

OPERATIONALIZING PREDICTIVE FACTORS OF  
SUCCESS FOR ENTRY LEVEL STUDENTS  
OF COMPUTER SCIENCE

---

A Dissertation  
Presented to  
the Graduate School of  
Clemson University

---

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy  
Educational Leadership

---

by  
Kenneth Allen Weaver  
August 2004  
Advisor: Dr. Russell A. Marion

UMI Number: 3142985

Copyright 2004 by  
Weaver, Kenneth Allen

All rights reserved.

### INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

**UMI**<sup>®</sup>

---

UMI Microform 3142985

Copyright 2004 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company  
300 North Zeeb Road  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

July 30, 2004

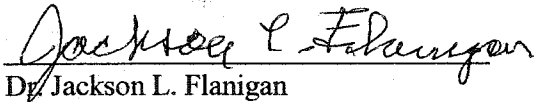
To the Graduate School:

This dissertation entitled "Operationalizing Predictive Factors Of Success For Entry Level Students Of Computer Science" and written by Kenneth Allen Weaver is presented to the Graduate School of Clemson University. I recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor Of Philosophy with a major in Educational Leadership.

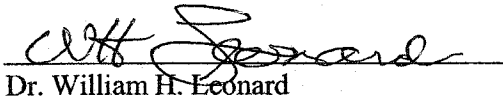


Dr. Russell A. Marion, Dissertation Advisor


We have reviewed this dissertation  
and recommend its acceptance:



Dr. Jackson L. Flanigan

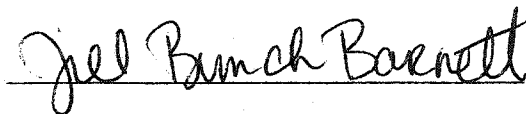


Dr. William H. Leonard



Dr. F. Catherine Mobley

Accepted for the Graduate School:



## ABSTRACT

This study undertakes to implement a predictive model of student success in the introductory course of computer science (CPSC 101) at a major southern university in the United States. The central issue is moving a predictive model from an “explanatory” state to an “operational” state within a student advising framework. The study’s premise is that what works for analytical purposes may diverge greatly when the model is implemented within a “real world” institutional framework and in the process encounters questions of data accuracy, and availability.

The study analyzes the achievement of 1,014 students who took CPSC 101 between the fall term of 1996 through the spring term of 2004. The primary independent variable under scrutiny is the Clemson Math Placement Test (CMPT) which is used to place students in their first calculus course by the mathematical sciences department, a co-requisite for taking the initial computer science course. The relationship between the student’s score on the CMPT and a student’s performance in computer science has been historically assumed by the computer science department, but never tested until this study.

The analytical design uses multiple and logistic regression processes, the former to define a predictive model for student achievement in an introductory computer science course and the latter to test the efficacy of the model. The model developed and tested shows weakness overall with an explanatory  $R^2$  values of .168 and a decided inability to deal with the case of predicting the unsuccessful student, incorrectly classifying the student outcome 60% of the time. Further, the underlying data elements supporting the

prediction are extremely limited and in some cases, questionable in their validity and utility. Much of the predictive model's failure can be traced to the differing environments separating a purely analytical or "explanatory" model, and the compromises that must be made to bring that model to bear "operationally" for predictions in real world situations.

## DEDICATION

I dedicate this work first to my parents: my father, Thomas Edison Weaver, CWO2 USN (Ret) (1917 – 1999), a member of that “greatest generation” who left the farm in the ninth grade, completed his high school education in the midst of war, and who throughout his life demonstrated and taught me that you can never get enough education, and to look for and appreciate it at every opportunity. To my mother, Ruth Marie Weaver, who taught me that opportunities and help are always present.

A special dedication to my wife Barbara, whose love and support is unconditional and undeserved.

## ACKNOWLEDGMENTS

I could not have completed this work without the encouragement of my committee chair, mentor, and friend, Dr. Russ Marion. I am grateful for his continued faith in me and his unique ability to know exactly which of my buttons to push and move me to the next level. I am especially grateful for his active modeling of what it means to be a scholar.

Thanks also to the other members of my committee, Drs. Jack Flannigan, Bill Leonard, Catherine Mobley and Wayne Madison. Thank you for spending your time making my work, and my learning better. A special thanks to Dr. Bonnie Holaday, Dean of the Graduate School, for her continued encouragement and friendship.

I am also grateful to my many classmates who through mutual encouragement, expanded the bounds of where and how learning takes place. I hope I was able to return the favor. Special thanks as well to Ms. Katherine Dobrenen, and Joan Alexander whose assistance and encouragement is appreciated and invaluable.

In no small measure could I understate the contributions and support of my family. No husband, father, or grandfather, has ever been more blessed than I for the continuing love and encouragement given during this endeavor. I would not have been able to undertake, much less complete, this work without their ever-present support. Thank you – Barbara, Michael, Joy, Patrick, Connie, Andy, Macy, Daniel, Isaac, and Carly.

## TABLE OF CONTENTS

	Page
TITLE PAGE .....	i
ABSTRACT .....	ii
DEDICATION .....	iv
ACKNOWLEDGMENTS .....	v
LIST OF TABLES .....	viii
 CHAPTER	
I. INTRODUCTION AND STATEMENT OF THE PROBLEM .....	1
Introduction .....	1
Statement of the Problem .....	2
Research Questions .....	5
Significance and Limitations of the Study .....	6
II. REVIEW OF THE LITERATURE .....	7
Introduction .....	7
Conceptual Construct of the Literature .....	8
Explanatory Studies .....	9
Testing Specific Skills or Traits .....	19
Predictive Studies .....	23
Summary of Prior Modeling .....	25
Revisiting the Conceptual Construct of the Literature .....	39
Revisiting the Problem .....	39
III. METHODOLOGY .....	43
Introduction .....	43
Data Sources .....	44
Descriptions and Limitations of the Data .....	45
The Model .....	54
Analysis and Procedures .....	55



## Table of Contents (Continued)

	Page
IV. RESULTS .....	57
Organization of the Results .....	57
Difference in Grades Before and After Implementation of CMPT .....	57
Analysis of CPSC 101 Data .....	58
V. DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS.....	64
Themes and Threads.....	64
Summary of the Research Questions.....	71
Limitations of the Study .....	77
Recommendations and Areas for Further Research .....	77
APPENDICES .....	82
A. Sample CMPT Test Questions.....	83
B. IRB Approval .....	88
LIST OF REFERENCES.....	89

## LIST OF TABLES

Table	Page
1. Model Matrix .....	26
2. Model Matrix – Significant Factors .....	30
3. Model Matrix – Significant and “Available” Factors .....	33
4. Model Matrix – Significant, “Available,” and Leading Factors .....	35
5. Model Matrix – Final .....	38
6. Data Sources .....	45
7. Data Variables.....	49
8. Added/Recoded Variables .....	50
9. Initial Analysis Model .....	55
10. Regression CPSC 101 – Full Model .....	59
11. Regression CPSC 101 – CMPT/Race Model.....	59
12. Regression CPSC 101 – CMPT Only Model.....	61
13. CPSC 101 – Full Model Classification Table.....	62
14. Classification Table – CMPT/Race .....	62
15. Classification Table – CMPT Only.....	63

CHAPTER I  
INTRODUCTION AND STATEMENT  
OF THE PROBLEM

Introduction

Helping students select courses that meet academic criteria is an advising function that is central to standards of performance for institutions of higher learning (CAS, 2003). Across disciplines and institutions, there have been different processes developed to match students with courses, including efforts with automated advising (McKendree & Zaback, 1988; Timmreck, 1968). The goal of any advisor is to correctly provide information that matches the student's academic goals and academic capacities at the moment; the student must be capable of performing the work required while at the same time being challenged to learn new skills in the process. The down side of this process is the danger of placing students in a position of being challenged to a point where they are unlikely to succeed because of inadequate preparation or development.

Coupled with this advising function is the desire for academicians to have a sense of how successful students will be in their chosen course of study. To that end, predictive models have been explored to identify those background characteristics, experiences, and traits that might suggest that a student would be successful in attempting certain courses or course of study. As will be show subsequently (Chapter II), the vast majority of research to date has not tested the leap from the analytical (explanatory) to the operational (predictive) where the "real world" issues of data availability or quality are associated with the analysis. In other words, the research to date tends to proceed from the

perspective of academic inquiry (explanatory) only, and not from the perspective of an advisor owing a student an answer (predictive).

### Statement of the Problem

This latter point is the heart of the problem under investigation. What happens to predictive models of student success in computer science when elements of the predictive model are not available at the time they are needed or not available at all? Does the student data available enable a prediction that is sufficiently accurate to be of value to both the advisor and the student? What are the resources associated with providing such prediction and do they provide marginal benefit over current advising processes in a way that justifies expenditure of additional resources? This study explores these questions by seeking to operationalize a predictive model of student success in introductory computer science courses at a major southern university.

At this institution, students of computer science (as well as other disciplines) come to their academic program with varied backgrounds and skill levels. The information about the student that is available to advisors is limited, usually limited to SAT/ACT math and verbal information, demographic data, and a predictive measure (internal to the institution) of potential grade point performance. While this is interesting data, and as will be noted later, some are elements in predictive modeling in the literature, the data are not always available to the academic advisor while the student is with the advisor. Nor are the data in a form that is particularly useful to either the student or the advisor.

At this institution and department, first year student advising takes place during 10 to 12 summer orientation sessions, where the matriculating students meet with the advising team and representative course schedules are discussed based on a series of most

common scenarios. The introductory course for most computer science students is CPSC 101, a course in the JAVA programming language and the first of a two-course sequence that covers both WINDOWS<sup>®</sup> and UNIX<sup>®</sup> environments. A third course, CPSC 212, introduces the student to algorithms and data structures, which when considered in concert with CPSC 101 and 102, represent the “gateway” introductory courses to the curriculum.<sup>1</sup>

Alternatively, CPSC 104 (an introductory course in programming logic and problem-solving) has been introduced within the last academic year to provide an exploratory course for those students who have not had any prior programming experience, who are unsure of their readiness for computer science, or who do not qualify to take the first calculus course. The latter point is important as it has been the computer science department’s long standing policy that a co-requisite for enrollment in CPSC 101 is enrollment in MTHSC 106, the first course in the calculus series required by the department’s curriculum. This requirement acknowledges the traditional linkage between computer science and competency in math (Konvalina, Wileman, & Stephens, 1983; Ralston & Shaw, 1980) and accreditation standards (Accreditation Board for Engineering and Technology, 2003).

The linkage has been strengthened in recent years because the mathematical sciences department at this university implemented a math placement test during the 2001 academic year (the Clemson Math Placement Test – CMPT) to determine the student’s readiness for various levels of math as required by a number of disciplines across the

---

<sup>1</sup> “Gateway” in the context of the curriculum does not suggest a process of “qualification” for the student to continue in the major, only that these three courses represent the base set of knowledge that the student should master in order to proceed.

institution (Department of Mathematical Sciences, 2004b). Student performance on this qualifying test is almost always available to the advising team, and even when it is not (the students will still know their performance prior to the start of classes), the results provide a common benchmark for all entering students and provides guidance as to their entry level computer science courses.

The Mathematical Sciences Department has validated the test for predictive use for math placement and has stated that the test (in conjunction with a first day skills quiz) successfully predicts success (a "C" or better in the course) 76% of the time, independent of SAT or ACT performance (Department of Mathematical Sciences, 2004a).

The CMPT has not been tested for relational significance to computer science performance however, and to date the CMPT results have influenced only the co-requisite requirement of the computer science curriculum. Student performance in the entry level course of computer science historically has been of concern with some terms anecdotally being attributed with a 50% non-success rate. "Success" is defined as a grade of "C" or better to progress to the next computer science course in the curriculum. Thus the non-successful student will have received a grade of "F" or "D," or alternatively will have withdrawn ("W") from the course before completing it. An analysis performed at the end of the fall 2003 term showed that the non-success (DWF) rate for CPSC 101 was 31.25%, a level that is on the margin of concern as to student performance (A. W. Madison, personal communication, May 3, 2004).

The purpose of this inquiry is to explore the CMPT's utility as a predictor of student performance in entry level computer science classes and to determine if it will perform in the absence of other indicators of success. An added purpose is to determine if, of the data collected prior to the student's arrival, there is an improvement of predictive

power that would justify the added resources and expense of including that data in a predictive process for pre-entry advising and placement.

### Research Questions

The overarching question of this study then is “What information about newly matriculating students is useful to predict their success in the introductory course of computer science at this university; and what role, if any, might the Clemson Math Placement Test (CMPT) play in such a prediction?” A series of six (which increases to seven in Chapter II) Research Questions suggest themselves to analyze and answer that question:

- RQ1: Is there a relationship between CMPT scores and CPSC 101 student outcomes?

While the CMPT has been validated for use with predicting potential outcomes for calculus and math courses, it has not been subject to scrutiny as to its relationship to computer science outcomes.

- RQ2: If there is a relationship, is it significant?

Is any relationship between the CMPT and computer science student outcomes statistically significant?

- RQ3: If significant, how well does the CMPT predict student performance?

Does the CMPT actually have predictive power for student success (“C” or better) in entry level computer science courses?

- RQ4: What other factors, in addition to the CMPT should be in the predictive model?

Are other student data available that would strengthen the predictive model?

- RQ5: How well do they (RQ4) predict performance?

What is the optimal predictive model of student success in entry level computer science courses at this institution based on available information?

- RQ6: Does the model have predictive power for other computer science courses in the curriculum sequence?

Does the model have any predictive merit as to the student's chance of success through the "gateway courses," that is, the first three introductory courses of the curriculum (CPSC 101, 102 and 212)?

### Significance and Limitations of the Study

This inquiry is obviously limited to the implementation of predictive models at one institution. Still, the results should be of interest to a number of audiences especially those involving technical skills such as in engineering and science disciplines.

For the institution, the study should be of interest because it adds to the literature concerning handling, manipulating, and archiving student data. Data are good only to the extent that they help in answering a question and there may be systemic issues regarding data availability and consistency that requires institutional attention.

This study should also add to the debate of whether this university's computer science department should consider creating and implementing a computer science-specific "qualification" exam or test for entering students. To do so requires an expenditure of resources and the question of cost to the institution and benefit to the student will have to be addressed. This study should aid that analysis.

Finally, this study should also be of significance specifically to advisors of computer science students as it represents an addition to the literature of using predictive models of success for student advising procedures and processes. More importantly it will represent one data point in the small group of literature where predictive models have morphed from academic exploration to operational use.



