Chapter-2

REVIEW OF LITERATURE

Although the theme of Assessment of Aptitude Tests runs throughout this research, three specific areas of inquiry emerged, viz. assessment practices and tools across countries and cultures, the determinants of aptitude test scores and the effectiveness of coaching in improving aptitude tests scores. All these three ‘topic foci’ are reviewed before conducting this empirical study.

2.1 Assessment Practices and Tools across Countries and Cultures

“Assessment is a process which involves the systematic collection, analysis, and integration of information; the main goal of the educational assessment process is to interpret learning patterns for students” (Hoy, 1994). One of the most important goals of education is assessment of knowledge and capabilities. Educating people from different cultures and their assessment is a challenge for global education. One of the goals of
“Education for All” refers to ‘recognized and measurable learning outcomes’ being achieved by all (Goldstein, 2004).

Assessment of student’s performance is often characterized as either formative or summative evaluation (kozleski.et.al, 2000). The former is conducted on a regular basis and later is conducted after a lengthy period of study. Cumming says the main goal of assessment is to evaluate how efficiently students perform (skill), what conviction they hold (values, attitudes), how much potential they have for learning (aptitudes).

### 2.1.1 Assessment Practices Related with GRE/GMAT/SAT

Sui (589-618 A.D.) emperors of China are believed to have first developed the rigorous forms of evaluation of formal educational achievement (Miyazaki, 1981). Then it expanded to west. According to Ikuo Amano (1990) Prussia in 1748 was the first European society to rely on written assessment tests and then middle nineteenth century he called as the age of examinations. University education became prerequisite for seeking government office, thus leading to competition for university entrance (Ringer, 1974). So every country has its history of evaluation and assessment.

The Graduate Record Examinations (GRE) General Test, widely used in assessing the academic qualifications of applicants for admission to graduate school, is designed to measure verbal, quantitative, and analytical abilities that are developed gradually through both formal and informal learning over a relatively long period of time. It is reasonably safe to assume that the scores are an adequate reflection of general ability or achievement at the time the applicant took the test. The problem lies in determining how much the applicant's competence has changed in either direction in the intervening years and how
the change affects present potential for graduate work. (Educational Testing Service, 1984, p. 12) After considering the ambiguities involved, test users could decide whether to waive the GRE requirement, request that the applicants involved retake the test, or use the scores. The desirability of consistency in the treatment of all applicants with older scores was emphasized. Beginning with the 1985-86 testing year, the GRE Program announced a policy, to be phased in gradually, of not reporting scores 5 years old or more (Educational Testing Service, 1985). The policy decision was based on (a) the possibility of differential change in tested abilities over relatively long time periods, and (b) the corresponding need for consistency in treating all individuals with older scores. Although such a policy was deemed to be prudent, the GRE Board recognized that it was important to obtain evidence bearing on how much change typically occurs in abilities measured by the GRE over time, especially in samples representative of the population of examinees (applicants) who ask ETS to report GRE scores that are 5 years old or more.

### 2.1.2 Assessment Practices in Particular with GMAT

In March 1953, deans from Columbia, Harvard, Northwestern, Rutgers, Seton Hall, University of Chicago, University of Michigan, University of Pennsylvania, and Washington University (St. Louis) met with representatives from the Educational Testing Service (ETS) to create an objective, national entrance exam for graduate business school. The exam would test relevant skills that lead to success in the core curriculum of graduate business schools. It would be long enough to be statistically reliable and would be administered and scored uniformly.
In February 1954, ETS administered 'The Admission Test for Graduate Study in Business' (ATGSB) to 1300 students. In the first year, ETS administered the ATGSB to 4288 test takers. At that time, only eight business schools required the exam as part of the application for admissions. Today the GMAT exam is required by more than 1800 business schools and the number of tests administered in 2007 had grown to 219,077.

In the fifty-five years since the first meeting of the Graduate Business Admission Council (changed to the Graduate Management Admission Council in 1976), nearly every aspect of the test has changed. In 1976, the name of the exam changed from the Admissions Test for Graduate Study in Business (ATGSB) to the Graduate Management Admission Test (GMAT). In 1955, one year after the exam was first administered, the Quantitative and Verbal sectional scores were added. In 1961 all but one of the question types were changed and they continued to change until 1994. The length of time required to take the test grew from two hours and twenty-five minutes in 1954 to four hours in 1994 and settled at the current three hours thirty minutes in 1997. Even the test developers have changed.

ETS developed GMAT questions until 2006 when GMAC changed vendors to ACT Inc. Since 1954 the only things that has not changed is the GMAC’s desire to test those skills necessary to succeed in graduate business schools’ core curriculums, the belief that the skills tested on the exam develop over a relatively long period of time, and quantitative problem solving questions.

