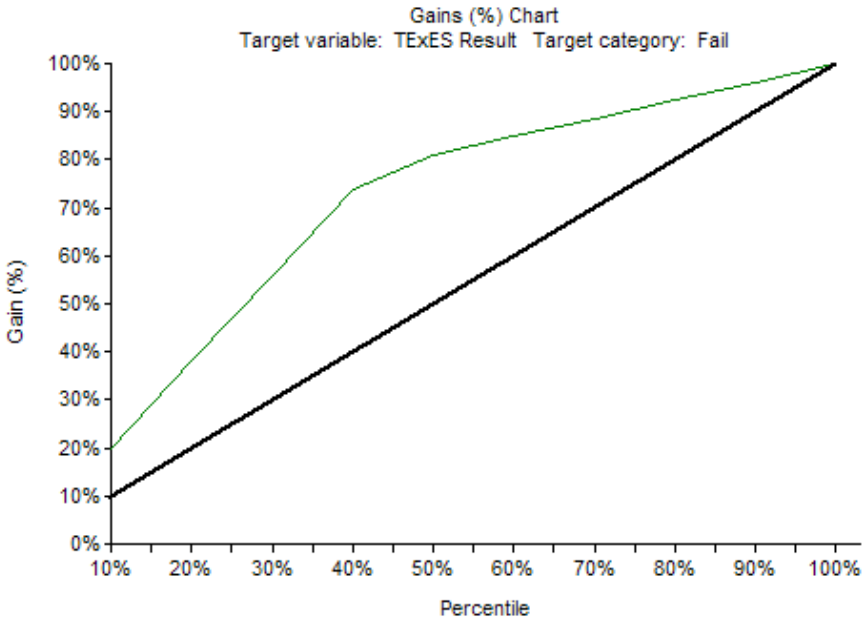

Figure B.4. Gains Summary and Gains Chart: TASP Scores model. Prune: one standard error. Misclassification risk: 19.9%.

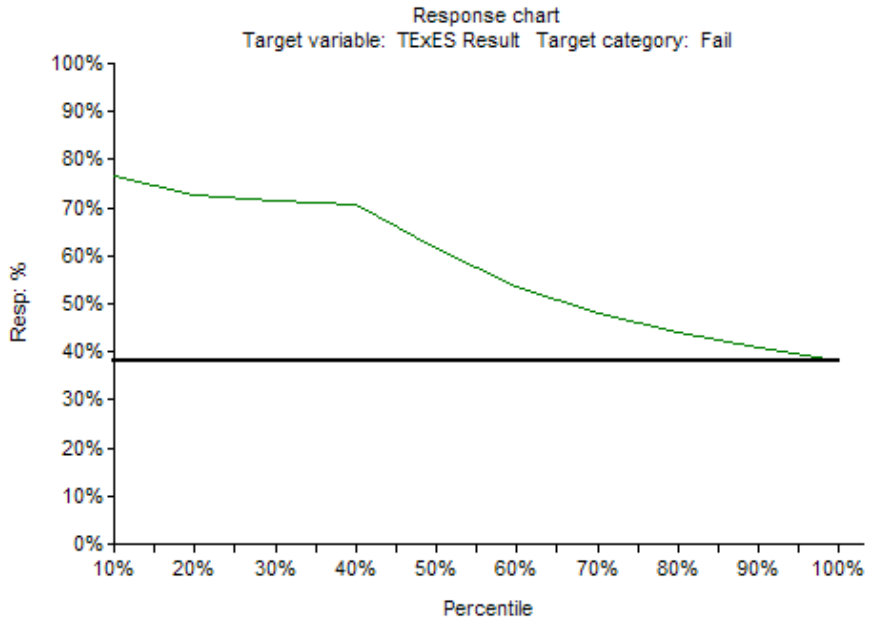

Figure B.5. Response chart: TASP Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