## CHAPTER II

### REVIEW OF THE LITERATURE

#### Introduction

The literature on academic success in computer science is extensive, somewhat repetitive, and for the purpose of this inquiry, classifiable into two categories (Pedhazur, 1982). First are studies which attempt to discover critical factors that explain a student's success in computer science. In one sense this group seeks to explore the widest possible range of potential variables to explain what contributes to student success. A subset of this group explores student success factors using a specific trait, such as gender, reasoning skill, or a common experience. The use of this sub-category does not mean to suggest that the researchers do not consider other factors, only that they approach the problem with the intent to test a specific trait or experience. All of the studies in this explanatory category to one extent or another can be described as "contextual studies" (Pedhazur, 1982, pp. 540-542); that is they seek to discover the impact of common backgrounds or experiences on student performance.

The second category of literature is those studies that attempt build on the explanation to formulate predictive models of student achievement. Studies in this grouping test a model for predictive value under a given set of circumstances. The study I propose falls into this latter group of studies, which of necessity includes elements of the former.

### Conceptual Construct of the Literature

At this point it seems appropriate to reflect on the literature in light of the initial taxonomy presented above. As noted, the literature falls into two groups, explanatory and predictive, following Pedhazur's (1982) classification of the purposes for undertaking a study.<sup>1</sup> Related to Pedhazur's construct, Flanigan, Marion, and Richardson (1996) undertook a causal study of student outcomes and increased educational funding, but in the process developed a student achievement model containing three levels of independent variables: contextual, demographic, and main or "performance" variables (R. A. Marion, personal communication, June 2, 2004). A visualization of that relationship is presented in Figure 1.

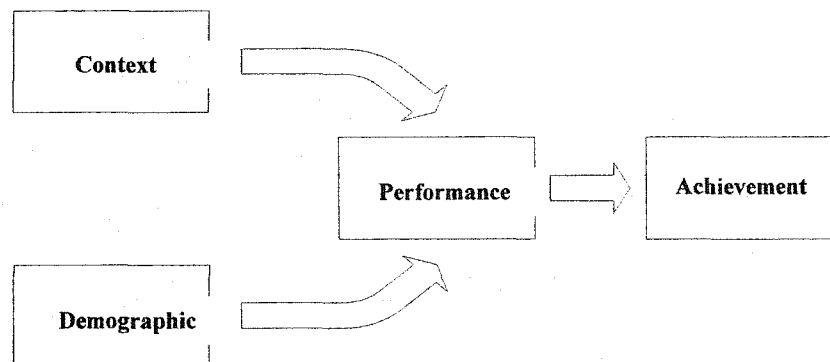


Figure 1. Relationship of Variables

With this visualization, the relationship of the variables to be explored in the literature, and their relationship to the predictive nature of the current study, may be more

---

<sup>1</sup> To which some add a third purpose – causality (R. A. Marion, personal communication, June 2, 2004).

evident. Phrased another way, contextual variables (high school background and preparation, high school or college courses taken, reasoning skill or intellect, etc.) and demographic variables (age, gender, race, etc.) contribute to the student's performance on intermediary measures (SAT/ACT/CMPT scores, specialized test instruments, etc.) which in turn impacts (and potentially predicts) the student's achievement on the measure of merit (dependent variable), in this case the bench mark of success in introductory computer science courses.

### Explanatory Studies

A series of prior work concentrates on identifying demographic and/or contextual factors present in a student's background or experience that explain level of success in introductory computer science courses, and by implication through the remainder of the student's program of study. The basic methodology of this body of work is to identify a group of pertinent factors about a sample population, usually students taking a computer science course, and whose performance in that course provides the bench mark against which the model is evaluated.

Konvalina, Wileman, and Stephens (1983) focused their work on discovering the differences in background or ability between those students who completed or withdrew from a beginning computer science course. They were not so much interested in how well the student performed in the class but only that the student persisted in completing it. Their model used nine demographic, contextual and performance factors: age, estimated high school performance, hours worked per week, prior computer education, prior non-programming computer work, prior programming work, years of high school mathematics, number of college mathematics courses, and total number of high school and college

math courses. In addition to the demographic and contextual data, the researchers collected performance data by testing five dimensions of computer science related skill through a specialized instrument (which contained what they termed the “computer science aptitude predictor” p. 377) previously developed by the researchers (Wileman, Konvalina, & Stephens, 1981). These skill factors were student performance regarding: number and letter sequences, logic-type questions, calculator simulator, algorithms, and word problems from high school algebra. Data were gathered from groups of students taking their first technical computer science course.

The researchers discovered that of these factors, students who persisted (did not withdraw) in the first computer science course were older, had better high school academic performance, had significantly more previous computer science education, had a more extensive background in college math, and more total math courses in high school and college than those students who did not persist. The other demographic/experiential factors were found not to be significant. Student performance on the specialized skills exam (aptitude predictor) showed that all elements except the section on algorithms were considered “good” discriminators between the groups of withdrawers and non-withdrawers.

From their analysis, the researchers concluded that the predictor portion of the instrument could be used as a placement tool. They also argued that, based on their analysis, their institution should see an estimated reduction from a 40 to 23 percent withdrawal rate from the first computer science course. The researcher’s argument regarding math skills and previous computer science education is an important finding, as it reinforces long held curriculum expectations (Accreditation Board for Engineering and Technology, 2003). Further, their use of the specific skills exam to predict actual results is

important validation that performance in computer science is predictable based on assessment of math, computer, and general academic skills.

The measure of merit for their work was to determine whether a student would withdraw from the course as opposed to completing the course “successfully” with some minimally acceptable grade. The authors noted the importance of the predictive instrument’s utility in the model as opposed to demographic or experiential factors and they concluded that their model would not be as predictive if it was not possible to administer a specialized exam to the students.

Campbell and McCabe (1984) focused on demographic, contextual, and performance factors by structuring their study to use information already available from their institution’s registrar by limiting their model’s factors to previously reported high school performance. These factors were: SAT-Math; SAT-Verbal, High School (HS) graduating rank, HS graduating class size, HS math semesters, HS science semesters, HS English semesters, average HS math grades, average HS science grades, average HS English grades, and gender. Their goal was to determine if there were differences in the background of those students who persisted in computer science courses compared to those who did not. Differing from the previous study, the researchers sought specifically not to administer any specialized instruments of skill or knowledge.

The measure of merit for the Campbell and McCabe study was whether a student persisted in the computer science major through the middle (second term) of their sophomore year as opposed to a specific “success” threshold of achievement in terms of grades achieved or scores on an outcome test. The researchers tested this by grouping the students (as of their sophomore year) into three groups based on academic major (computer science, engineering or other science, and all other majors) and then looking for

significant differences between the students' major classification and the model's factors regarding persistence. The researchers also looked for which combination of the model's factors best predicted a student's outcome; that is persisting or not persisting in computer science through the middle of their sophomore year. As with the previous study, the researchers did not set a standard for "successfully" completing the course based on a minimally acceptable grade or score; they were only interested in whether students persisted in the course.

The analysis associated with the Campbell and McCabe work focused on two areas: (1) determining significant differences between persisting and non-persisting computer science/engineering and science students (the researchers in fact found no significant difference between these two groups) and persisting and non-persisting students in the "other" categories of majors; and (2) determining the appropriate contextual classification predictors. For the former, SAT scores (math and verbal), high school rank and background in high school math and science were found to be significantly higher for the persisting group compared to the non-persistors, regardless of major. For the latter contextual factors, the researchers found that gender became a significant predictor for persistence in computer science and engineering.

The work is significant in that it is based on previously known demographic, contextual, and performance information without administering a separate specialized test or exam and illustrates that predictive power is possible from pre-existing data. The finding regarding gender's predictive power may be surprising at the time of the original study, but over time the role of gender and success in computer science has become better understood (see following discussions) and its predictive power is now not as significant as it was at the time of this study.

About the same time as the Campbell and McCabe study, Butcher and Muth (1985) focused on pre-existing high school data and ACT scores.<sup>2</sup> The study differed from earlier work in that the purpose was to determine qualification of the student for entry into the computer science major as opposed to persistence in the major. The study came about because the institution where it took place had been using a four-semester series of qualification courses (pre-computer science) that had to be passed successfully prior to official entry as a computer science major. The research question at the heart of their inquiry was whether high school performance and ACT test data could successfully predict completion of the four-course qualification regime that controls entry into the major.

Independent variables for their work included ACT-related scores, including mathematics, English, natural science, social science, and a calculated composite score based on the other scores. The researchers also gleaned high school data from transcripts that included students' high school class rank, class size, grade point average, level of mathematics taken, programming courses taken, number of math and science courses completed, and a computed student percentile rank. The dependent variables and measure of merit for predictive success in the analysis was to match the outcomes from the sample's examination average, laboratory average, and final course grade in the first semester introductory computer science course. Additionally, the researchers used the student's first semester grade point average to measure overall success in college.

---

<sup>2</sup> For the purpose of this study, ACT and SAT scores will be assumed to be interchangeable as are the terms Grade Point Average (GPA) and Grade Point Ratio (GPR).

The analytical results led the researchers to conclude that any two of the ACT composite score, the ACT math score, and HS-GPA provided a significant fit against all four dependent variables (final course grade, examination grade, lab examination grades, and college GPA).<sup>3</sup> On additional analysis, they determined it explained 36.6 percent of the variance in the dependent final grade variable when used as a singular outcome. The researchers were surprised to find that neither success in computer science nor college GPA were influenced by previous exposure to computer science, at the time a relationship that was expected to be significant.

The researchers also noted that using the model for admission decisions must be done with care as admission decisions are usually done prior to the student completing high school, thus impacting the availability of HS-GPA data. The lack of these data in turn weakens the fit of the HS-GPA and ACT-Math combination as best predictors of both computer science and first semester college performance. In that light, they concluded that although ACT-Math and ACT-Composite as predictor variables were “acceptable,” the combination was not as powerful as when HS GPA was included in the model. Thus, the researchers concluded that it was possible to predict performance in the introductory course based only on ACT scores and high school transcript information.

This realization was one of the very few acknowledgements of problems that would be encountered when converting an analytical model to a predictive model in an operating environment. Specifically at issue is the data mining required to obtain the needed independent variable data as not all institutions have such information easily retrievable (i.e., the data act as a “trailing” indicator and are not available at the time

---

<sup>3</sup> Acknowledging perhaps the danger of multi-colinearity associated with using all of the ACT factors.



needed), a factor that impacts utility for advising as opposed to admission purposes. Second, it is encouraging that the study does not rely on data that require specialized testing and it is data that could be available at many institutions. Finally, even though significant, their model left over 64 percent of the variance in the qualifying course grade unexplained.

Evans and Simkin (1989) sought to consolidate the multiple models available at the time of their inquiry. They purposefully sought to define explanatory variables, using student performance in an introductory course in business computing as the measure of "proficiency." The researchers also employed a 100-item questionnaire administered to the same group of students.

The model employed by the researchers is of interest. Building on the literature base available at the time of their work, they created a large set of 34 independent variables grouped into general categories of demographic, academic, prior computer training and/or experience, and behavioral factors. As a result, the model included virtually all the factors (or proxies) identified by prior research studies at the time of their work. In effect, the researchers intended to consolidate the body of literature in their model and at the same time allow for a new emphasis on the possible impact of cognitive processes on student achievement.

To accommodate the cognitive portion of their model, the researchers used an abbreviated Myers-Briggs type indicator coupled with a specific problem-solving test within the survey instrument. When the results of the problem-solving test are included in the model, it grows to 49 independent variables. The model uses six dependent variables, all associated with performance benchmarks within the introductory course used as

the performance measure of merit. These include grades received on homework, midterm and final exams and grades received on two programming assignments.

The researchers declared three important outcomes. First, with such a large model, the associated  $R^2$  values (no more than 24%) reflected a low explanatory power, although not out of line with previous studies on which they based their work. Second, they concluded that very few of the demographics were “particularly strong” predictors of performance. Third, they held that cognitive factors emerged as important “explanatory” variables in several of their alternative models (p. 1326).

Goold and Rimmer (2000) took a cohort approach to their search for predictive factors, the first study found in the literature to do so. They took this approach in an attempt to accommodate what they saw in other study approaches as uncontrolled variables, such as types of students, instructors, courses, and learning environments endemic to analysis of student groups located in different settings. Their rationale maintained that the cohort approach provides consistency of environment and experience across time, which controls for the variability they argue was introduced in other studies.

Thus, their sample consisted of a group of students over a two-semester experience in introductory computer science at an Australian university. The first term consisted of two possible courses: an introductory IT (Information Technology) course where applications software was emphasized and a Basic Programming Concepts course taught in the C programming language. In the second term, the cohort came together in a data structures course. The researchers analyzed seven independent indicators: average score in other classes, relative abstraction, did programming before entering university, dislike of programming, problem solving, and gender. These indicators were evaluated for the cohort against performance in each of the two-semesters in the first year of their

computer science curriculum. The instrument was given to the students during the second half of the data structures course.

Goold and Rimmer's analysis shows different groupings of the factors based on the course (IT or Programming Concepts) taken in the first term by the students. They found that their models provided explanatory power of between 42 and 65 percent of the variability with the highest  $R^2$  applicable to the introductory IT course. The only consistent factor across all models was Average Score Across Other Units (i.e., previous college coursework outcomes). By their analysis, the researchers concluded that, over time, performance of computer science students "come to conform to other undergraduate grades" (p. 42); that is, performance in computer science tended to reflect performance in other classes. They also noted that other factors changed in importance to the model between the two terms. For example, dislike of programming had little effect in the first term courses, but showed a larger impact in the second, while at the same time the importance of problem-solving had diminished to being non-significant in the second course. Likewise, the impact of gender was apparent in the first term but not in the second. Overall the researchers concluded that the model's factors were dynamic; that is, changing over time and environment.

The implication of Goold and Rimmer's study is that the measure of merit, i.e., course performance against which a model is applied, may be changed by the course's temporal position in the curriculum. Phrased differently, a model developed for prediction of student performance in the very first course, may not be significant when applied to a subsequent course.

Wilson and Shrock (2001) built their model on some of the demographic, contextual, and performance traits already identified in the literature; however, they took a

new tack in that they added a heavier behavioral component. One unique aspect of this model was using attribution theory as a basis for identifying persistence in the endeavor. The researchers included self-attribution for success or failure in their model on the rationale that “. . . theory suggests that when people attribute their successes to unstable causes (luck or effort) and their failure to stable causes (ability or task difficulty), the probability of persistence is low” (p. 184). In addition to attribution, their model also included other motivational factors such as encouragement to study computer science and assessment of the subject's “comfort level” in the discipline.