So why has the test changed? In a 1984 paper, 'The Graduate Management Admission Test: Technical Report on Test Development and Score Interpretation for GMAT users',
William Schrader suggests that from the beginning the association of business schools that became GMAC recognized that the ATGSB and later the GMAT should change as educational researchers discovered new techniques for testing those skills necessary to succeed in graduate business schools. GMAC largely introduced new questions or modified old question types as a result of testing research. For example, the introduction of Data Sufficiency questions resulted from research indicating that Data Sufficiency questions would increase the reliability of GMAT scores and the introduction of the Analytical Writing Analysis resulted from research indicating that writing ability, as tested by a writing task, was necessary to succeed in most business schools’ MBA programs. Although research has been the impetus for many of the changes to the GMAT exam, GMAC, like a business, has had to control costs while responding to the needs of both its board members and customers.

Many of the board members of GMAC are deans of business schools. As a result, some of the changes to the GMAT exam have resulted from business schools requesting change. In an October 2006 report, 'Use of the GMAT Analytical Writing Assessment: Past and Present', Karen Owens cites surveys that indicate admission committees wanted a writing sample from the AWA to be able to personally validate the AWA score. Additionally, on GMAC’s website under the title 'Why You Can Rely on GMAT Scores', GMAC states the GMAT exam has been repeatedly studied, tested, and modified to ensure that it continues to meet the assessment needs of graduate management programs. GMAC surveyed its board members and responds to their needs.
There are many theories as to why GMAC chose the Computer Adaptive format. However, when interviewed by BusinessWeek in April 1998, David Wilson the president and CEO of GMAC cited only two reasons for changing to the computer adaptive format: improved customer service and increased flexibility. Prior to the computer adaptive format ETS administered the GMAT only four times a year. Mr. Wilson cited flexibility as being important to MBA applicants who frequently have busy schedules. GMAC understood its customers and responded to their needs.

By keeping in close contact with board members and sympathizing with customers, GMAC has continued to provide an objective assessment that is relevant to business schools. It is surprising not that the GMAT exam has changed over the last fifty-eight years, but rather how much it has stayed the same. While the format of different questions has changed over time, the GMAC has continued to test skills that develop over a long period and that are necessary to succeed in business schools' core curriculums. Largely the same skills tested in 1954.

Problem Solving questions have not changed since 1954. Data Sufficiency questions have not changed since 1961. It is safe to assume that good quantitative skills have always been necessary to do well in business school. Thus, little change should be anticipated in quantitative question types. However, GMAC has considered making the quantitative section harder. In the April 1998, David Wilson revealed that some business schools would like to see differential equations tested on the exam, implying that the skills taught in some MBA programs require a greater knowledge of mathematics than what is tested. Mr. Wilson suggested that rather than add differential equations to the GMAT exam,
GMAC was considering a second test in addition to the GMAT exam. Ten years later the second test has not been introduced. Just as soon as business schools began thinking that more mathematics was required to succeed in business school, they realized what too much mathematics could lead to monumental real world business mistakes.

In September 1998 (five months after the BusinessWeek interview), Long Term Capital Management made financial history by turning 2 billion dollars into 6 hundred million in 3 weeks. Nobel Laureate professors at Stanford GSB, the University of Chicago GSB, and Harvard pioneered the sophisticated mathematical models that drove Long Term Capital Management. Two of those professors were partners in Long Term Capital Management. Many business schools took note of the Long Terms demise as evidenced by the preponderance of case studies on the topic. After all comprehending differential equations is not necessary to realize that turning 2 billion dollars into 6 hundred million is not a business savvy move. So perhaps just as GMAC began considering changes to the quantitative portions of the test, the spectacular failure of a hedge fund caused business schools to question whether those changes were necessary. While the quantitative section has remained constant, there have been far more changes to the verbal section.

In 1954, there was a Sentence Completion, not Sentence Correction, section on the GMAT exam. The questions asked test takers to choose a set of two words to fill in two blank spaces of a sentence. The Sentence Completion questions were first removed from the ATGSB in 1961 and restored in 1966. In 1976, the sentence completion questions were forever discarded in favor of Usage questions. Usage questions asked students to choose which of four words or phrases in a sentence contained an error. If no choice
contained an error students were directed to choose the last option, which stated “No Error.” In 1983, Sentence Correction question appeared in their present form. Although there have been small changes to this section of the GMAT exam, there has been a section that tested editing ability, and thus writing ability, on the exam since 1966. However, the portion of the verbal section that is made up of questions that test writing and editing ability has increased significantly. On the first ATGSB, Sentence Completion questions were one of three types of questions tested in a 25 minute section. With the introduction of the Usage section, test takers were given just 15 minutes to complete the Usage section. When Sentence Correction questions were introduced, test takers were asked to complete 22 questions in 25 minutes.

The final changes to the length of time allotted for Sentence Correction questions occurred with the introduction of computer adaptive testing. On the computer adaptive GMAT exam roughly half of all verbal questions are Sentence Correction questions. It seems clear that writing and editing abilities have increased in importance since 1954. The trend of increasing emphasis being placed on writing skills is consistent with the introduction of the Analytical Writing Assessment (AWA) in 1994. If the AWA and the Sentence Correction sections are understood both to test writing ability, then approximately 90 minutes of the three and a half hour exam are spent testing writing skills as opposed to approximately 8 minutes on a two hour twenty-five minute exam.