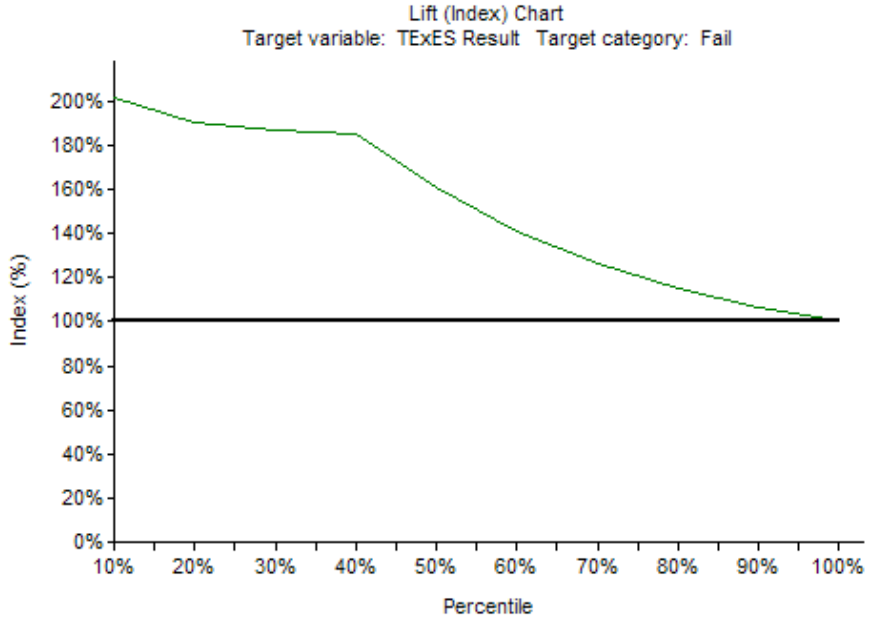

Figure B.6. Lift (Index) chart: TASP Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

Gain Summary

Target variable: TExES Result Target category: Fail

Statistics

Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Resp: %	Index (%)
27	10	18	13	19.5	74.6	195.8
27	20	36	27	38.9	74.6	195.8
27	30	54	40	58.4	74.6	195.8
27;28	40	72	52	74.9	71.8	188.2
28	50	91	58	84.1	63.7	167.2
23	60	109	61	88.1	55.8	146.4
23;22	70	127	63	91.9	49.9	131.0
22	80	145	65	94.6	45.0	118.1
22	90	163	67	97.3	41.2	108.0
22	100	181	69	100.0	38.1	100.0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

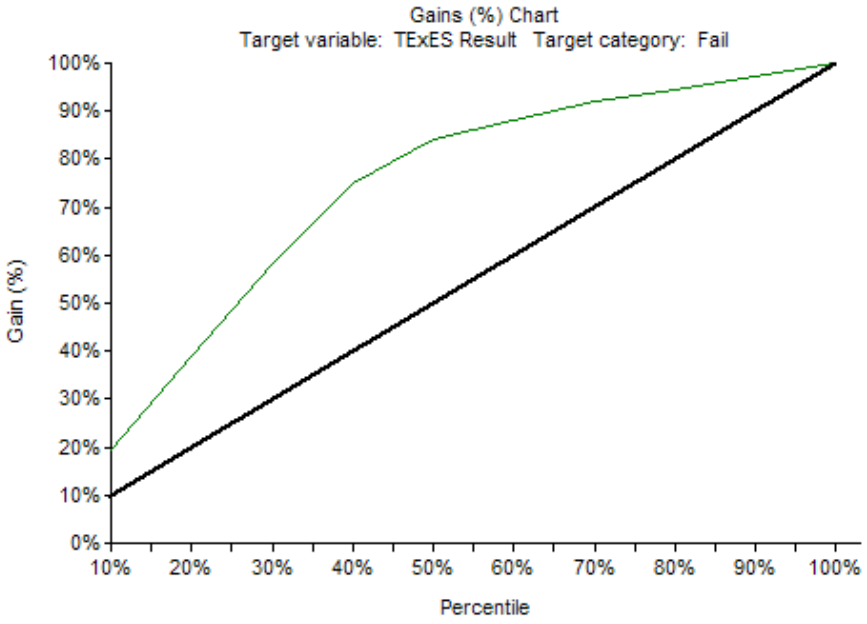

Figure B.7. Gains Summary and Gains Chart: ACT Scores model. Prune: one standard error. Misclassification risk: 19.9%.