Their 12-factor model includes the following independent variables: gender; previous non-programming experience; previous programming experience; math background; encouragement to pursue computer science; comfort level; work style preference; attributions which were sub-categorized as ability, task, luck, and effort; and, self-efficacy. The dependent variable was the student's performance on the midterm exam of an introductory computer science course. This variable was chosen in an attempt to measure the independent variable's impact prior to students dropping the course after the mid-term.

Interestingly, results showed that comfort level was the most significant factor in predicting success in the index course and math background was the second most important factor. A third factor, attribution to luck, was also found significant. Using stepwise regressions, the researchers created a five-factor prediction model based on adding work style preference and attribution to task difficulty to the previous three variables. The two models (three-factor and five-factor) demonstrated  $R^2$  values of 0.44 and 0.40, respectively.

While the significant factors were interesting, the researchers also concluded that what was not significant was also interesting, particularly the absence of prior programming experience in the full model, which tended to fly in the face of previous studies.<sup>4</sup> Further, they concluded that a teaching style which inculcated a “comfortable” environment as far as the course content was concerned was something that should be brought to the attention of faculty.

LeJune made a unique contribution to the literature with a qualitative rather than quantitative study of success factors (LeJeune, 2000). LeJune relied on interview data to supplement transcript and course performance data for a small n-group (4). The researcher found consistency in the data for four factors: encountering a disrupting life event outside of the academic environment; a general lack of motivation; inadequate academic preparation; and, the course content being outside of the student’s zone of proximal development, in other words, a sense of being totally lost in the course. LeJune’s work suggests that other factors may be at work in an “unsuccessful” outcome that may or may not have any relation to course performance and may represent a perfectly logical outcome based on the student’s situation.

#### Testing Specific Skills or Traits

The explanatory literature also contains a limited body of work in which the researcher was intent on testing the explanatory power of specific demographic, contextual, or performance characteristics on student success. As noted earlier, this does not suggest that the researchers disregarded other factors, but rather was intent on subjecting

---

<sup>4</sup> An analysis focusing only on prior experience factors showed that prior computer courses were significant as was game playing experience. In this case, however, the former was positively correlated to performance on the midterm while game playing was negatively correlated (p. 187).

selected demographic or contextual traits (Figure 1) to analysis. The rationale of why the researcher selected the demographic or contextual element is not of particular interest here but the process of focusing on one specific trait or factor is, as it relates to a similar focus on the relationship between the CMPT (intermediate performance variable) and student success in this study.

### Reasoning Skill

Kurtz (1980) argued that variables based on past performance both in high school and college had not been (at that time) useful predictors and suggested that a better approach would be to base student performance prediction on the intellectual development and reasoning skill of the student. Accordingly, he developed a specialized instrument and administered it to introductory computer science students at a university in the western region of the United States. The approach was to use the testing instrument to classify students by developmental levels, and, in turn, use those levels as predictors of performance in an introductory computer science course.

Kurtz found that intellectual development was a predictor of performance (80% explanation of variation) and particularly for performance on tests administered during the course. He concluded that the specialized instrument was useful in identifying the exceptional student, as well as the student who needs to approach the course at a slower pace. In other words, he was able to identify the extremes of the continuum, but the model was not able to differentiate levels of performance. Kurtz also noted that his work was hampered by a small sample (23 students) and suggested that a larger follow-on effort be undertaken.

Barker and Unger (1983) responded to Kurtz's challenge and modified his instrument to shorten the time it took students to complete the intellectual development instrument, and administered it to a larger sample ( $n = 353$  students instead of 23). Barker and Unger also included self-reported data which included prior programming experience and courses, GPA, and "enjoyment" of programming. Their purpose was to develop a screening instrument that could be administered to entering students and the results used for class placement. The instrument was administered after the students had begun work in the introductory class.

The researchers found that when the modified Kurtz instrument was administered to a larger population across multiple sections of students the predictive value of the instrument dropped sharply, exhibiting an  $R^2$  of 0.12 as opposed to Kurtz's reported 0.80. In spite of the differences in their work and that of Kurtz, they believed the instrument, when used in conjunction with the other self-reported data, could be useful for student placement and of classifying outcomes into dichotomous groups.

### Gender

The question of gender difference relating to success in computer science has been raised both in the context of careers and persistence in industry (Ahuja, 1995; DeClue, 1997; Natale, 2002; Teague, 2002) and in terms of success for students of computer science in higher education (Beyer, Rynes, Perrault, Hay, & Haller, 2003; Wilson, 2000). Of interest is the latter group, not only for the student success context, but also the construction of models to test the inferences of the researchers regarding student success and gender.

Wilson (2000) posited a 12-factor model that included not only gender but additional contextual factors. These included math background, previous programming experience, previous non-programming computer experience, attribution for success/failure, self-efficacy, encouragement, comfort level, and work style preference. Wilson's exploration had the explicit purpose of evaluating gender as a contextual variable contributing to a student's success or failure in an introductory computer course (p. i). She found no significant relationship regarding gender, but did find that comfort level, math background and attribution for causes of success or failure were significant. This work clearly was the basis for the 12-factor model developed and published a year later (Wilson & Shrock, 2001) and cited earlier in this review. In that later work, she de-emphasized gender and focused on the other factors as being more important to the explanation of student success or failure.

Beyer et al. (2003) took a slightly different approach and evaluated gender factors for computer science majors and non-computer science majors in a multivariate analysis. Their model consisted of 11 factor groups (reflecting 19 specific factors): demographic, ability in quantitative areas, educational goals and interests, experience with computers, stereotypes and knowledge about computer science, confidence, personality, support and encouragement, stress and financial issues, gender discrimination, attitudes toward computer science courses and instructors. Specifically, the researchers sought to compare the results between the female majors and non-majors, seeking factors with explanatory value as opposed to predictive value. The focused purpose of the research was to specifically test gender as a root cause of the "dearth" of women majoring in computer science.

From their analysis, Beyer et al. concluded that there were no gender differences between majors and non-majors regarding factors of demography, ability in quantitative



areas, stereotypes and knowledge about computer science, gender discrimination, educational goals and interests, or attitudes toward computer science courses and instructors. Differences were noted between majors and non-majors regarding knowledge about exactly what “computer science” is, reflecting a general ignorance among women of career opportunities and quality of life in the profession, which acted as a deterrent to women entering the field. Additionally, factors that reflected a student’s confidence and support structure were also an important deterrent to entry to computer science.

### Predictive Studies

A final grouping in this literature review is a small body of work that seeks to create models that predict student outcomes in computer science programs. Two studies are of particular interest because they highlight characteristics of the problem addressed by this current work.

Glorfeld and Fowler (1982) sought to validate a model they had previously developed (Fowler & Glorfeld, 1981) that specified a predictive model for classifying student aptitude in introductory computing. The original model was based on three categories of data: personal, academic, and aptitude. Within the categories were the following factors: personal – gender, race, age, veteran status, marital status; academic – major, classification, number of math courses completed, GPA (at the college level); aptitude - SAT math, SAT verbal, and a score from a programming aptitude test. Evaluation of the model yielded significant factors of age, SAT math, number of math courses completed, and GPA, and yielded a  $R^2$  value of 0.808.

The later validation effort (1982) was based on a new student sample using the same selection methodology as the earlier (1981) work. The “target” for the model was

to correctly predict student performance into “high” aptitude (a grade of A or B) or “low” aptitude (a grade of C, D, or F) for the introductory computing course. To validate their model, the researchers used a logistic discriminant analysis (more commonly called “logistic regression”) as the predictive instrument to produce the classification. The researchers also modified the original model significantly to focus on four discrete continuous variables – age, SAT math score, number of math courses taken, and GPA.

Their validation testing showed that age was not a significant predictor in the model, affirming earlier findings that age was of marginal value to a predictive model. GPA, on the other hand, was the most important factor in the model, as they phrased it “. . . the best indicator of a student’s future performance is their past academic performance” (p. 143). The model did correctly classify students into “high” or “low” aptitude categories approximately 75 percent of the time. The researchers also noted that their validation model required a constant flow of data to keep it current “. . . as a source of additional input in counseling students,” and went on to remark that “It would be interesting to compare models developed at various universities to see if the predictive performance of the model and variables included in the model were similar” (p. 143).

Chowdhury et al. (1987) also used a logistic regression process to predict student outcomes in a beginning computer course. In this instance the dichotomized dependent variable indicating success was grouping course grades into two categories, “acceptable” (A, B, C) and “not acceptable” (D, F, or Withdrawn). The independent variables included outcome of an introductory calculus course, SAT verbal and math scores, high school class rank, gender, term when the computing course taken, and instructor of the course. Using a logistic regression technique, the researchers found that the “best” predictor (p.

449) of computing success was previous success in a calculus course and the second best predictor was the high school class rank.

### Summary of Prior Modeling

To summarize the literature to this point and to relate it to the problem at hand, a matrix of variables derived from the literature was constructed (Table 1). The matrix captures the independent variables used in the models described by the literature base and groups them into five general categories: demographic, academic experience, specialized experience, behavior, and specialized tests.

With this structure it is easier to see the multiplicity of factors deemed by the literature to have potential to explain student performance, and the almost unlimited number of factors that could be brought to bear as independent variables. Moreover, it becomes clearer that acquiring the information to flesh out the variables is a major undertaking.

### Significant Factors

But simply consolidating the possible factors is not enough and ignores the reality of the situation when constraints are placed on the choice of variables during replication, such as availability of existing information or what it would take to gather new information to support the variables. Further, the literature suggests that large models are not always informative (e.g., Evans and Simkin, 1989). To address this, several consolidating actions are performed to the initial matrix. First, the matrix is reduced to only the factors found in the literature to be significant in explaining variance in the dependent variables; second, the one qualitative study is removed; and third, thematically linked

Table 1. Model Matrix

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavior	Specialized Test
Barker (1983)		GPA	previous programming experience; previous programming classes	enjoyment of programming	intellectual development*
Beyer (2003)	demographic		ability in quantitative areas; educational goals and interests; experience with computers; stereotypes and knowledge about computer science*	personality ; confidence; support and encouragement*; gender discrimination; stress and financial issues; attitudes toward computer science courses and instructors	
Butcher (1985)		HS class rank; HS class size; HS grade point average*; level of HS math; HS programming courses; number of math, science, and computer courses; HS class percentile rank; ACT-Math* ACT-English; ACT-Natural; Science; ACT-Social Science; ACT-Composite*			

Table 1. Model Matrix (Continued)

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavior	Specialized Test
Campbell (1984)	gender*	HS graduating rank*; HS graduating class size; HS math semesters*; HS science semesters*; HS English semesters; average HS math grades*; average HS science grades*; average HS English grades; SAT-Math*; SAT-Verbal*			
Chowdhury (1987)	gender	SAT math; SAT verbal; HS class rank*; term of course; instructor of course; class status; success in beginning calculus course*			
Evans (1989)	age*; gender*; race, parents graduate?; present living situation; place of birth; citizenship; first language; second language*; mother's occupation*; father's occupation	declared or preferred major; HS GPA*; SAT Verbal; SAT Math; number of HS science courses*; number of HS math courses; class in logic or problem-solving?; typing ability*	own computer?*; prior experience; amount of formal computer training; programming in – LOGO; BASIC*; COBOL; FORTRAN; PASCAL	hours watching TV now; hours watching TV growing up; hours spent playing video or computer games*; hours worked at outside job*; type of work last summer; type of job wanted after graduation	Abbreviated Myers-Briggs*; Problem-Solving Test*
Fowler (1981) Glorfeld (1982)	gender; race; age*; veteran status; marital status	SAT math*; SAT verbal; major; classification; number of math courses completed*; (college) GPA*			Wolfe programming aptitude test

Table 1. Model Matrix (Continued)

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavior	Specialized Test
Goold (2000)	gender*	did programming before university; raw secondary scores*; average scores in other (college) units (classes)*		dislike programming*	relative abstraction*; problem solving*
Konvalina (1983)	age*	high school performance*; prior computing education*; years of HS math; number of college math courses*; total number of HS and college math courses*	hours worked per week; prior non-programming computer work; prior programming work		computer science placement exam "predictor" (sequence & logic; calculation; algorithms; word problems)
Kurtz (1980)					intellectual development*
LeJune (2000) [qualitative study]		academic preparation		motivation; outside life events	zone of proximal development (understanding)
Wilson (2000)	gender	math background*;	previous programming experience; previous non-programming computer experience	attribution for success/failure*; self-efficacy; encouragement; comfort level*; work style preference	
Wilson (2001)	Gender	math background*;	previous programming experience; previous non-programming computer experience;	encouragement to pursue computer science; comfort level*; work style preference*; attributions: ability, task*, luck*, effort; self-efficacy	

\*Significant factor in the model.

studies (i.e., Kurtz – Barker and Wilson – Beyer) are combined. The resulting new matrix (Table 2) helps clarify the scope of possible variables and brings the factors into better focus.

With this perspective, the table reduces to a 10 by 5 matrix and the dominance of Academic Experience factors becomes evident, being present in every model. The other categories are present only in one-half of the models. It also becomes clear that the most often expressed measures of Academic Experience are SAT/ACT scores and/or accomplished performance in selected academic environments such as math courses or college level GPA. Regarding the other factors, age and gender are the most common demographic factors. In the context of the literature, the factors contained in the remaining three categories suggest more of a “one-of-a-kind” set of factors important for that particular study.

#### Data Availability

There is more to operationalizing a model in an advising context and that revolves around the actual availability of data to support predictive models and whether the data available actually can be collected, processed, and be available for use prior to the student’s arrival for predicting, classifying, or advising. For most of the literature cited, data availability was not a key issue of study as the researchers sought to define (as opposed to use) explanatory variables in light of the situation they were testing.