The ATGSB administered in 1954 tested logical reasoning in the Best Arguments section of the exam. Best Argument question consisted of a short passage describing a hypothetical situation followed by questions that asked about the logic of the arguments.
made in the passage. These questions essentially asked test takers to play judge to resolve a conflict. The Best Argument questions were replaced with Organization of Ideas questions in 1961. Organization of Ideas questions directed test takers to classify statements as central ideas, supporting ideas, illustrative facts, or irrelevant statements. These questions were removed from the GMAT test in 1966. In 1972 Practical Business Judgment questions appeared on the GMAT test in two sections Data Evaluation and Data Application. Data Evaluation questions asked test takers to read a long passage and to classify facts that referenced the passage as major objectives, major factors, minor factors, major assumptions, or unimportant issues. The Data Application section directed students to apply some of the data in the passage to answer question. The Data Application questions were dropped from the GMAT in 1984. Practical Business Judgment questions, which had become known as Analysis of Situation questions, were replaced by Critical Reasoning questions in the late 1980s.

In addition to the reading in the Best Arguments section of the first ATGSB test takers reading ability was tested in the Quantitative Reading section. These questions consisted of a passage and possibly a chart. Test takers were asked to infer logical conclusions from the passage. Interestingly, these questions were only used to calculate test takers overall score and were not used to calculate either the verbal or the quantitative sectional scores. Reading Recall or Directed Memory questions replaced Quantitative Reading questions in 1961 and were used in calculating the verbal sectional score. Reading Recall questions directed test takers to read long prose passages and answer questions without referring back to the passage. Preventing test takers from referring back to questions was difficult for test administrators and research suggested that the results of the Reading Recall
section wouldn’t change significantly if test takers were permitted to look at the passage. So in 1977, Reading Recall questions were replaced with Reading Comprehension questions.

Analogy and Antonym questions were eliminated from the ATGSB in 1961, reappeared in 1966, and were eliminated again in 1976. They have not returned. Today the only questions on the GMAT that test vocabulary skills associated are Reading Comprehension questions that ask test takers to identify the answer choice that most nearly provides a definition of a word in the passage. It is interesting to note that GMAC made the decision to remove these questions nearly thirty years before they were removed from the SAT in 2005. In 2005 analogy questions were removed due to complaints that the questions were biased and ambiguous. It seems that GMAC’s constant research and improvement has served all stakeholders well and allowed GMAC to anticipate a problem with a question type thirty years in advance.

2.1.3 Assessment Practices in Particular with GRE

The Graduate Record Examination (GRE) is a standardized test that is an admissions requirement for many graduate schools in the United States, in other English-speaking countries, and for English-taught graduate and business programs world-wide. Created and administered by Educational Testing Service (ETS) in 1949, the exam aims to measure verbal reasoning, quantitative reasoning, analytical writing, and critical thinking skills that have been acquired over a long period of time and that are not related to any specific field of study. The GRE General Test is offered as a computer-based exam administered at Prometric testing centers.
In the graduate school admissions process, the level of emphasis that is placed upon GRE scores varies widely between schools and between departments within schools. The importance of a GRE score can range from being a mere admission formality to an important selection factor.

The GRE was significantly overhauled in August 2011, resulting in an exam that is not adaptive on a question-by-question basis, but rather by section, so that the performance on the first verbal and math sections determine the difficulty of the second sections presented. Overall, the test retained the sections and many of the question types from its predecessor, but the scoring scale was changed to a 130 to 170 scale (from a 200 to 800 scale). ETS does not release scores that are older than 5 years, although graduate program policies on the acceptance of scores older than 5 years will vary.

The computer-based GRE General Test consists of six sections. The first section is always the analytical writing section involving separately timed issue and argument tasks. The next five sections consist of two verbal reasoning sections, two quantitative reasoning sections, and either an experimental or research section. These five sections may occur in any order. The experimental section does not count towards the final score but is not distinguished from the scored sections. Unlike on the computer adaptive test prior to August 2011, the examinee is free to skip back and forth within sections. The entire testing procedure lasts about 3 hours 45 minutes. One-minute breaks are offered after each section and a 10-minute break after the third section.

The paper-based GRE General Test consists of six sections and is only available in areas where computer-based testing is unavailable. The analytical writing is split up into two sections, one section for each issue and argument task. The next four sections consist of
two verbal and two quantitative sections in varying order. There is no experimental section on the paper-based test.

The computer-based verbal sections assess reading comprehension, critical reasoning and vocabulary usage. The verbal test is scored on a scale of 130-170, in 1-point increments (Before August, 2011 the scale was 200–800, in 10-point increments). In a typical examination, each verbal section consists of 20 questions to be completed in 30 minutes.

Each verbal section consists of about 6 text completion, 4 sentence equivalence, and 10 critical reading questions. The changes in 2011 include a reduced emphasis on rote vocabulary knowledge and the elimination of antonyms and analogies. Text completion items have replaced sentence completions and new reading question types allowing for the selection of multiple answers were added.