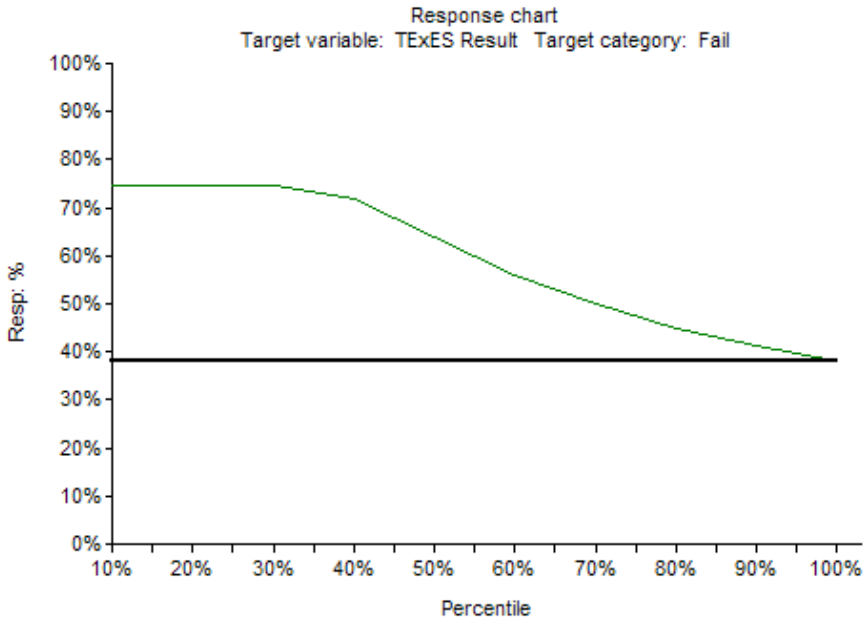

Figure B.8. Response chart: ACT Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

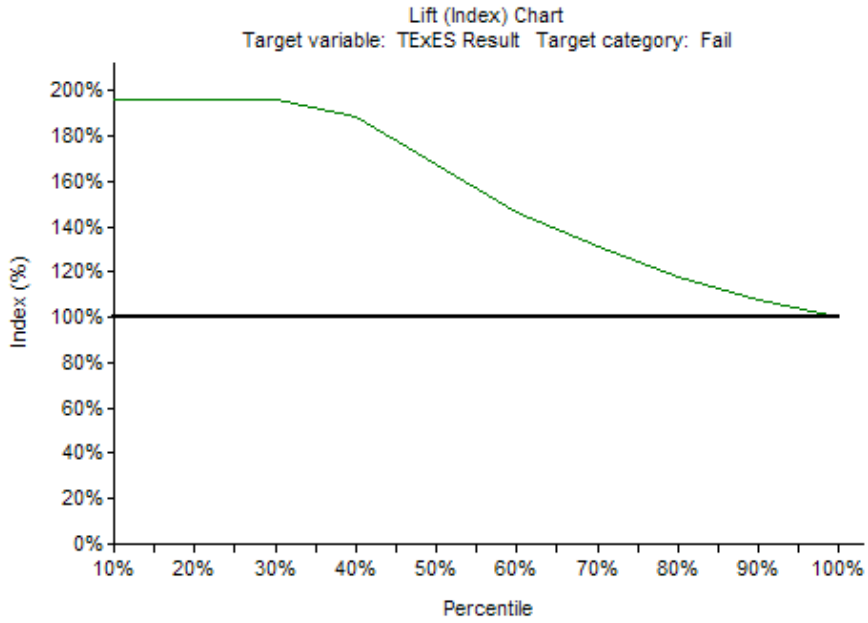

Figure B.9. Lift (Index) chart: ACT Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

Gain Summary

Target variable: TExES Result Target category: Fail

Statistics

Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Resp: %	Index (%)
8	10	18	14	20.4	78.3	205.5
8	20	36	28	40.9	78.3	205.5
8	30	54	42	61.3	78.3	205.5
8;7;4	40	72	51	73.6	70.5	184.9
4;2	50	91	55	79.1	60.0	157.4
2	60	109	57	83.3	52.7	138.3
2	70	127	60	87.5	47.5	124.7
2	80	145	63	91.7	43.6	114.4
2	90	163	66	95.8	40.6	106.4
2	100	181	69	100.0	38.1	100.0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