However, there were exceptions. For example, Campbell (1984) attempted to limit data requirements to that available from a registrar’s office. Even so, realistically one can expect wide variation as to what exactly “available” might mean, depending on the institution where replication is attempted. The significant predictive variables for

Table 2. Model Matrix – Significant Factors

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavioral	Specialized Test
Barker (1983) and Kurtz (1980)					intellectual development*
Beyer (2003) and Wilson (2000)		math background*	stereotypes and knowledge about computer science*	attribution for success/failure*; comfort level*; support and encouragement*	
Butcher (1985)		HS class rank; HS class size; HS grade point average*; ACT-Math* ACT-Composite*			
Campbell (1984)	gender*	HS graduating rank*; HS math semesters*; HS science semesters*; average HS math grades*; average HS science grades*; SAT-Math*; SAT-Verbal*			
Chowdhury (1987)		*success in beginning calculus course*; HS class rank*			
Evans (1989)	age*; gender*; second language*; mother's occupation*	HS GPA*; number of HS science courses*; typing ability*	own computer?*; programming in – BASIC*	hours spent playing video or computer games*; hours worked at outside job*	abbreviated Myers-Briggs*; problem-solving test*
Fowler (1981); Glorfeld (1982)	Age*	SAT math*; number of math courses completed*; (college) GPA*			



Table 2. Model Matrix – Significant Factors (Continued)

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavioral	Specialized Test
Goold (2000)	gender*	raw secondary scores*; average scores in other (college) units (classes)*		dislike programming*	relative abstraction*; problem solving*
Konvalina (1983)	Age*	high school performance*; prior computing education*; number of college math courses*; total number of HS and college math courses*			
Wilson (2001)		math background*		comfort level*; work style preference*; attributions: task*; luck*	

\*Significant factor in the model.

their model were found to be SAT scores (usually available in an advising context), the student's background in high school math and science (usually not so readily available in an advising context), and gender (readily available). Depending on the specific institution, to operationalize a similar model processes or procedures have to be in place (or developed and implemented if not already existing) to have the data available for analysis prior to the student's arrival for advising. Accomplishing these changes may not be as easy as it seems for the time prior to matriculation is often the most uncertain and unstable as potential students attempt to complete their application packages and data are constantly flowing into the admissions data system. Having information available when needed to perform the predictive analysis boils down to a question of timing. This component was not addressed. Campbell's (1984) model tried to deal with information available for analytical purposes, i.e., to developing the model, but not necessarily attempting to use it in an operating or advising environment.

Likewise, Butcher and Muth (1988) used ACT test data and high school performance data in constructing their model and, in one sense, succeeded in improving the likelihood of having data available compared to Campbell (1984). However, their factors were manipulated beyond the "normally" reported types of scores. The significant predictors from their work were high school class rank, class size, HS grade point average, ACT-Math, and ACT-Composite (calculated from other ACT scores).

As discussed above and in Chapter I, data availability is a key factor in attempting to predict student success. When Table 2 is re-evaluated in that light, it collapses even more (Table 3). The criteria for collapsing the factors as to their "availability" are driven largely by the situation in the current analysis and may not be totally representative of all institutions, but nonetheless seems to be instructive.

Table 3. Model Matrix – Significant and “Available” Factors

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavioral	Specialized Test
Barker (1983) and Kurtz (1980)					
Beyer (2003) and Wilson (2000)					
Butcher (1985)		ACT-Math*			
Campbell (1984)	gender*	SAT-Math*; SAT-Verbal*			
Chowdhury (1987)		*success in beginning calculus course*; HS class rank*			
Evans (1989)	gender*				
Fowler (1981); Glorfeld (1982)		SAT math*; (college) GPA*			
Goold (2000)	gender*	average scores in other (college) units (classes)*			
Konvalina (1983)		number of college math courses*			
Wilson (2001)					

\*Significant factor in the model.

Of interest is the vacating of Specialized Experience, Behavioral, and Specialized Test factors completely from the matrix, and only gender being present in the Demographic category. Of the remaining factors in the category of Academic Experience, operationalizing them into a form that can be used for predictive purposes is a question that still remains.

### Leading Indicators

Another way to evaluate the modeling matrix (Tables 1-3) is to evaluate the factors on the basis of which are leading indicators. A “leading” indicator is defined as a factor that represents information suggestive of predictive merit, and is available for use before the student comes to college. In Table 1, seven of the 13 models used specialized data collection or testing of students once they are already in college – not particularly helpful for providing information before the student enters the institution or program. In Table 2, three of the 10 models had significant results for the specialized tests used. This is not to say that specialized testing is not appropriate for building predictive models, but it should be approached from the perspective that such derived factors will be (in most cases) institutional dependent, and not readily replicable across higher education institutions.

Of the other factors in the Academic Experience category, three models (Konvalina, Goold, and Chowdhury) are based on student performance after arriving at college (i.e., trailing indicators). In the current instance, factors based on a student’s performance once entering college is of little utility when the student base consists of students who have not taken college work yet. When Table 3 is again collapsed based on whether or not the factor is pre- or post-college entry, Table 4 results.

Table 4. Model Matrix – Significant, “Available,” and Leading Factors

Factor/Study	Demographic	Academic Experience	Specialized Experience	Behavioral	Specialized Test
Butcher (1985)		ACT-Math* **			
Campbell (1984)	gender*	SAT-Math*; SAT-Verbal*			
Evans (1989)	gender*				
Fowler (1981); Glorfeld (1982)		SAT math*			
Goold (2000)	gender*				

\* Significant factor in the model.

\*\* SAT data are assumed exchangeable for ACT data and vice versa.

This analysis suggests a starting model consisting only of gender, SAT math and verbal scores. Upon reflection, the number of factors has been significantly reduced, and begs the question of whether they have been reduced to the point where their utility as predictive factors is usable in any predictive context.

### Operationalizing a Model

At the risk of being trite, it is one thing to intellectually “explain” what factors contribute to the success of a student in computer science; it is an entirely different matter to take that explanation into an “operational” environment and attempt to explain (or predict) performance. This happens in the more “normal” situation when not all the information deemed significant in explanatory models is available, or in useful form to present to the prospective student (or parents for that matter). It is appropriate then to visit the inventory of information available to the current study that might support predictive factors within the context of the literature analysis.

At this institution, data are collected from student application data upon which the admission decision is based. These data included SAT/ACT test results and self-reported (via the high school counselor) high school performance data (GPA, Class Rank, and Class Size). Official high school transcript data are not collected. Of these data, identification, demographic and selected academic information (HS class rank, class size and SAT/ACT scores) are then transmitted informally to the department of the declared major once the student has been accepted for admission. The data “trickle” in during the course of the admissions “season” (perhaps up to one year before matriculation), usually in groups of three to four at a time for computer science. There is no “cutoff” point at which time a department can ensure that such a grouping represents the “class” that will present

itself for advising. In fact, of the data received, not all will implement a decision to attend this institution and across the institution about one-third of those accepted actually enroll (Clemson University, 2004), nor is there anything to prevent (as often happens) students being admitted and present for advising before any admissions data are received by the department.

Adding to the issue, the historical data base maintained at this institution seems to diverge from the concurrent reporting during the admissions process in that self-reported high school GPA, rank and class size are not maintained in the central data base available to advisors. The practical effect is that some high school information is perishable, resulting in potential predictor information being lost. In the current situation, data available to support a revised model based on previous high school performance and universal (SAT) testing are limited to the SAT Math and Verbal test scores and an institutionally generated prediction of potential GPA based on SAT/ACT test results.

Taken as a whole and again in context of the current study, the review of the literature seems to indicate that there is no “ideal” set of data that will predict student success in computer science courses. The review also suggests that individual situations regarding data availability will have a significant impact, accompanied by choices regarding the use of specialized testing (either for general tendencies; i.e., Evans, Kurtz, and Barker) or testing for special skills or inclinations (i.e., Konvalina, Evans, Goold, LeJune, and Glorfeld).

This is exactly the situation in the current study where each entering student is required by the institution to take a math placement test (Clemson Math Placement Test – CMPT) to determine the first math class that will be taken by the entering student. The Mathematical Sciences Department reports the results of this test to the student and is

retrievable by other departments (such as Computer Science) upon request. As discussed in Chapter I, these results are used to place incoming students in the appropriate computer science class as a certain minimum score is required (a “4” on a scale of 1-6) as a co-requisite for entry into CPSC 101, the introductory course of the computer science curriculum. Lacking the minimum score, the student is advised not to attempt CPSC 101, but a course in problem-solving and programming logic is offered (CPSC 104) while the students hone their precalculus skills through an appropriate math course other than MTHSC 106.

Bringing all of the preceding together and collapsing again the previous matrices, a model for testing and predicting student performance in an introductory computing course at this institution takes the form presented in Table 5. Race is added to the model since the information is available, and given the impact in the literature of race on specific situations in various models, it does no harm to include the factor in at least the initial model.

Table 5. Model Matrix – Final

Demographic	Academic Experience	Specialized Experience	Behavioral	Specialized Test
gender; race	SAT-Math; SAT-Verbal; SAT-PGPR			CMPT



### Revisiting the Conceptual Construct of the Literature

At this point it seems appropriate to pause and reflect again on the literature as it influenced the evolution of the initial model presented above. As noted at the beginning of this chapter, the literature falls into two main groups, explanatory and predictive, and can be visualized using a diagram of variables developed by Flanigan, Marion, and Richardson (1996) (Figure 1), specifically the three levels of independent variables: contextual, demographic, and main or “performance” variables.

All of the variables from the literature (Table 1) fall into either contextual, demographic, or performance variables, while seeking to explain or predict the outcome achievement in computer science. When the literature is evaluated and the potential variables are collapsed into an initial working model (Table 5) for this study, it is interesting that the contextual variables fall away leaving only demographic (gender, race) and intermediate performance factors (SAT-Math, SAT-Verbal, SAT-PGPR, and CMPT scores) as independent variables.

The current study occurs in an operational environment (with the concurrent lack of some data), where, in fact, the student’s context (from a college perspective) has yet to be developed. How much that will impact any predictive power of the model is indeed at the heart of the current study.

### Revisiting the Problem

It is also appropriate at this point to revisit the problem statement from Chapter I. While the predictive model described previously is a central structure for this inquiry, it is not the complete purpose of the study. As developed in the preceding section, the literature suggests that the predictive model is but one step in a line of inquiry that has other

important questions to be answered, including the “logistical” implications of gathering data on a scale needed to support operational use in a student advising environment and how well any predictive model might perform with such constraints. The research questions from Chapter I are listed below and additive comments based on findings from the literature review are added in *italics*.

- RQ1: Is there a relationship between CMPT scores and CPSC 101 student outcomes?

While the CMPT has been validated for use with predicting potential outcomes for calculus and math courses, it has not been subject to scrutiny as to its relationship to computer science outcomes.

- RQ2: If there is a relationship, is it significant?

Is any relationship between the CMPT and computer science student outcomes statistically significant?

- RQ3: If significant, how well does the CMPT predict student performance at each level?

Does the CMPT actually have predictive power for student success in entry level computer science courses? *This suggests evaluating the model at one level with only the CMPT score as the independent variable. This represents the “lowest cost” option since the data to support a one factor model are available and usable within the advising context.*

- RQ4: What other factors, in addition to the CMPT should be in the predictive model?

Are other student data available that would strengthen the predictive model? *Other data are available, but limited to gender, race, SAT scores (math and verbal) and an institutionally generated prediction of potential GPR once matriculated. This represents a slightly “higher cost” option as the data to be added must come from separate sources and processes to acquire that information which may not be available.*

- RQ5: How well do they (RQ4) predict performance?

What is the optimal predictive model of student success in entry-level computer science courses at this institution? Should other student data be considered and are they significant to the model? *For the purposes of this study, the initial model should include the other factors (i.e., gender, Race, SAT-Math., SAT-Verbal, SAT-PGPR, as well as the CMPT score) until such time they are deemed to be non-significant, or not available in an operational environment for advising.*

- RQ6: Does the model have predictive power for other computer science course in the curriculum sequence? *Analysis of outcomes from the last course in the introductory sequence (CPSC 212) will be analyzed.*
- RQ7: *What is the impact of using a predictive model in an “operational” environment, i.e., is there a “best mix” of variables that has predictive power, yet at the same time can be used within a framework of constrained data such as found in an actual advising situation?*

*Even if the model has predictive power, the data to support the model must be available at “reasonable cost” (in terms of improving advising), and in a temporal framework that improves the existing situation.*

The added research question (RQ7) reflects an over-arching question that must be considered as an important construct for the current analysis. The literature, while for the most part silent concerning them, suggests that there are issues and problems associated with using predictive models as opposed to explaining those variables or factors that might be useful for predictive purposes. Restated, a critical element in operationalizing a predictive model is that it must operate in an environment that has constraints placed upon it. It is one thing to explain what factors contribute to (or even determine) success in an endeavor from a purely academic perspective, and quite another to create an operational environment where such information is routinely available, processed, and delivered to the appropriate place or actors, and at an appropriate time in the matriculation cycle to be of use to both the institution and student. Evaluating the transition from the explanatory to the operational (predictive) will highlight problems of such a change

involving availability of data to support an operational implementation, and from a resource perspective, what processes must be changed, developed and implemented to make any outcome useful. This study attempts to highlight those constraints and helps to define the issues associated with making a decision on whether such expenditure of resources is warranted, and whether the effort adds value to the advising process of students over what is currently in use.

## CHAPTER III

### METHODOLOGY

#### Introduction

This chapter deals with the sources of data, a description of the data and its limitations, an explanation of the model to be evaluated, and finally, a description of the analysis and procedures used in the analytical process. The analytical design uses multiple and logistic regression processes, the former to define a predictive model for student achievement in an introductory computer science course and the latter to test the efficacy of the model. One goal of this study is to explore the process and difficulties associated with implementing a predictive model of student success. Accordingly the methodology not only includes the empirical evaluation of data and its predictive power, but also relates that data to the process of moving from an analytical framework to an operational environment. As such, some recognition of these additional issues is contained within the methodological discussion.

This study focuses on the achievement of students who take an introductory course in computer science (CPSC 101) at a major southern university in the United States. The population for analysis is 1,014 students who took CPSC 101 during the fall term of 1996 through the spring term of 2004 and who received a final grade in that course.

The population of students who took CPSC 212 (the final course in the introductory sequence of courses – see Chapter I and discussion following) was also analyzed, representing 349 student grades for the course.

An additional word is appropriate here about the primary independent variable under scrutiny, the Clemson Math Placement Test (CMPT). The stated purpose of the exam is to function as a “diagnostic tool for basic algebra and college mathematics skills” (Department of Mathematical Sciences, 2004c). As such, the mathematical sciences department has relied on it heavily during the time since its institution-wide implementation in 2001. The score that the student makes, in relation to the first required math course of their specific curriculum, determines whether the student actually takes that course, or is referred first to a preparatory course. As the department states on its descriptive Web page:

You don't pass or fail the CMPT -- it measures your preparation in mathematics, not your innate ability or intelligence. Your CMPT score is used to determine the appropriate initial mathematics course for you at Clemson. It is intended to improve your chances of successfully completing your mathematics courses at Clemson University” (Department of Mathematical Sciences, 2004e).

Results of the CMPT are enforced, and as advisors are cautioned: “Students not having the requisite CMPT score will be dropped from their math class unless they present evidence on the first day of classes of having satisfactorily completed the required preparatory course or having AP or IB credit” (Department of Mathematical Sciences, 2004d). Sample questions from the CMPT are included as Appendix A.