The computer-based quantitative sections assess basic high school level mathematical knowledge and reasoning skills. The quantitative test is scored on a scale of 130–170, in 1-point increments (Before August 2011 the scale was 200–800, in 10-point increments). In a typical examination, each quantitative section consists of 20 questions to be completed in 35 minutes. Each quantitative section consists of about 8 quantitative comparisons, 9 problem solving items, and 3 data interpretation questions. The changes in 2011 include the addition of numeric entry items requiring the examinee to fill in a blank and multiple-choice items requiring the examinee to select multiple correct responses.

The analytical writing section consists of two different essays, an "issue task" and an "argument task". The writing section is graded on a scale of 0-6, in half-point increments. The essays are written on a computer using a word processing program specifically
designed by ETS. The program allows only basic computer functions and does not contain a spell-checker or other advanced features. Each essay is scored by at least two readers on a six-point holist scale. If the two scores are within one point, the average of the scores is taken. If the two scores differ by more than a point, a third reader examines the response.

The test taker is given 30 minutes to write an essay about a selected topic. Issue topics are selected from a pool of questions, which the GRE Program has published in its entirety. Individuals preparing for the GRE may access the pool of tasks on the ETS website.

The test taker will be given an argument (i.e. a series of facts and considerations leading to a conclusion) and will be asked to write an essay that critiques the argument. Test takers are asked to consider the argument's logic and to make suggestions about how to improve the logic of the argument. Test takers are expected to address the logical flaws of the argument, not to provide a personal opinion on the subject. The time allotted for this essay is 30 minutes. Arguments are selected from a pool of topics, which the GRE Program has published in its entirety. Individuals preparing for the GRE may access the pool of tasks on the ETS website.

The experimental section, which can be either a verbal, quantitative, or analytical writing task, contains new questions ETS is considering for future use. Although the experimental section does not count towards the test-taker's score, it is unidentified and appears identical to the scored sections. Because test takers have no definite way of knowing which section is experimental, it is typically advised that test takers try their best on every section. Sometimes an identified research section at the end of the test is
given instead of the experimental section. There is no experimental section on the paper-based GRE.

An examinee can miss one or more questions on a multiple-choice section and still receive a perfect score of 170. Likewise, even if no question is answered correctly, 130 is the lowest possible score.

### 2.1.4 Assessment Practices in Particular with SAT

The SAT is a standardized test for most college admissions in the United States. The SAT is owned, published, and developed by the College Board, a nonprofit organization in the United States. It was formerly developed, published, and scored by the Educational Testing Service which still administers the exam. The test is intended to assess a student's readiness for college. It was first introduced in 1926, and its name and scoring have changed several times. It was first called the Scholastic Aptitude Test, then the Scholastic Assessment Test.

The current SAT Reasoning Test, introduced in 2005, takes three hours and forty-five minutes to finish, and costs US$51 (US$91 International), excluding late fees. Possible scores range from 600 to 2400, combining test results from three 800-point sections (Mathematics, Critical Reading, and Writing). Taking the SAT or its competitor, the ACT, is required for freshman entry to many, but not all, universities in the United States.

In the late 1800s, a group of leading American universities was concerned about not having a universal way to determine if students were prepared for college-level course
work. They formed the College Entrance Examination Board, and working together they administered the first standardized exam in 1901.

For the first time, students could take one entrance exam for several universities instead of taking a separate exam for each university to which they applied.

The College Board states that SAT measures literacy and writing skills that are needed for academic success in college. They state that the SAT assesses how well the test takers analyze and solve problems—skills they learned in school that they will need in college.

The SAT is typically taken by high school sophomores, juniors and seniors.

Specifically, the College Board states that use of the SAT in combination with high school grade point average (GPA) provides a better indicator of success in college than high school grades alone, as measured by college freshman GPA. Various studies conducted over the lifetime of the SAT show a statistically significant increase in correlation of high school grades and freshman grades when the SAT is factored in.

There are substantial differences in funding, curricula, grading, and difficulty among U.S. secondary schools due to U.S. federalism, local control, and the prevalence of private, distance, and home schooled students. SAT (and ACT) scores are intended to supplement the secondary school record and help admission officers put local data—such as course work, grades, and class rank—in a national perspective.

Historically, the SAT has been more popular among colleges on the coasts and the ACT more popular in the Midwest and South. There are some colleges that require the ACT to be taken for college course placement, and a few schools that formerly did not accept the SAT at all. Nearly all colleges accept the test.
Certain high IQ societies, like Mensa, the Prometheus Society and the Triple Nine Society, use scores from certain years as one of their admission tests. For instance, the Triple Nine Society accepts scores of 1450 on tests taken before April 1995, and scores of at least 1520 on tests taken between April 1995 and February 2005.

The SAT is sometimes given to students younger than 13 by organizations such as the Study of Mathematically Precocious Youth, who use the results to select, study and mentor students of exceptional ability.