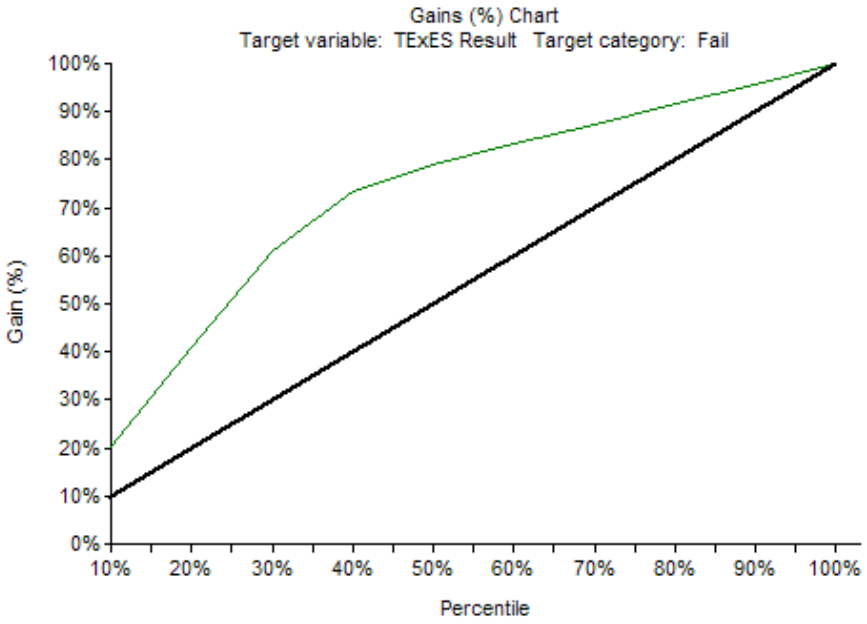

Figure B.10. Gains Summary and Gains Chart: SAT Scores model. Prune: one standard error. Misclassification risk: 19.3%.

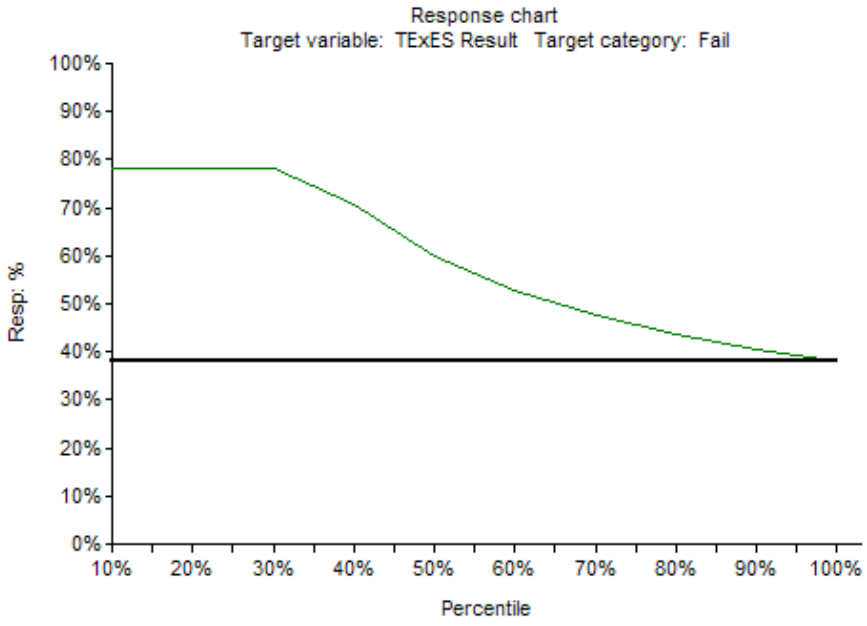

Figure B.11. Response chart: SAT Scores model.
Prune: one standard error. Misclassification risk: 19.3%.