### Data Sources

The model (Table 5) developed in Chapter II requires three categories of data: demographic information for gender and race, contextual information about previous academic experience represented by SAT scores and the institutionally calculated predicted Grade Point Ratio (GPR), and specialized test information in the form of scores from the Clemson Math Placement Exam (CMPT). A fourth category of outcome data is needed

in the form of grade data (dependent variable) for the index courses against which the predictive nature of the model will be evaluated. A summary of the data required by the model and their sources is presented in Table 6.

Table 6. Data Sources

Data Source	Demographic	Context: Academic Experience	Specialized Test	Outcome: Grade Data
Admissions Office Data Base	Gender; Race	SAT-Math; SAT-Verbal; Predicted GPR		
Mathematical Sciences Department Data Base			CMPT Score	
Registrar's Student Data Base				CPSC 101; CPSC 212

One of the objectives of this study is to determine the process involved in implementing predictive measures in an "operating" environment, in the instant case for the purposes of using the model's outcome for advising newly matriculated students of computer science. Table 6 clearly shows that the operative components of the model (all but the grade data) must come from disparate sources and manually linked for analysis. Of these data sets, only the grade data (contained in 21 separate data tables) are directly available to the researcher, the other categories of data require the assistance of the office of record that "owns" and maintains the data.

#### Description and Limitations of the Data

Outcome data are grades earned by students in the course(s) of interest. The Student Data Base maintained by the Registrar's Office contains the grades earned by every

student in the institution. The data consist of many separate items and include the following components of direct interest to this study: student identification number (unique to each student), the student's major at the time the course was taken; course identification information, the academic term when the course was taken, and the grade earned in the course. These data are extractable directly from the central data base in the form of Microsoft Access<sup>®</sup> data base files.

The institution's Admissions Office maintains two categories of data (Demographic and contextual Academic Experience) required by the model. These include: student identification number, gender, race, SAT Math and Verbal test scores, the predicted GPR (calculated from SAT scores and other information from the student's application for admission), and the academic term indicated by the student for arrival at the institution. These data are available as a text file after being extracted from the data base by the Admissions Office staff.

Lastly, the model requires the student's score on the math placement test administered and maintained by the mathematical sciences department. The test is administered via the World Wide Web and data are input to the department's Web server. Data are extracted by the department and provided in a text format including student identification number, a timed test score, a total test score, and the date the exam was taken by the student.

The various data sets (a total of 23 separate data files) were combined by indexing on the student's identification number and output into a combined file suitable for export to SPSS<sup>®</sup> data analysis software.



### Population and Sample

As presented in Chapter I, the introductory sequence for students of computer science at this institution begins with CPSC 101, an introductory course in the JAVA programming language within the WINDOWS<sup>®</sup> operating environment, followed by a second course of the JAVA language, this time within the UNIX<sup>®</sup> environment. A course in algorithms and data structures (CPSC 212) is the third course in the sequence and taken together, these three courses represent the “gateway” sequence, that is, the base set of skills and knowledge a student should master if they intend to continue in the major. Although there is no formal process of eliminating students from the major based on performance in these courses, it is departmental policy that a student must pass any curriculum course with a grade of “C” or better to progress to the next required course.

The population under study for this work consists of students who have taken CPSC 101 and CPSC 212. As noted above, CPSC 101 is the initial course in the “gateway” sequence in computer science and CPSC 212 is the last course. Of primary interest to the study are the data associated with CPSC 101, but CPSC 212 data are collected also to perform an alternative analysis in response to Research Question 6 presented in Chapter I.

In its final form, the data set for CPSC 101 contained a total of 1,014 valid grades earned by students during the period from fall 1996 through spring 2004. Within that data, 213 valid CMPT scores were associated with students who also had a valid grade for CPSC 101 (recall that the CMPT was only started beginning with fall term of 2001). In a similar fashion, the CPSC 212 data set contained 349 valid grades, associated with 130 valid CMPT scores.

### Data Variables

The merged data set for analytical purposes then contains the operative variables and their associated attributes presented in Table 7.

### Audit and Recoding of Data

Several new variables were added, some based on recoding an original variable and others added to support evaluation and validation of the data. Table 8 summarizes those changes and additions.

Tables 7 and 8 then represent the working data for the purposes of analysis. Implied in using the data are other issues that stem from the nature of the problem under study and are discussed below.

### Status of the "Withdraw" Outcome

The literature review indicates that previous work had treated the "W" outcome (a student withdrawing from the course before a final grade is earned) several ways. To some (e.g., Campbell, 1984) a grade of "W" represented an outcome variable indicating the student did not persist in completing the course. On the other hand, Konvalina (1983), who also studied persistence as an outcome, excluded the "W" grade in favor of evaluating those students who actually completed the course. Still others, (e.g., Butcher, 1985, Evans, 1989, and Wilson, 2001) defined their study outcome variable(s) to exclude impact from any course grade; that is, they defined the outcome variables from events during the course like interim and mid-term exams, and exercises or assignments, and did not use the final course grade in the analysis.

Table 7. Data Variables\*

Variable Name	Description	Use in Model
COURSE_STUDENT_ID	Student Identification from the grade data base.	Unique student identifier, match records between data sources.
ADMIT_STUDENT_ID	Student Identification from the admissions data base.	Unique student identifier, match records between data sources.
CMPT_CUID	Student Identification from the CMPT data base.	Unique student identifier, match records between data sources.
COURSE_CODE_MAJOR	Student major at time of taking course.	Sorting variable.
COURSE_ABBRV	Department offering the courser.	Course differentiator for identification.
COURSE_NUM	Course identifier.	Course differentiator for identification.
COURSE_GRADE	Student grade in course.	Dependent (outcome) variable – categorical variable.
COURSE_TERM	Academic term when the course was taken.	Sorting variable.
ADMIT_TERM	Term when student anticipated to begin study.	Sorting variable.
ADMIT_SEX	Student gender.	Independent categorical variable.
ADMIT_RACE	Student race.	Independent categorical variable.
ADMIT_SATPGPR	Institutionally calculated predicted GPR based on SAT & admissions factors.	Independent continuous variable (scale).
ADMIT_SATVERB	Student score on verbal component of the SAT exam.	Independent continuous variable (scale).
ADMIT_SATMATH	Student score on math component of the SAT exam.	Independent continuous variable (scale).
CMPT_Day	Date when CMPT exam taken on-line.	Sorting variable.
CMPT_CMPT	Student score on the CMPT exam.	Independent continuous variable (scale).

\*The format of the variable name includes the source of the variable data, i.e., COURSE from the institutional student data base; ADMIT from the Registrar's admissions data base; and CMPT from the Department of Mathematical Sciences records of CMPT outcomes.

Table 8. Added/Recoded Variables

Variable Name	Description of Change	Use in Model
GradeTest	Interim scale value to calculate $E^2$ .	Support White's test
SRE_1	Interim value.	Support White's test
E2	Calculated regression term.	Support White's test
PrePostCMPT	Categorical variable to classify record as being before or after the CMPT exam was implemented.	Sorting/grouping variable.
DichotCMPT	Dichotomized results of CMPT exam into "qualified for taking CPSC 101" or "not qualified to take CPSC 101."	Categorical independent interim value for exploratory analysis of different CMPT classifications.
Course_AY	Academic year student took the course.	Sorting/grouping variable.
Admit_Gender	Change letter classification (M/F) to numerical equivalent, M = 0, F = 1.	Categorical independent variable.
All_Race	Race data adjusted for missing values, retains all categories of classification.	Categorical independent variable.
Race_Test	Categorical grouping of race classifications into "White," "Foreign," and "NonWhite".	Categorical independent interim value for exploratory analysis of different race classifications.
Race_three	Categorical grouping of race classifications into "White," "Asian/Foreign," and Non-White."	Categorical independent interim value for exploratory analysis of different race classifications.
Race_Collapsed	Categorical grouping of race classifications into "White" and "African American.	Categorical independent interim value for exploratory analysis of different race classifications.
NumGrade	Change letter grade to numerical equivalent, i.e. A=4, B=3, etc.; adjust for missing values and adjust for "W" (withdrawn) outcome.	Continuous scaled dependent (outcome) variable for multiple regression.
Grade_SU	Dichotomized version of NumGrade to indicate student success on non-success in the course, "S" = grade $\geq 3$ (C) and "U" = grade $\leq 2$ (D).	Categorical dependent (outcome) variable for logistic regression.

At this institution one measure of merit for course and curriculum outcomes is the “DWF” rate, the percentage of students who receive a D or F as a grade, or who withdraw from the course before completion. In effect, this approach treats a “W” on the same plane as an unfavorable outcome. Yet the one qualitative study examined in Chapter II (LeJune, 2000) questions this approach, indicating that reasons for leaving a course may in fact have little or nothing to do with class performance at the time.

Based on LeJune’s argument, and on discussion with this institution’s computer science department (A. W. Madison, personal communication, May 3, 2004), it was decided to ignore any record that contained an outcome grade of “W” and to recode it to a missing value. This action resulted in the “NumGrade” variable (Table 8) which also formed the basis for the dichotomized variable of the same data as “Grade\_SU.” This action in effect treats the “W” as a non-outcome and not relevant to the question of predicting performance.

#### Delayed Grades for Unsuccessful Students

The research plan calls for multiple regression procedures to be run against the data to test the strength of the model, and in turn to run logistic regression against a dichotomized (categorical) outcome variable. The logistic regression poses a problem however. Students must qualify for calculus (MTHSC 106) with a given score on the CMPT test; those who fail to qualify for calculus are advised to take an alternative introductory computing course instead of CPSC 101. That is, students who score low on the key independent variable in the logistic regression are excluded from taking the initial introductory course (CPSC 101), thus largely pre-empting the category of students whom

the logistic regression would predict to fail. If these students are not allowed to take the course, how do we determine non-success?

It is not possible to change the policy of the computer science department to accommodate this study so there is no clean solution to this problem. A partial solution is available, however. Many students (roughly 55%) who fail to qualify for CMPT will complete the remediation courses and subsequently take CPSC 101 successfully. Thus, we can pick up a “delayed” CPSC grade by keying on the unique identifier of the student identification number without regard to the date when the CMPT was taken. If the student eventually takes CPSC 101, then the grade is included in the analysis.

As suggested above, the data were reviewed for this situation and it was found that there were 213 valid CMPT scores associated with a student who also had a valid CPSC 101 grade. Of this group, 49 students did not qualify (score of 1, 2, or 3) to take CPSC 101 and 164 did (score of 4, 5, or 6). Of the “not qualified” group, 27 of the 49 students (55%) were eventually successful in the course, that is made a “C” or better as a grade. We conclude then there is a sample of students who took the course with a delay--that is, they took the pre-calculus course before taking CPSC 101 and prediction of “non-success” is possible for the purposes of this analysis.

### Repeated Courses

Under the current curriculum at this institution, it is possible that a student will take one of the introductory courses more than once. This comes about when a student withdraws from the course or makes an unsatisfactory grade (D or F). The student then has the option to leave the major, or, more likely, to repeat the course to improve their

grade over the first attempt. To ascertain the impact of possibly double counting grades in the data set, an analysis was conducted to identify repeat grades.

This analysis revealed that of all grades for CPSC 101 there were 74 instances of the course being taken more than one time, and in a few instances three times. For CPSC 212, there were 49 unique instances of the course being repeated. Since the objective of this study is to evaluate a predictive model for the student taking the course for the first time, it was decided to treat additional grades after the first attempt as missing values thus ignoring them in the analysis and using only the initial grade.

More out of curiosity than any thing else, a short analysis of the repeating grades was done. Of those repeating CPSC 101 ( $N = 74$ ), 30 received the same or a worse grade (41%), and 44 improved to a successful grade above the C threshold (59%). For CPSC 212, ( $N = 49$ ) 17 stayed the same or got a worse result (35%) and 32 improved to a successful grade (65%). This analysis did not include any consideration of CMPT score or outcome.

### Correlations

To ensure that the course grade was suitable for analysis, a Pearson correlation coefficient was generated for grade data associated with CPSC 101 and 212. The correlations were significant at 0.01 level for both cases (Pearson value was 0.261,  $N = 194$ ,  $p < 0.000$  for the former, and 0.288,  $N = 112$ ,  $p = 0.002$  for the latter). From this, it is considered reasonable that the course grade is a suitable outcome measure.

### Linearity of Data

For CPSC 101, a test for linearity was done with the SPSS<sup>®</sup> Means function that uses an ANOVA procedure for the dependent variable NumGrade (scaled data) and the

independent variable CMPT score. This resulted in a deviation F-value of 0.285 with a p-value of 0.836 which fails to reject the null hypothesis that the means of course grade and CMPT score are located on a straight line. For the CPSC 212 data, the test resulted in a F-value of 2.969 and  $p = 0.023$  which rejects the null hypothesis that the means are located on a straight line.

#### Co-linearity

To check for co-linearity, diagnostics for the full model (CMPT Score, SATPGPR, SATVERB, SATMATH, Gender, and race) were calculated. For CPSC 101 data, the variables ADMIT\_SATMATH and ADMIT\_SATVERBAL demonstrated colinearity (tolerance values of 0.127 and 0.138, respectively), a condition that was suspected. For the CPSC 212 data, a similar result was evident with tolerance values of 0.172 and 0.176 for the same variables being evident. We conclude from this that SAT math and verbal scores will probably not be particularly useful as independent variables in the regression analyses.

#### Heteroskedacity

The data were checked for heteroskedacity by using White's Test. This resulted in a Chi Square value of 5.376 which does not exceed the critical value of Chi Square at 4 degrees of freedom of 9.488. Since this is the case, we fail to reject the null hypothesis that variances are equal and assume there is no heteroskedacity in the data.

#### The Model

With all of the above in mind, the final model for initial testing is presented in Table 9.



Table 9. Initial Analysis Model

Variable Name	Description of Change	Use in Model
NumGrade	Change letter grade to numerical equivalent; adjust for missing values and adjust for "W" (withdrawn) outcome.	Continuous scale dependent variable (outcome variable) for multiple regression.
Grade_SU	Dichotomized variable to indicate student success or non-success in the course.	Categorical dependent (outcome) variable for logistic regression.
Admit_Gender	Change letter (M/F) to numerical equivalent.	Categorical independent variable.
All_Race	Race data adjusted for missing values. All categories included.	Categorical independent variable.
ADMIT_SATPGPR	Institutionally calculated predicted GPR based on SAT & admissions factors.	Continuous independent variable.
ADMIT_SATVERB	Student score on verbal component of the SAT exam.	Continuous independent variable.
ADMIT_SATMATH	Student score on math component of the SAT exam.	Continuous independent variable.