While the exact manner in which SAT scores will help to determine admission of a student at American institutions of higher learning is generally a matter decided by the individual institution, some foreign countries have made SAT (and ACT) scores a legal criterion in deciding whether holders of U.S. high school diplomas will be admitted at their public universities.

SAT consists of three major sections: Critical Reading, Mathematics, and Writing. Each section receives a score on the scale of 200–800. All scores are multiples of 10. Total scores are calculated by adding up scores of the three sections. Each major section is divided into three parts. There are 10 sub-sections, including an additional 25-minute experimental or ”equating” section that may be in any of the three major sections. The experimental section is used to normalize questions for future administrations of the SAT and does not count toward the final score. The test contains 3 hours and 45 minutes of actual timed sections; most administrations (after including orientation, distribution of materials, completion of biographical sections, and fifteen minutes of timed breaks) run for about four and a half hours. The questions range from easy, medium, and hard depending on the scoring from the experimental sections. Easier questions typically
appear closer to the beginning of the section while harder questions are toward the end in certain sections. This is not true for every section (the Critical Reading section is in chronological order) but it is the rule of thumb mainly for math, grammar, and the 19 sentence completions in the reading sections.

The Critical Reading (formerly Verbal) section of the SAT is made up of three scored sections: two 25-minute sections and one 20-minute section, with varying types of questions, including sentence completions and questions about short and long reading passages. Critical Reading sections normally begin with 5 to 8 sentence completion questions; the remainder of the questions are focused on the reading passages. Sentence completions generally test the student's vocabulary and understanding of sentence structure and organization by requiring the student to select one or two words that best complete a given sentence. The bulk of the Critical Reading section is made up of questions regarding reading passages, in which students read short excerpts on social sciences, humanities, physical sciences, or personal narratives and answer questions based on the passage. Certain sections contain passages asking the student to compare two related passages; generally, these consist of shorter reading passages. The number of questions about each passage is proportional to the length of the passage. Unlike in the Mathematics section, where questions go in the order of difficulty, questions in the Critical Reading section go in the order of the passage. Overall, question sets near the beginning of the section are easier, and question sets near the end of the section are harder.
The Mathematics section of the SAT is widely known as the Quantitative Section or Calculation Section. The mathematics section consists of three scored sections. There are two 25-minute sections and one 20-minute section, as follows:

One of the 25-minute sections is entirely multiple choice, with 20 questions. The other 25-minute section contains 8 multiple choice questions and 10 grid-in questions. For grid-in questions, test-takers write the answer inside a grid on the answer sheet. Unlike multiple choice questions, there is no penalty for incorrect answers on grid-in questions because the test-taker is not limited to a few possible choices. The 20-minute section is all multiple choice, with 16 questions.

The SAT has done away with quantitative comparison questions on the math section, leaving only questions with symbolic or numerical answers.

New topics include Algebra II and scatter plots. These recent changes have resulted in a shorter, more quantitative exam requiring higher level mathematics courses relative to the previous exam.

The writing portion of the SAT, based on but not directly comparable to the old SAT II subject test in writing (which in turn was developed from the old Test of Standard Written English (TSWE)), includes multiple choice questions and a brief essay. The essay subscore contributes about 28% to the total writing score, with the multiple choice questions contributing 70%. This section was implemented in March 2005 following complaints from colleges about the lack of uniform examples of a student's writing ability and critical thinking.
The multiple choice questions include error identification questions, sentence improvement questions, and paragraph improvement questions. Error identification and sentence improvement questions test the student's knowledge of grammar, presenting an awkward or grammatically incorrect sentence; in the error identification section, the student must locate the word producing the source of the error or indicate that the sentence has no error, while the sentence improvement section requires the student to select an acceptable fix to the awkward sentence. The paragraph improvement questions test the student's understanding of logical organization of ideas, presenting a poorly written student essay and asking a series of questions as to what changes might be made to best improve it.

The essay section, which is always administered as the first section of the test, is 25 minutes long. All essays must be in response to a given prompt. The prompts are broad and often philosophical and are designed to be accessible to students regardless of their educational and social backgrounds. For instance, test takers may be asked to expand on such ideas as their opinion on the value of work in human life or whether technological change also carries negative consequences to those who benefit from it. No particular essay structure is required, and the College Board accepts examples "taken from [the student's] reading, studies, experience, or observations." Two trained readers assign each essay a score between 1 and 6, where a score of 0 is reserved for essays that are blank, off-topic, non-English, not written with a Number 2 pencil, or considered illegible after several attempts at reading. The scores are summed to produce a final score from 2 to 12 (or 0). If the two readers' scores differ by more than one point, then a senior third reader decides. The average time each reader/grader spends on each essay is less than 3 minutes.
In March 2004, Les Perelman analyzed 15 scored sample essays contained in the College Board's *ScoreWrite* book along with 30 other training samples and found that in over 90% of cases, the essay's score could be predicted from simply counting the number of words in the essay. Two years later, Perelman trained high school seniors to write essays that made little sense but contained infrequently used words such as "plethora" and "myriad". All of the students received scores of "10" or better, which placed the essays in the 92nd percentile or higher.