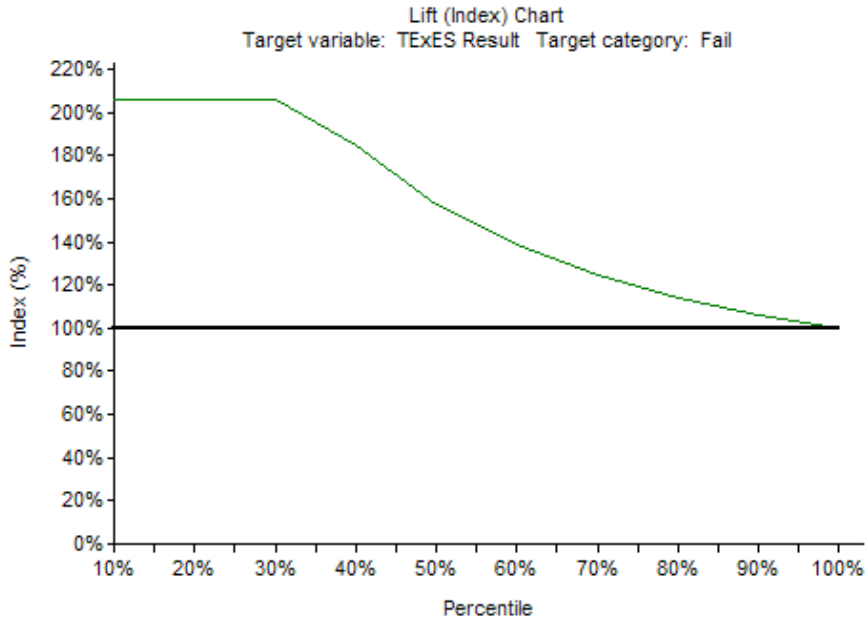

Figure B.12. Lift (Index) chart: SAT Scores model.
Prune: one standard error. Misclassification risk: 19.3%.

Gain Summary

Target variable: TExES Result Target category: Fail

Statistics

Nodes	Percentile	Percentile: n	Gain: n	Gain (%)	Resp: %	Index (%)
15	10	18	14	20.9	80.0	209.9
15	20	36	29	41.7	80.0	209.9
15	30	54	43	62.6	80.0	209.9
15;16	40	72	50	72.6	69.6	182.6
16;8	50	91	55	79.4	60.2	158.0
8	60	109	58	83.7	53.0	139.0
8	70	127	61	88.0	47.8	125.4
8	80	145	64	92.3	43.9	115.2
8;9	90	163	66	96.3	40.8	106.9
9	100	181	69	100.0	38.1	100.0

In versions prior to AnswerTree 3.0 the Gains column was known as Responses and vice versa.

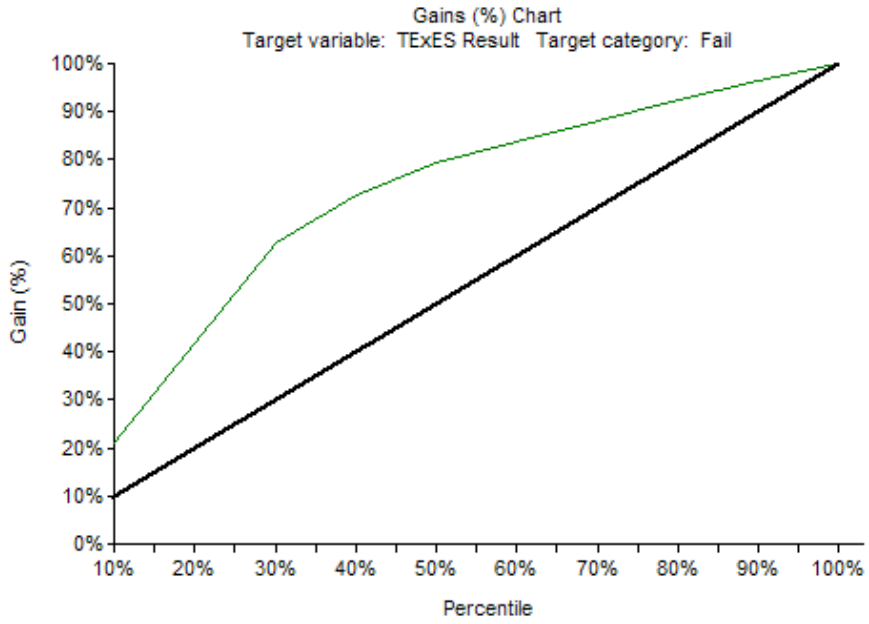

Figure B.13. Gains Summary and Gains Chart: All Scores model. Prune: one standard error. Misclassification risk: 19.9%.