### Analysis and Procedures

#### Explanatory Analysis

To test the model for analytical significance, the SPSS<sup>®</sup> General Linear Model (ANCOVA) procedure is used with the full model (CMPT\_CMPT, ADMIT\_SATPGPR, ADMIT\_SATMATH, ADMIT\_SATVERB, All\_Race, Admit\_Gender) as the base case. An alpha of 0.05 was used to determine significance. Within the SPSS<sup>®</sup> procedure, a Univariate analysis was used placing the model's variables into the following categories:

- Dependent = NrGrade;
- Fixed Factor = ADMIT\_GENDER, All\_Race;
- Covariates = CMPT\_CMPT, ADMIT\_SATPGR, SATMATH, SATVERBAL.

### Prediction Analysis

Following Kurtz (1980), Glorfeld, (1984), and Chowdhury (1987), prediction is accomplished using a logistic regression procedure (SPSS: Regression: Binary Logistic), which requires the use of a dichotomous dependent variable. For this process, the dependent variable NumGrade (scaled grade) was recoded to a dichotomized variable (Grade\_SU) to indicate Successful (grade of A, B, or C) or Unsuccessful (grade of D or F) completion of CPSC 101. A similar recoding process was accomplished for the CPSC 212 data set to support logistic regression analysis. Variations of the base (full model) case are then run, removing non-significant variables in turn.

## CHAPTER IV

### RESULTS

#### Organization of the Results

Results of the analyses are presented for both CPSC 101 and CPSC 212 data, the former being the prime course of interest, the latter being a response to Research Question 6, i.e., whether the model has utility for courses other than CPSC 101. For the regression analyses, the explanatory analysis (multiple regression) is presented first, followed by the predictive analysis (logistic regression).

#### Difference in Grades Before and After Implementation of CMPT

The initial question that arises is whether there was a difference in CPSC 101 grades after the CMPT was implemented in the 2001 academic year compared to before. The null hypothesis would be stated: there is no difference between the grade means between students who took CPSC 101 before or after the implementation of the CMPT test. For analysis involving CPSC 212, the null hypothesis would be stated the same, substituting CPSC 212 for CPSC 101. For the answer, a T-test was used to determine differences between grade means in CPSC 101 and 212, before and after implementation of the CMPT test.

Using the dummy or categorically coded variable PrePostCMPT to group the data, it was found that the means in grades for CPSC 101 before and after implementation of the CMPT were significantly different. A rise in grade mean from 2.18 to 2.52 was observed, which was significant at  $p < 0.000$ .

For CPSC 212, the opposite was true, a reduction in the grade mean from 2.84 to 2.55 was not significant ( $p = 0.191$ , equal variances not assumed), indicating there was no statistical difference in grades before and after implementation of the CMPT for these students. This finding, in concert with the rejection of the null hypothesis that the means lie on a straight line (Chapter III), functionally ends any need for further analysis of CPSC 212 data and answers the associated Research Question 6 regarding the model's utility for courses other than CPSC 101.

### Analysis of CPSC 101 Data

#### CPSC 101 Explanation Model (ANCOVA)

To test the model for significance, the SPSS<sup>®</sup> General Linear Model, Univariate procedure of regression with categorical and continuous independent variables (ANCOVA) was used to evaluate each component of the full model as described in Chapter III. Only the CMPT score was shown to be a significant factor ( $p = 0.011$ ), but it only explained 17 percent ( $R^2 = 0.168$ ) of the variance in the data (Table 10).

It was interesting that race was “close” to being significant ( $p = 0.069$ ) and based on this finding, alternative versions of the model were run varying the different constructs of the race variable with the CMPT score. These included collapsing race into white and non-white, and white, non-white, and foreign. Of these, the most interesting was a model which excluded all the variables but CMPT score and all categories (All\_Race) of race that resulted in the following regression (Table 11).

In this restricted model, both CMPT and race become significant ( $p = 0.005$  and  $0.019$ , respectively), and the  $R^2$  value falls slightly to  $0.157$ , or roughly one percent from

Table 10. Regression CPSC 101 – Full Model

Dependent Variable: NumGrade

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	46.269 <sup>a</sup>	14	3.305	1.977	0.024
Intercept	10.381	1	10.381	6.211	0.014
CMPT_CMPT	11.138	1	11.138	6.664	0.011
ADMIT_SATPGPR	.488	1	0.488	0.292	0.590
ADMIT_SATVERB	.336	1	0.336	0.201	0.655
ADMIT_SATMATH	.515	1	0.515	0.308	0.580
Admit_Gender	1.447	1	1.447	0.866	0.354
All_Race	20.101	6	3.350	2.004	0.069
Admit_Gender * All_Race	1.797	3	0.599	0.358	0.783
Error	228.994	137	1.671		
Total	1392.000	152			
Corrected Total	275.263	151			

a.  $R^2 = 0.168$  (Adjusted  $R^2 = 0.083$ ).

Table 11. Regression CPSC 101 – CMPT/Race Model

Dependent Variable: NumGrade

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	43.082 <sup>a</sup>	7	6.155	3.817	0.001
Intercept	11.110	1	11.110	6.891	0.010
CMPT_CMPT	13.108	1	13.108	8.130	0.005
All_Race	25.518	6	4.253	2.638	0.019
Error	232.181	144	1.612		
Total	1392.000	152			
Corrected Total	275.263	151			

a.  $R^2 = 0.157$  (Adjusted  $R^2 = 0.116$ ).

the full model. Of note is the improvement of the p value for CMPT score from 0.011 to 0.005.

To attempt further clarity on the race variable, a second alternative model using race was done, this time recoding the race categories. The categories associated with the race variable, other than African American, represent small percentages of the data population, ranging from 1.1 percent for "Other" to 4.1 percent for Asian American. Thus, African Americans represented the largest minority group at 14.4 percent ( $n = 63$ ). A new variable is created with white and African American as the categories and ignoring other race categories. When the regression is run again with the new collapsed race variable, both CMPT score and Race were significant,  $p = 0.011$  for the former and 0.050 for the latter. The resultant  $R^2$  falls again to 0.137, about a five percent drop from the full model but only 0.02 percent from the regression using the all category race variable. The changes in p value are interesting in that for CMPT it gets larger by an order of magnitude (100%), and for the race variable it drops by one-half.

From these alternative excursions we conclude that race is at its most powerful state when included as a component variable of the full model, even though as a separate variable it is not statistically significant.

To deal with the central question of the CMPT's value as a singular variable the model is collapsed further using only the CMPT score as the independent variable. CMPT stays significant at the 0.001 level ( $p < 0.000$ ) but the  $R^2$  falls dramatically to 0.068 or a difference of nearly 10 percent (Table 12) from the original CMPT/All\_Race model and seven percent from the CMPT/Collapsed\_Race alternative model.

Of the models evaluated, the full model produces the highest  $R^2$  value (0.168), followed by the CMPT/race models at 0.157 and 0.137, and lastly the CMPT only model

Table 12. Regression CPSC 101 – CMPT Only Model

Dependent Variable: NumGrade

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	23.571 <sup>a</sup>	1	23.571	14.013	0.000
Intercept	34.876	1	34.876	20.733	0.000
CMPT_CMPT	23.571	1	23.571	14.013	0.000
Error	322.970	192	1.682		
Total	1735.000	194			
Corrected Total	346.541	193			

a.  $R^2 = 0.068$  (Adjusted  $R^2 = 0.063$ ).

at 0.068. In terms of significant factors, only CMPT is consistently significant at the 0.05 level across all models; race is only significant in the two alternative race-based models.

#### CPSC 101 Prediction Model (Logistic Regression)

For the initial prediction portion of the study, the full model was used; of the model's variables, none of them were significant at the 0.05 level. Still, the model produced the following classification table (Table 13) with an overall accuracy of 94 percent and correctly predicting 100 percent of the successful and 40 percent of the unsuccessful student outcomes.

As with the earlier explanatory analysis, the model was then re-executed eliminating all variables but CMPT and race with the result that the classification table (Table 14) showed a drop in overall accuracy to 82.9 percent. Compared to the full model, the p values for CMPT and race became significant at  $p = 0.011$  and  $0.033$ , respectively.

Table 13. CPSC 101 – Full Model Classification Table<sup>a</sup>

Observed			Predicted		
			Grade_SU		Percentage Correct
			Unsuccessful	Successful	
Step 1	Grade_SU	Unsuccessful	4	6	40.0
		Successful	0	90	100.0
Overall Percentage					94.0

a. The cut value is 0.500.

Table 14. Classification Table – CMPT/Race<sup>a</sup>

Observed			Predicted		
			Coded Outcome		Percentage Correct
			Unsuccessful	Successful	
Step 1	Coded Outcome	Unsuccessful	10	22	31.3
		Successful	4	116	96.7
Overall Percentage					82.9

a. The cut value is 0.500.

To answer the research questions posed in Chapters I and II, another model was analyzed using only the CMPT as the predictor variable ( $R^2 = 0.068$ ). The result is the following classification table (Table 15) which showed a further drop in overall accuracy to 79.4 percent while the significance of the CMPT variable improved to  $p = 0.002$ . Of interest is the failure of the model to correctly predict any of the “unsuccessful” cases.



Table 15. Classification Table – CMPT Only<sup>a</sup>

Observed			Predicted		
			Coded Outcome		Percentage Correct
			Unsuccessful	Successful	
Step 1	Coded Outcome	Unsuccessful	0	40	0.0
		Successful	0	154	100.0
Overall Percentage					79.4

a. The cut value is 0.500.

From the foregoing it appears that the full model is the best overall predictor of student outcomes in CPSC 101. In terms of overall accuracy, the full model has the highest successfully predicted value at 94 percent but suffers from having no significant factors. The next best model is the CMPT/Race model, with an overall accuracy of 82.9 percent with both CMPT and Race having significant p-values. The least accurate model is CMPT only with an overall accuracy of 79.4 percent.

In terms of accurate prediction of successful and unsuccessful students, the full model produces the best combination of predicting 100 percent of successful students and 40 percent of unsuccessful students. The CMPT/Race model falls to a combination of 96.7 and 31.3 percent for the same categories, and the CMPT only model demonstrates the worse combination with successful/unsuccessful predictions of 100 and 0 percent.

CHAPTER V  
DISCUSSION, CONCLUSIONS, AND  
RECOMMENDATIONS

Themes and Threads

Several themes and threads emerge from the previous chapters that must come together here as they impact and influence the nature and interpretation of the results of this study. Some are thematic, that is they provide the conceptualization within which facts and data take their meaning. Others are threads, that is a construct that provides a structure for “hanging” the bits of data and facts. Both of them together provide a framework of meaning, or the place where the analytical and real world meet, and where research finds meaning and utility.

While individual findings are discussed below, it needs to be stated that this study shows that predictive modeling of student success in CPSC 101 is very risky given the limited amount of data typically available. The model developed and tested in this study shows overall weakness and a specific inability to predict the unsuccessful student. Further, the underlying data elements supporting the prediction are limited and in some cases, questionable in their validity and utility. Much of the predictive model’s failures can be traced to the differing environments separating a purely analytical or explanatory model, and the compromises that must be made to bring that model to bear operationally for predictions in real world situations.

### Why do this?

As I stated in Chapter I:

The goal of any advisor is to correctly provide information that matches the student's academic goals and academic capacities at the moment; the student must be capable of performing the work required while at the same time being challenged to learn new skills in the process. The down side of this process is there is a danger in placing students in a position of being challenged to a point where they are unlikely to succeed because of inadequate preparation or development.

Here then is the first theme of this study, that of providing something of value to another, for which the provider (advisor) has a responsibility, and in some cases a fiduciary one. It is not unheard of for advisors to be held, or to be accused of being, culpable for shirking that responsibility or providing inaccurate or misleading information, as a recent lawsuit against this institution demonstrates (R. J. Hendricks, II v. Clemson University, 2003). This responsibility and consequence theme is one of the important motivating reasons for my focus on the operational rather than the analytical. Another way to think about this is to consider what it feels like to look students in the eye, and tell them they are likely to, or not likely to, succeed. This is the real world of the advisor; here the question turns from "I wonder?" to "What happens if I'm wrong?" This is the heart of the challenge in taking explanation into an "operational" or predictive environment. Not only is there a consequence, but it brings forth the real issues of what happens in the more "normal" situation when not all the information (variables) deemed significant in explanatory models is available or, if available, not in sufficiently useful form to present to the prospective student or their parents.

### What Does Operationalization Mean?

Again from Chapter I:

What happens to predictive models of student success in computer science when elements of the predictive model are not available at the time they are needed or available at all? Does the student data available support a predictive function of value to both the advisor and the student? What are the resources associated with providing such prediction and does it provide marginal benefit over current processes in a way that justifies expenditure of additional resources?

At the heart of this second theme is the question: "Because you can do a thing, should you do it?" What are the issues that cause pause? For example, the results presented in Chapter IV suggests that the logistic predictive model under scrutiny here correctly classified 100 percent of the successful students in CPSC 101 based on a model where CMPT score was the only significant variable, but was only correct at predicting 40 percent of the unsuccessful students. With a 60 percent differential between the two possible outcomes, how confident will the advisor be suggesting to students that they will probably not be successful in CPSC 101? Expressed another way, it is unlikely that the advisor would say anything at all beyond "If you make a high score on the CMPT, you will probably do ok in CPSC 101." Even though the advisor now may have additional information based on the predictive model, the uncertainty is so great that full revelation of the prediction may actually do more harm than good.

From the two main themes above we now can evaluate some of the threads that suggest themselves from the analysis and the literature.

### The Importance of Variables

In Chapter II, I suggested that the framework of variables presented by Flanigan, Marion, and Richardson (Flanigan, 1996) provides a means by which the supporting data

in the literature could be structured (Figure 1). Summarized, the demographic and contextual elements (independent variables) of a student's background combine to influence the students' performance on intermediary performance measures (independent variables) which, in turn, form the basis for achievement on the measure of merit, their introductory course in computer science (dependent variable).

The validity of any research effort is dependent on the quality of the data on which it is based. This study shares that dependency, not only in terms of internal or external validity or applicability to the subject under study, but also availability, timing, and even format. Phrased differently, all of these characteristics must come together to be of utility in an operational situation enabling the research to move from explaining to using. This becomes a second thread of the study; unless the data supporting the analysis can be accessed and used when it is needed, it suffers just as much as if the data was not available at all.

#### Data Sources and Availability

Table 6 (Chapter III) clearly shows that the operative components of the model come from disparate sources and must be manipulated into a form appropriate for analysis. Interestingly, only the grade data (dependent achievement variable) are available to the researcher (and the advisor who must make critical predictions) directly; the other categories of data require the assistance of the office of record, who "owns" and maintains the data.