Most of the questions on the SAT, except for the essay and the grid-in math responses, are multiple choice; all multiple-choice questions have five answer choices, one of which is correct. The questions of each section of the same type are generally ordered by difficulty. However, an important exception exists: Questions that follow the long and short reading passages are organized chronologically, rather than by difficulty. Ten of the questions in one of the math sub-sections are not multiple choice. They instead require the test taker to bubble in a number in a four-column grid.

The questions are weighted equally. For each correct answer, one raw point is added. For each incorrect answer one-fourth of a point is deducted. No points are deducted for incorrect math grid-in questions. This ensures that a student's mathematically expected gain from guessing is zero. The final score is derived from the raw score; the precise conversion chart varies between test administrations.

The SAT therefore recommends only making educated guesses, that is, when the test taker can eliminate at least one answer he or she thinks is wrong. Without eliminating any answers one's probability of answering correctly is 20%. Eliminating one wrong answer
increases this probability to 25% (and the expected gain to 1/16 of a point); two, a 33.3% probability (1/6 of a point); and three, a 50% probability (3/8 of a point).

Testing has been the subject of heated controversy in recent years, particularly testing for admissions to educational institutions. Criticisms have been directed toward the appropriateness, accuracy, validity, and social functions of tests. The controversy has become intense and polemical because many people believe that admission tests are determinants of students' educational progress throughout their careers in higher education. They think that tests play the dominant role in determining the students' admission to college, the college they may attend, admission to graduate or professional school, and the quality of the schools the students may attend. (See Nairn et al., 1980; a good summary of the issues is provided by Garvey, 1981). Considering the controversy concerning tests and their assumed importance, it is striking that there is so little information about the attitudes of test takers themselves, and, more importantly, the role the tests play in their decisions. The Russell Sage studies (Brim et al., 1969; Goslin, 1963, 1967) provided some information about people's general opinions about tests, including the facts that the majority of secondary school students did not believe it fair to use intelligence tests to decide upon eligibility for certain colleges, and that the majority of those who had taken college entrance tests felt nervous, lacked confidence, and did not enjoy the experience.
2.2 The Determinants of Aptitude Test Score

2.2.1 Understanding Aptitude Tests

Today's students will live and work in the 21st century, in an era dominated by computers, by worldwide communication, and by a global economy. Jobs that contribute to this economy will require workers who are prepared to absorb new ideas, to perceive patterns, and to solve unconventional problems. Basic aptitude is the key to opportunity for these jobs (Steen 1989). Aptitude testing involves understanding of Verbal, Quantitative, and Logics. It is not simply a collection of facts and procedures, and this is not simply recalling these facts, nor performing memorized procedures.

The three aptitude tests administered by Educational Testing services and other such bodies are Graduate Record Exam, Graduate Management Test and Scholastic Aptitude Test have two basic sections—"The Verbal section" which measures the ability to reason with words in solving problems. The section has analogy, antonym, sentence completion, and reading comprehension items. There are typically 30 items; second, "The Quantitative section" measures basic mathematical skills, understanding of elementary mathematical concepts, and the ability to reason quantitatively and to solve problems in a quantitative setting. The section has quantitative comparison and problem solving items. There are typically 28 items. The tests are administered within particular time limit.

2.2.2 Criticism of standardized tests

A oft-heard, related criticism of standardized tests is that they tap only a very narrow set of skills and abilities and therefore do not reflect the many traits and abilities that are
equally important for successful performance in academia or elsewhere. For example, a former U.S. Secretary of Labor recently criticized standardized tests because of their inability to measure elements of creativity (out-of-the-box thinking, originality, and flair) on which many new jobs depend (Reich, 2001). On this point, most test makers would agree. They would be less likely to agree, however, that selection based on such tests may actually penalize applicants who have other desirable traits. This latter belief is illustrated in the following statement by a distinguished Harvard University professor (Buchbinder, 1998): It is likely that most reasonable people would agree about the desirability of such traits as empathy and altruism. They might, however, take exception to the implicit assumption that the use of Graduate Record Examinations (GRE) scores systematically screens out applicants who possess these worthwhile characteristics. Some time ago, Lloyd Bond (1986) pointed out the misconception underlying this kind of logic, which assumes that people who have some desirable traits cannot also have others.

Bond noted that this fallacy often surfaces in connection with academic admissions tests there are two kinds of people in the world: those who have good grades and high test scores, and those who are responsible, have common sense, and make sound decisions everyday experience. Belief in such dichotomies is popular, Bond noted, although they are contrary to. Although some desirable traits may be essentially uncorrelated, there is little reason to suspect that people who are strong with respect to some personal qualities must necessarily be weak with regard to others.

In addition to everyday experience, there is also empirical evidence that runs counter to the assumption mentioned above. As Zeidner (2001) has pointed out, psychologists have, for some time, explored the relations between intelligence and personality. And, much
recent work (see, for example, Collis and Messick, 2001, for a summary) has helped substantially to elucidate the relationship between personality and cognitive ability.