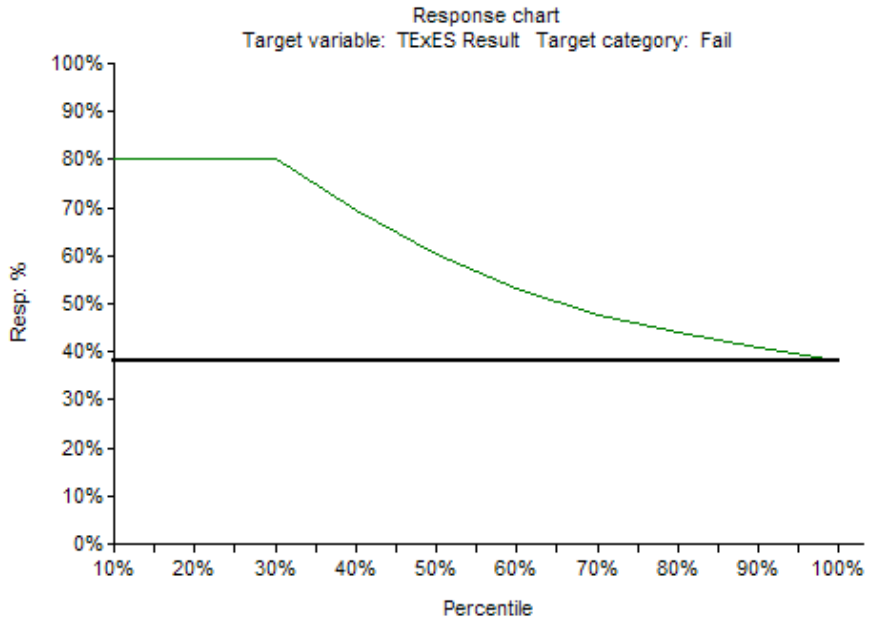

Figure B.14. Response chart: All Scores model.
Prune: one standard error. Misclassification risk: 19.9%.

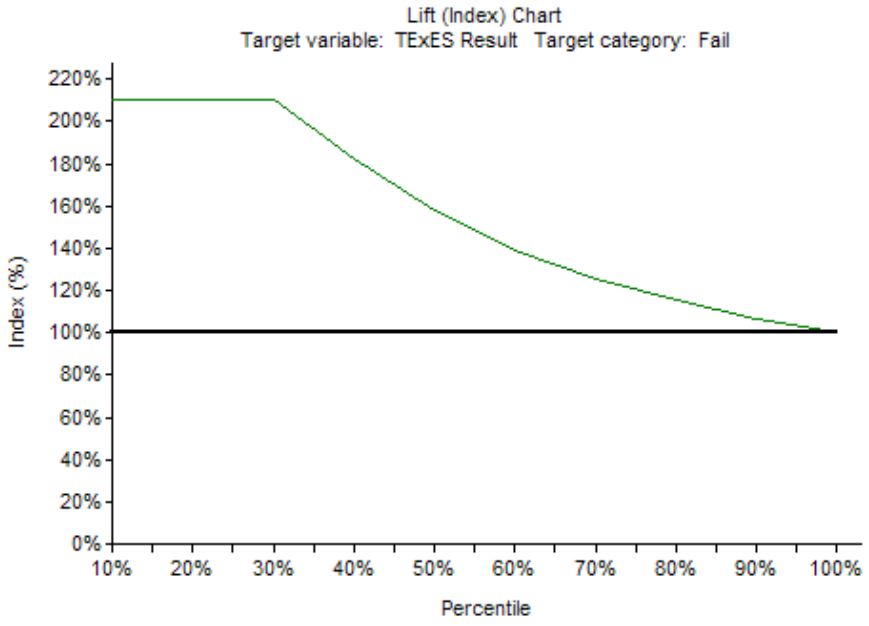

Figure B.15. Lift (Index) chart: All Scores model.
Prune: one standard error. Misclassification risk: 19.9%.