Other explanatory models in the literature (e.g., Butcher and Muth, 1985) have included many more demographic and background variables than were included in this study, including, importantly, math courses taken, high school grade point average and

other transcript-related information. In Chapter II, it was noted that high school class size and student rank information seem to fall out of the data archiving process on its way to the database available to the departmental advisor at this institution. Additionally, high school grade point averages (HS GPA) were not part of this analysis but other studies have shown that the data does exist at this institution (Lane, 2003). However, this information is not available in the database accessible by advisors and thus not recoverable for the purposes of this study. Thus, the base model arrived at for analysis represents the entire inventory of student data available as a routine matter to advisors at this institution.

To be used on a routine basis for advising students (the environment of this study), processes and procedures would have to be implemented to make sure data are available and in a useable form during the summer before the student arrives at the institution. Such processes and procedures do not currently exist to support routine use of potentially useful data, but rather are recoverable through one-time queries or in response to specialized requests. To correct this implies a decision to expend resources across several university elements to provide ready access to the data. The bottom line is that to move the model from analysis to operation requires an infrastructure that does not exist, and the costs of creating such a structure may in fact outweigh the value of implementing the model.

### Leading Indicators

Another key thread is the literature's reliance on variables that simply do not exist in an operational environment, or are not available until the student has matriculated to the institution. For example, Barker and Unger (1983) focused on the administration of a

specialized test of intellectual development but more importantly suggest its utility as a “useful predictor of course performance” for advising is improved “when [it is] used in conjunction with other advising information,” including trailing factors like college GPA (p. 156). Similarly, Glorfeld and Fowler (1982) also used college-level GPA, a “trailing” indicator that is not available until the students have at least one term of college work behind them.

The analysis by Chowdhury et al. (1987) pre-supposes that the student has already taken an introductory calculus course prior to enrolling in the introductory computing course, making it effectively a “trailing” indicator in an operational environment. In the current analysis, however, students are enrolled in calculus at the same time as, rather than prior to, enrollment in the computing course. Thus, the predictor variable used by the Chowdhury study would not be available for the model at this institution.

As a further example, the specialized test used in this study (CMPT), is taken by prospective students on-line, preferably before they come to the institution for summer orientation. However, there is no guarantee that they will take, or have the results, before they arrive for their orientation and initial advising. From the mathematical sciences department’s perspective, this is acceptable as they enforce the results of the test only after the last day of the “drop/add” period has passed (Department of Mathematical Sciences, 2004f). This fulfills their objective of managing class sizes and making sure the students are in the appropriate level math class for their demonstrated skill level. For computer science advisors, however, this timing issue has impact because even though the test may be available for the student to take prior to orientation, there is no enforcement mechanism that ensures the student will take the test prior to the summer orientation sessions and in effect, sets up the “trailing indicator” situation. As a result, advising

comments based on the CMPT in those cases have to be conditional; that is: “. . . if you get a score of such and such, then consider these factors.”

These examples point out the impact of leading and trailing factors for utility in the real world situation of advising. In all the above examples, were these studies performed without the trailing indicators, their models would be significantly weakened. Using trailing variables is not an option in this analysis nor is it practicable for advising newly matriculated students, again reinforcing the conceptual thread that what may work from a purely analytical perspective in actuality is not convertible to the operational environment.

#### Quality of Variables

A final thread deals with the quality of the variables available. It is becoming clear that the operational environment places constraints on the type of variable that the advisor can use to predict student success. But there is also an issue regarding the quality of available data; the question is whether the data is accurate and actually represents the variable intended.

For example, the predicted grade point ratio (PGPR) calculated by the admissions office at this institution is intended to help admissions officers decide whether applicants will be reasonably successful in their work at the college level. For CPSC 101 students, an exploratory T-test showed there was no significant difference ( $p = 0.041$ ,  $n = 414$ ) between the mean PGPR (2.31) and the mean grade earned in the introductory computer science course (2.48), suggesting favorable utility as a predictor. But, as noted in Chapter IV, the PGPR was not a significant variable in the regression analysis on the CPSC 101 grade, a conflicting finding.



Other research at this same institution has also questioned the quality of the PGPR variable. While exploring factors for predicting first-year success of new students (regardless of major or curriculum), Lane (2003) found that the predicted grade point ratio (PGPR) was within the statistical variance for predicting the successful (GPR > 2.0 for the first term) student, but diverged greatly in predicting the non-successful student (GPR < 2.0 for the first term). This finding is consistent with the pattern encountered by my logistic regressions on success and non-success in CPSC 101, which was able to predict success but not non-success.

Further, Lane found that the High School GPR (HS GPR) was a significant predictor of success in college. The problem is that HS GPR was made available for her study by a special request and is not part of the database normally accessible by advisors. The few variables available to the advisor at the institution in the current study have been shown to be of questionable use for predicting success, but at least one potentially better variable is not available.

This finding is consistent with Campbell and McCabe's (1984) observation that not all institutions that might like to predict student persistence (or performance) have a great deal of detail available on the student's high school performance. Specifically, SAT data is usually available, but details about the specific curriculum followed by each student in high school (number of HS math and science courses for example) are not.

#### Summary of the Research Questions

We can now revisit the Research Questions presented in Chapters I and II, and, based on the themes and threads identified above, put the analytical results from Chapter IV into perspective. The Research Questions are:

- RQ1: Is there a relationship between CMPT scores and CPSC 101 student outcomes?
- RQ2: If there is a relationship, is it significant?
- RQ3: If significant, how well does the CMPT predict student performance?
- RQ4: What other factors, in addition to the CMPT should be in the predictive model?
- RQ5: How well do they (RQ4) predict performance?
- RQ6: Does the model have predictive power for other computer science courses in the curriculum sequence?
- RQ7: What is the impact of using a predictive model in an “operational” environment, i.e. is there a “best mix” of variables that has predictive power, yet at the same time can be used within a framework of constrained data such as found in an actual advising situation?

#### CMPT Score and CPSC 101 Relationships (RQ1, RQ2, and RQ4)

The finding that there is a significant rise in the means of the student's outcome (grade) in CPSC 101 after the CMPT test was implemented in 2001 suggests that there may be a relationship between the CMPT outcome and the student's grade in CPSC 101. The full regression model using all available independent variables showed that only the CMPT score was significant at the .05 level but at the same time the analysis showed that the explanatory power of the model was very weak ( $R^2 = 0.168$ ), a result somewhat expected based on previous studies reviewed in Chapter II. But even with reduced expectations of explanatory power, the model showed a much weaker explanation than other studies in the literature whose  $R^2$  values normally ranged from 0.24 to 0.40. This certainly suggests that the rise in grades is probably not due solely to implementation of the CMPT test.

To explore this further, an evaluation of SAT scores for those students in the CPSC 101 data set was performed. It was found that the overall average SAT score of students entering the university increased from 1128 in 1996 to 1204 in 2003 (Clemson University, 2004) and the average SAT score for students taking CPSC 101 also increased from 1115 to 1217 during the same period. This increase suggests that the university is recruiting increasingly brighter students and this may account for some of the rise in CPSC 101 grade means before and after the CMPT was implemented.

It is also interesting that the only way to garner additional explanatory power from the model under study is to collapse independent variables to the CMPT score and Race, excluding the predicted GPR, gender, and SAT math and verbal scores. Yet this outcome tends to contradict other studies which have not found race to be a significant variable (Table 1) for explaining success in computer science. Also, in this collapsed model of regression, an additional reduction of one percent in  $R^2$  results when compared to the full model, rather than improving the explanatory power of the model which one might expect when reducing the number of variables.

The weakness of the model in the current application is dramatically demonstrated when the model is again further collapsed to using only the CMPT as the independent variable by the precipitous drop in  $R^2$  value (to 0.034) and, more importantly, the variable itself becomes non-significant ( $p = 0.053$ ). This result suggests that the CMPT explains student performance only when in the presence of other factors (i.e., Race) or as a part of a larger model with all available factors being considered.

### Quality of Prediction (RQ3 and RQ 5)

The analysis suggests that prediction is risky and indeed predictive (logistic regression) modeling showed that the best results were obtained when the analysis included all the variables that were available (this demonstrated the highest  $R^2$  value), even when many of the contributing variables were found not to be significant. Further, the prediction results were uneven in that the prediction model did a good job of predicting the successful students, but was not a good fit for predicting the unsuccessful student.

The significance of the race variable is particularly interesting; as noted above, it contradicts results found in the literature. Further, the outcome suggests a negative impact, i.e., race may help predict non-success, but not success.

For prediction purposes, it would seem that the model consisting of all the available variables is the “best” predictor of success in CPSC 101. In effect however, it becomes a one-dimensional prediction in that it can only be stated to a student “we expect that you will do well” if they get a “passing” score on the CMPT, but at the same time the advisor must remain silent when they do not “pass” the CMPT, since the negative outcome regarding CPSC 101 cannot be confidently predicted.

### Utility for Other Courses (RQ6)

CPSC 212 (the last of the three “gateway” introductory courses) is the test case for evaluating the relationship of CMPT scores on courses other than CPSC 101. Grade outcomes in this course were found not to have a significant relationship with a student’s CMPT score, suggesting that the intervening academic courses and experience of the students (two additional college terms) may have introduced other factors that were not evaluated, a result that tracks with expectations from the literature (Goold, 2000).

When considered in light of the themes and threads presented earlier, this outcome demonstrates the CMPT scores can only be used in predicting success in the introductory course of computer science. In one sense, this may be an advantage since the issue of “leading” and “trailing” indicators becomes moot when predicting later success in college courses, and the data to support a mid- or late-curriculum analysis is more readily available. Additionally, it follows observations in the literature, particularly Konvalina (1983), Goold (2000), and Chowdhury (1987), who noted that the best predictor of performance is prior performance. In the CPSC 101 case, that college level track record has not been established, but it would have been for CPSC 212 or later courses.

#### Operationalizing Models (RQ7)

Much has already been said regarding the problems associated with moving from the analytical to the operational and as a theme of this study, this movement defines the overall utility and suitability of the predictive model. In this study, the operational logistics associated with gathering, manipulating and analyzing existing data at this institution argues forcibly against the utility of the model without some additional changes to that model and the processes that support it. Simply collecting, handling, manipulating, and merging the different data files in order to make one run of the prediction model is daunting enough, but to do so on a recurring basis to accommodate and keep up with the ever-changing admissions process makes it a virtual impossibility under the existing conditions at this institution. Further, as the data are now housed, student-related data are not routinely available to advisors without the intervention of other University offices. This effectively could have a huge impact on the work load of those offices were predictive models to be employed by even a small set of departments. These issues could be

overcome, but would require resource decisions and commitments, as well as a total change in the handling and processing of student-related data.

Of more direct implication is the unlikelihood that any analytical model could be applied unchanged in an operational environment given the logistics involved in producing data for the different variables in analytical models. Compromises are made by the operational user who must make value judgments about what is available to support the model, make decisions about which variables are directly applicable or which need proxies in order to accommodate data differences between the analytical and operational environments. It is possible to make such adjustments, establish a process for reliably obtaining and processing the data elements, revalidate the efficacy of the model and make it useable in that particular environment. The questions that arise then focus on whether or not such operational changes can be made without destroying the power of the analytical model, and whether or not the institution is willing to commit those necessary resources. Compounding the nature of the problem is the likelihood that one singular model would not be predictive across disciplines or departments, suggesting that the institution would have to establish processes and procedures that can, and would support multiple variations of the predictive model.

Specifically, educational leaders must weigh the cost of operationalizing explanatory models against the potential benefits of doing so. Virtually every study reviewed in Chapter II of this study makes some sort of claim that its proposed explanatory model could have operational application. However, it is not until such implementation is actually attempted that the true extent of the compromises and cost needed to accomplish this becomes evident. Thus, the idea of generalization, an objective of all research, essentially becomes self-limiting when the "generalized explanation"

moves to the specific operational environment where it is needed. Not quite a conundrum, but close.

#### Limitations of the Study

This study suffers the “singular” event limitation in that it is constrained by the environment existing at this institution. At the same time, it does point out the type and kinds of problems that additional attempts at replication might encounter elsewhere.

The study is also limited by the data available, which in the sense of the themes introduced at the beginning of the Chapter, is one of the critical elements being explored. Still, the data do not allow leeway for exploring alternatives without changing dramatically the existing processes and procedures of the institution. Studies such as this may ultimately bring about such changes, making additional data available to support a predictive model, but such changes will come about only when additional studies such as this and that of Lane (2003) can demonstrate a reasonable cost/benefit ratio.

#### Recommendations and Areas for Further Research

In one sense a good study raises more questions than it answers. Were that the only criteria, then this analysis would be considered a raging success. The recommendations and areas suggested for further research are for the most part, directed toward what can be done at this institution to improve the quality of student advising. For this additional research to be accomplished, however, decisions must be taken at several levels within the institution regarding changing processes and committing the resources needed for such change.

### Student Advising - Variables

It appears from the data for this study that this institution has not made a significant effort to provide data to support the advising function. This is not suggested to be a culpable state of affairs, but rather it is a recognition that the types of data collected and maintained are intended to support admissions decisions, not advising based on routine analytical or predictive procedures. In order for the data to support more functions like advising, the types of high school data retained from the application process may need to be re-evaluated, and more importantly, made retrievable by other departments and disciplines within the institution. For example, this study suggested that the high school GPA would be a desirable variable for advisors to have.

At the same time, the study begs the question of what additional variables might be brought to bear. The answer seems to be very few, without some new collection process. This is an area for further exploration across the institution.

### Data Perishability and Compilation

Using pre-admission data for advising purposes places additional requirements on how data is retained, presented, and made accessible to advisors. The current central repository of student data (called the Student Data Warehouse) is retrievable by advisors for only a certain period of time, after which the data file is replaced by data for the next term. Unless a departmental representative downloads and archives the data before term changes, it can only be retrieved by special request to the Registrar's Office. The Registrar maintains a web site where the data elements contained in the various data sets (student data, course data, professor data) are described, but as noted earlier very little of the pre-admission data is maintained and is often duplicated in several different data sets.



For example, the predicted GPR (as calculated by the admissions office) is maintained in two separate data files.

As noted in Chapter II, data on newly matriculating students is distributed informally during the course of the admissions “season” as admission decisions are made. It then falls to the individual departments to archive this data and convert it to a form suitable for further analysis should they so desire. Such a methodology gains some additional data elements (i.e., HS rank and class size) but does not provide the HS GPA which Lane (2003) demonstrated does exist upon special request and may be a potent predictor of success.

Additional analysis is needed to determine the cost/benefit factors associated with changing, modifying, or altering the way student data is processed and archived. Also, there is the issue of how much interest is there across the colleges and departments of the institution to even consider such changes. If there is no demand, there is no need for resource expenditure.