For instance, Goff and Ackerman (1992) investigated relations between the “big five” personality traits (neuroticism, extroversion, openness, agreeableness, and conscientiousness) and both American College Test (ACT) scores and performance on measures of crystallized and fluid intelligence. They found only slight correlations between most ability measures and personality scales, with the largest being a correlation of $\sqrt{.25}$ between agreeableness and scores on the ACT science reasoning subtest. Conscientiousness (a trait that we have included in our study) correlated from $\sqrt{.14}$ to .02 with ACT scores and reasoning ability measures.

With respect to conscientiousness in particular, in their comprehensive meta-analysis of the prediction of job performance, Schmidt and Hunter (1998) showed that conscientiousness correlates significantly (.31 on average) with overall job performance. Furthermore, measures of conscientiousness were found to be second only to (a) work sample tests, (b) integrity tests, and (c) structured employment interviews in improving the prediction of job performance when used in combination with tests of general mental ability.

In his summary, Eysenck (1994) discussed a number of studies that investigated the relations between the Wechsler Adult Intelligence Scale and profiles on the Minnesota Multi-Phasic Personality Inventory. Based on his review, he concluded, ‘For normal samples there should be little correlation, and that is exactly what is usually found’” (p. 5).
Many other individual studies have also been summarized elsewhere. For instance, Ackerman and Heggestad (1997) conducted a meta-analysis of relations between personality traits and intelligence. Averaged across studies, correlations with a variety of intelligence and ability measures ranged from .05 to .14 for extroversion, from √.05 to .17 for agreeableness, from √.11 to .33 for openness, and from √.19 to .07 for conscientiousness. In short, there seems to be little if any evidence to suggest that personality traits and cognitive abilities are negatively related, or for that matter, strongly related at all.

Another significant body of work has examined the relationship between creativity and general intelligence, finding consistent but weak correlations between these two constructs (Barron and Harrington, 1981; Sternberg and O’Hara, 2000) and a somewhat stronger relationship between creativity and verbal abilities (Qureshi and Qureshi, 1990). The correlation between creativity and IQ tends not to be linear, appearing to diminish beyond an IQ of approximately 120 (Fuchs-Beauchamp, Karnes, and Johnson, 1993; Getzels and Jackson, 1962). Research relating creativity and achievement has produced somewhat conflicting results.

Ai (1999) found that teacher assessments (but not psychometric tests) of creativity were related to self-reports of academic achievement. Toth and Baker (1990), however, found that highly creative students (as identified by high scores on psychometric tests of creativity) were not as likely as less creative students to be successful in a traditional academic setting. Lubinski, Webb, Morelock, and Benbow (2001) found that profoundly gifted students were more likely than their less intellectually able peers to make exceptionally creative or innovative accomplishments later in life. Although the
exact nature of the accomplishments was not predictable from cognitive ability test scores, Lubinski et al. concluded that highly cognitively able people possess more than just “bookish strength” (p. 727) and indeed have the potential for extraordinary achievement. Some personality traits have exhibited somewhat puzzling relationships with other traits. For example, creativity and intelligence have been shown to relate to one another, as have conscientiousness and job performance/ability (e.g., Barrick and Mount, 1991; Hurtz and Donovan, 2000). Yet, creativity in the arts has been found to be negatively associated with conscientiousness for students in creative fields (Dudek, Berneche, Berube, and Royer, 1991).

No such relationship has been reported for scientific creativity. If anything, the tendency is in the opposite direction (Feist, 1999). In summary, there is a good deal of information about the relationship of personality traits to cognitive abilities and performance, although not all of it is entirely consistent. Apparently, however, it is sufficiently widely accepted so as to lead one of the reviewers of our article to feel that we might be “preaching to the choir.”

At first blush, the results seem to suggest that the aforementioned critics of standardized testing, as well as the proponents, are both partly right and partly wrong. For instance, there was no evidence that “deeper-thinking” students are penalized by a standardized multiple-choice test like the GRE General Test. Correlations of depth with GRE scores were near zero (except for a slight positive correlation with GRE verbal scores). Nor was there any indication that more creative students do less well on the GRE
General Test than do their less creative counterparts. In fact, GRE scores correlated consistently, although modestly, with creativity. That these correlations were modest may be a function of the generally high level of intellectual ability of GRE test takers, a hypothesis that is consistent with previous research showing a diminishing relationship at high levels of ability. That quickness exhibited modest correlations with GRE scores would seem, on the face of it, to mean that the test does privilege test takers who work rapidly. An inspection of the questions that comprise this scale, however, would give the critics less reason to gloat. Advocates of standardized testing would probably not be at all unhappy to see that quick test takers—ones who, for example, can “‘handle complex problems’” and “‘love reading challenging material’”—also perform better than their peers on a test like the GRE General Test.