#### Computer Science Department

In terms of the department desiring to use compiled student data and predictive modeling to improve student advising, a critical question becomes, “If we implement such predictive modeling, do we actually improve our advising process over current methods, and is it worth the additional resources (and time) to bring about such changes?”

Coupled with this question is a related issue – this study has shown the relationship of the CMPT outcome to student outcomes in CPSC 101 to be weak at best. In essence, we use as an indicator an instrument intended to measure something other than

performance in CPSC 101. The question that logically follows then: “Should the department consider implementing a specialized computer science instrument for placement purposes?” This question has resource implications and, as above, the question of cost vs. benefits becomes operative and can only be decided after additional analysis and reflection on the cost/benefit ratio.

#### University/Institution Issues

There are several issues that the institution itself may consider for further research and analysis, in addition to the issues raised above in the context of advising and data manipulation and storage.

#### Validity of the PGPR

This study has shown that the logistic model tested was inconsistent in predicting the “non-success” case of student performance in CPSC 101. Even in the “best” model, there was a 60 percent differentiation of accuracy between classifying the successful and unsuccessful student (correctly classifying students 100% of the time for the former and 40% for the latter). The predicted grade point ratio (PGPR) was found not to be a significant factor in this prediction. Additionally, as discussed earlier, Lane’s (2003) work demonstrated that the PGPR also suffered in its ability to predict unsuccessful students. In other words, Lane demonstrated that for successful students ( $GPA > 2.0$ ), the difference in actual vs. predicted GPRs was within statistical limits of acceptability; among unsuccessful students ( $GPA < 2.0$ ) there were statistically significant differences between actual and predicted GPRs by nearly a full grade point and in a direction of overestimating the students performance (actual GPR = 1.36, predicted GPR = 2.54).

This finding questions the validity of the PGPR for any analytical purpose and causes one to consider in what direction would the current predictive model for computer science success move were the variable truly validated. In other words, this analysis indicates that the PGPR is not a significant variable for predicting computer science success. Even so, this question needs further investigation.

#### Maintaining High School GPR (HS GPR) Data

The literature bases underlying this investigation all point toward previous student performance as being significant predictors of continued success. This model did not include high school performance data because these data were not available in the operational environment for use in this study. Yet Lane (2003) demonstrated that the data exist within the institutional data on special request, and more importantly the student's HS GPR was a significant predictor of student success during their first term (p. 55). The institution needs to evaluate the need to make this data element retrievable through some other method that would allow it to be used by other researchers and advisors.

#### Generalization

There are generalizable components that evolve from this study whereby other institutions can take stock comparing their own unique situation in the context of the problems encountered here. Such problems brought to light in this study should provide at minimum a starting point from which other institutional environments may be evaluated. The experience of other institutions moving to implement predictive models will add to the base of knowledge, particularly in exploring cost effective ways of making data available to a broader range of consumers and perhaps suggest other, more potent variables suitable for predictive purposes.

## APPENDICES

Appendix ASample CMPT Test Questions**Sample CMPT Questions**

The CMPT consists of 50 multiple choice questions divided into two parts of 25 questions each. The questions in Part I cover elementary and intermediate algebra and some geometry. The questions in Part II cover precalculus, algebra, and trigonometry.

A satisfactory score on Part I is required for placement into MthSc 101, 102, 103, and 117. A satisfactory score on both Parts I and II is required for placement into MthSc 106.

Your score is the number of correct responses minus one-fourth the number of incorrect responses. On the average, guessing at an answer will neither raise nor lower your score unless you can eliminate some of the possible responses as being incorrect.

**Part I Sample Questions**

1.  $2y - 3(4y - y) =$

A.  $-13y$

B.  $-11y$

C.  $-9y$

D.  $-7y$

E.  $2y - 12$

2.  $(3x^3y)(-2x^2y^3) =$

A.  $-6x^5y^4$

B.  $-6x^6y^3$

C.  $xy^{-2}$

D.  $x^6y^3$

E.  $6x^5y^3$

3. One factor of  $18x^2 - 32$  is

A.  $9x - 32$

B.  $9x - 16$

C.  $3x - 2$

D.  $3x + 4$

E.  $9x + 4$

4.  $\frac{a}{1 + \frac{1}{a}} =$

A.  $\frac{1}{2}$

B. 2

C.  $\frac{a^2}{2}$

D.  $\frac{1}{a+1}$

E.  $\frac{a^2}{a+1}$

5.  $\frac{8}{x^2-4} + \frac{6}{3x-6} =$

A.  $\frac{5}{x-1}$

B.  $\frac{3}{x^2+3x+2}$

C.  $\frac{14}{3x^2-12}$

D.  $\frac{14}{x^2+3x-10}$

E.  $\frac{2x+12}{x^2-4}$

6.  $\frac{xy+y}{y} \cdot \frac{x^2}{x^2+x} =$

A.  $x$

B.  $y$

C.  $x(x+y)$

D.  $\frac{x+1}{x}$

E.  $\frac{x^2y}{x+y}$

7. Which line is parallel to the line having equation  $2y + 4x = 3$ ?

A.  $y + 2x = 3$

B.  $y - 2x = 3$

C.  $2y - 4x = 3$

D.  $-2y + 4x = 3$

E.  $4y + 2x = 3$

## Part II Sample Questions

8. If  $\log_3 x = 5$ , then

A.  $x = 15$

B.  $x = 3^5$

C.  $x = 5^3$

D.  $x^3 = 5$

E.  $x^5 = 3$

9. For what value of  $k$  does the system of equations

$$\begin{cases} 3x + 2y = 5 \\ 12x + ky = 9 \end{cases} \text{ have no solution?}$$

A.  $-2$

B.  $0$

C.  $1$

D.  $4$

E.  $8$

10.  $\tan \frac{2\pi}{3} =$ 

A.  $\frac{\sqrt{3}}{2}$

B.  $-\frac{\sqrt{3}}{2}$

C.  $\sqrt{3}$

D.  $-\sqrt{3}$

E.  $\frac{1}{\sqrt{3}}$



## Answers to Sample Questions

1. D

2. A

3. D

4. E

5. E

6. A

7. A

8. B

9. E

10. D

Source: <http://www.math.clemson.edu/CMPT/sample.html>

Appendix BIRB Approval

To: marion2@CLEMSON.EDU  
Subject: IRB Proposal # EX-0403-009 entitled Operationalizing  
Predictive Factors of Success for Entry Level Students of Computer  
Science  
Cc: weaver3@CLEMSON.EDU

Dear Dr. Marion:

The Chair of the Clemson University Institutional Review Board (IRB) validated the proposal identified above using exempt review procedures and made a determination on March 22, 2004 that it qualifies as Exempt under 46.101 of the Code of Federal Regulations. Activities involving human subjects may now commence for this study.

The IRB requires that you notify this office immediately of any modification to this study. No change in this research protocol can be initiated without review by the IRB. Any unanticipated problems involving risk to subjects, any complications, and/or any adverse events must be reported to the Office of Research Compliance immediately. Please notify this office if your study has been completed or terminated.

Please feel free to contact the Office of Research Compliance at 656-6460 if you have any questions. The protocol number and title of the project should always be referenced in communications regarding this study.

Sincerely,

Benilda P. Pooser, Ph.D.  
IRB Coordinator  
Office of Research Compliance  
223A Brackett Hall  
Clemson University  
Clemson, SC 29634-5704  
Phone 864.656.6460  
Fax 864.656.4475  
pooserb@clemson.edu

## LIST OF REFERENCES

- Accreditation Board for Engineering and Technology. (2003). *2004-05 Computing Criteria*. Accreditation Board for Engineering and Technology, Inc. Retrieved May 3, 2004, from the World Wide Web: <http://www.abet.org/cac1.html>.
- Ahuja, M. K. (1995). *Information technology and the gender factor*. Paper presented at the Special Interest Group on Computer Personnel Research Annual Conference, Nashville, TN.
- Barker, R. J., & Unger, E. A. (1983). *A predictor for success in an introductory programming class based upon abstract reasoning development*. Paper presented at the Proceedings of the 14<sup>th</sup> SIGCSE Technical Symposium on Computer Science Education, Orlando, FL.
- Beyer, S., Rynes, K., Perrault, J., Hay, K., & Haller, S. (2003). *Gender differences in computer science students*. Paper presented at the Technical Symposium on Computer Science Education, Reno, NV.
- Butcher, D. F., & Muth, W. A. (1988). Predicting performance in an introductory computer science course. *Communications of the ACM*, 28(3), 263-268.
- Campbell, P. F., & McCabe, G. P. (1984). Predicting the success of freshmen in a computer science major. *Communications of the ACM*, 27(11), 1108-1113.
- CAS. (2003). *Academic Advising*: Council for the Advancement of Standards in Higher Education.
- Chowdhury, A. A., Van Nelson, C., Fuelling, C. P., & McCormick, R. L. (1987). *Predicting success of a beginning computer course using logistic regression*. Paper presented at the 15<sup>th</sup> Annual Conference on Computer Science, St. Louis, MO.
- Clemson University. *Clemson University Fact Book* [Web Page]. Office of Institutional Research. Retrieved May 19, 2004, from the World Wide Web: <http://www.clemson.edu/oir/factsBook.htm>.
- Clemson University. (2004). *SAT Scores Clemson, National and South Carolina Averages 1990 - 2003* [WebPage]. Clemson University. Retrieved June 7, 2004, from the World Wide Web: <http://www.clemson.edu/oir/factBook/student/freshmenSAT1.htm>.

- DeClue, T. H. (1997). *Academic computer science and gender: A naturalistic study investigating the causes of attrition*. Unpublished Dissertation, Southern Illinois University at Carbondale, Carbondale, IL.
- Department of Mathematical Sciences. (2004a). *How well does the CMPT work for calculus placement?* [Web Page]. Dept of Mathematical Sciences, Clemson University. Retrieved January 22, 2004, from the World Wide Web: <http://www.math.clemson.edu/CMPT/why.html>.
- Department of Mathematical Sciences. (2004b). *Why Math Placement?* [Web Page]. Dept of Mathematical Sciences, Clemson University. Retrieved May 3, 2003, from the World Wide Web: <http://www.math.clemson.edu/CMPT/why.html>.
- Department of Mathematical Sciences. (2004c). *Clemson Math Placement Test (CMPT)* [Web Page]. Dept of Mathematical Sciences, Clemson University. Retrieved June 10, 2004, from the World Wide Web: <http://www.math.clemson.edu/CMPT/toc.html>.
- Department of Mathematical Sciences. (2004d). *CMPT Scores and Math Placement* [Web Page]. Dept of Mathematical Sciences, Clemson University. Retrieved May 3, 2003, from the World Wide Web: <http://www.math.clemson.edu/CMPT/advisors.html>.
- Department of Mathematical Sciences. (2004e). *How Your CMPT Score Will Be Used* [Web Page]. Dept of Mathematical Sciences, Clemson University. Retrieved May 3, 2003, from the World Wide Web: <http://www.math.clemson.edu/CMPT/use.html>.
- Department of Mathematical Sciences. (2004f). *Clemson Mathematics Placement Test (CMPT) 2004* [Web Page]. Dept of Mathematical Sciences, Clemson University. Retrieved June 13, 2004, from the World Wide Web: <http://www.math.clemson.edu/CMPT/advworkshop.html>.
- Evans, G. E., & Simkin, M. G. (1989). What best predicts computer proficiency? *Communications of the ACM*, 32(11), 1322-1327.
- Flanigan, J. L., Marion, R. A., & Richardson, M. D. (1996). Causal and temporal analysis of increase funding on student achievement. *Journal of Research and Development in Education*, 30(4), 222-247.
- Fowler, G. C., & Glorfeld, L. W. (1981). Predicting aptitude in introductory computing: A classification model. *Association for Educational Data Systems Journal*, 14(2), 96-109.
- Glorfeld, L. W., & Fowler, G. C. (1982, February). *Validation of a model for predicting aptitude for introductory computing*. Paper presented at the 13<sup>th</sup> SIGCSE Technical Symposium on Computer Science Education, Indianapolis, IN.

- Goold, A., & Rimmer, R. (2000). Factors affecting performance in first-year computing. *ACM SIGCSE Bulletin*, 32(2), 39-43.
- Konvalina, J., Wileman, S. A., & Stephens, L. J. (1983). Math proficiency: A key to success for computer science students. *Communications of the ACM*, 26(5), 377-382.
- Kurtz, B. L. (1980). *Investigating the relationship between the development of abstract reasoning and performance in an introductory programming class*. Paper presented at the Proceedings of the 11<sup>th</sup> SIGCSE Symposium on Computer Science Education, Kansas City, MO.
- Lane, C. O. (2003). *Predictor variables of academic success for first-time freshmen at Clemson University*. Unpublished Dissertation, Clemson University, Clemson, SC.
- LeJeune, N. F. (2000). *Student Perceived Causes of Attrition in CSI 1300* [Web Page]. Retrieved March 31, 2004, from the World Wide Web: [http://ouray.cudenver.edu/~nflejeun/doctoralweb/Courses/REM6100\\_Qualitative\\_Research/Qual\\_Research\\_Project.htm](http://ouray.cudenver.edu/~nflejeun/doctoralweb/Courses/REM6100_Qualitative_Research/Qual_Research_Project.htm).
- McKendree, J., & Zaback, J. (1988, May). *Planning for advising*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems, Washington DC.
- Natale, M. J. (2002). The effect of a male-oriented computer gaming culture on careers in the computer industry. *ACM SIGCAS Computers and Society*, 32(2), 24-31.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral reserach: Explanation and prediction* (2nd ed.). Fort Worth: Hacrcourt Brace College Publishers.
- Ralston, A., & Shaw, M. (1980). Curriculum '78—is computer science really that unmathematical? *Communications of the ACM*, 23(2), 67-70.
- Teague, J. (2002). Women in computing: what brings them to it, what keeps them in it? *ACM SIGCSE Bulletin*, 34(2), 147-158.
- Timmreck, E. M. (1968). *ADVISER—a program which advises students on courses*. Paper presented at the 1968 23<sup>rd</sup> ACM National Conference.
- Wileman, S. A., Konvalina, J., & Stephens, L. J. (1981). Factors influencing success in beginning computer science courses. *Journal of Educational Research*, 74(4), 223-226.
- Wilson, B. C. (2000). *Contributing factors to success in computer science: A study of gender differences*. Unpublished Dissertation, Southern Illinois University at Carbondale, Carbondale, IL.

Wilson, B. C., & Shrock, S. (2001). *Contributing to success in an introductory computer science course: a study of twelve factors*. Paper presented at the 32<sup>nd</sup> SIGCSE Technical Symposium on Computer Science Education, Charlotte, NC.