Perhaps the most puzzling finding is a slightly negative, but consistent correlation between GRE scores and conscientiousness. It is unclear why test takers who describe themselves as being “‘able to accomplish work on time’” and “‘careful to avoid making mistakes,’” for example, should do slightly less well on the GRE General Test. Possibly, conscientiousness does not serve test takers especially well on a timed test like the GRE, when time is at a premium.

This hypothesis is, however, only slightly consistent with the data. If conscientiousness does interact with test speededness in a negative way, then we would expect its correlation to be significantly higher with performance on the most speeded section of the test—the analytical portion—rather than with performance on the other two less speeded sections of the test. The differences, however, are slight, although they are consistent with the hypothesis. No compelling evidence here of any strong negative relationships
between personality traits and cognitive abilities as measured by one widely used national academic admissions test. In fact, our evidence suggests that admissions based on the GRE General Test would also, incidentally, more often select test takers who rank higher with respect to other desirable personality traits on personality inventories. Thus, for carefully constructed tests like the GRE General Test, criticisms such as those voiced by Banesh Hoffman and others may not apply with much force.

Several significant variables have been used to predict success in graduate school performance. One factor is the Graduate Record Examination (GRE) score. The GRE has been found to be a valid predictor of graduate student performance in education (Sampson, 2001). This examination has also been shown to be significance predictors of student success in specific courses (Sampson, 2001). House (1997, 1999) asserted that GRE scores can predict overall grade point average and student achievement of degree completion. Morrison and Morrison (1995) examined the predictive validity of the quantitative, verbal, and analytical scores for students’ success in 30 graduate programs. They found that the tests only explained 6% of the students’ performance in their respective field of graduate study.

Another relevant predictor of graduate student performance is age. Peiperl and Trevelyan’s (1997) research on age showed that more younger students than older students performed better in graduate school. Other researchers (Ekpenyong, 2000; Ahmadi, Raiszadeh, and Helm, 1997) have supported this notion by explaining that younger students have recent experience with collegiate environments. Consequently, they are more likely to be able to adjust to graduate education than older students. As another example, classroom instruction has been cited as a determinant of graduate
students’ success. **Knapper (2004)** asserted that most college classroom instruction consists of lectures, group or class discussions, and field based experiences. He continued that the shortcoming of the lecture lies in the method’s passive involvement of students into the learning process. Group and class discussions, however, do allow students to be actively involved in the learning process (Knapper, 2004). Knapper posited that this instructional method allows students to talk with each other about the main points of the lesson. He further denoted that this method provides students with insight on other’s students lived experience.

The GRE General Test measures verbal and quantitative reasoning and assesses academic knowledge and skills relevant to graduate study. The content of the GRE General Test does not relate to any specific area of study (**ETS, 2008**). The number of items and time allowed for completion vary depending on computer-based or paper-based administration. Both the GRE-V and GRE-Q scores are reported on a scale ranging from 200 to 800 points allotted in 10-point increments. The GRE-V section "assesses the ability to analyze and evaluate written material and synthesize information obtained from it, analyze relationships among component part of sentences, and recognize relationships between words and concepts" (ETS, 2008, p. 3).

The GRE "tests basic mathematical skills and understanding of elementary mathematical concepts, as well as the ability to reason quantitatively and solve problems in a quantitative setting" (**ETS, 2008, p. 3**). Educational assessment continues to be a major topic on college campuses as programs strive to achieve or maintain accreditation. In a review, **Allen (2004)** concluded that published tests may not be very useful as a direct measure for program assessment unless those tests are aligned with
program learning objectives. In addition, Firestone, Monfils, and Schorr (2004) expressed concern with the possibility that there will be pressure on instructors to improve student scores on assessment tests. This could lead to more teaching of material found in test questions as opposed to providing a broader range of educational experiences, which might not improve test scores.

In their article concerning marketing curriculum content at AACSB-accredited schools of business, Barnett, Dascher, and Nicholson (2004) reported that written and oral communications were the most important core areas for marketing. While published tests may be effective at measuring subject matter knowledge, they may not be sufficient to assess the specific program objectives that are most important for each of the various departments in a college of business.

Pfeffer and Fong (2002) argued that attainment of a business degree should be related to the various measures of career success. Graduates of business programs should be better prepared for a career and should be compensated at a higher level. They found that, except for graduates from prestigious or top-ranked programs, research indicates that there is little correlation between having a business degree and attaining economic success. They even argued that what is being assessed might be the quality of the student rather than the quality of the program. They believed that assessment methods should be redirected toward meeting the needs of business in order to enhance the career prospects of business school graduates.

Educational Benchmarking, Inc. (EBI) took a different approach to the assessment of business programs. They have concentrated on the methods to use in measuring learning outcomes. Instead of direct measures of actual cognitive learning, they have developed a
method of measuring student self-perceived cognitive learning. They developed this approach from research that reports a high degree of correlation between direct and indirect measures of student learning (Cheseboro and McCroskey, 2000; Richmond, Gorham, and McCroskey, 1987). Students were surveyed as to their satisfaction with all aspects of their business school experience.

Instead of looking at factual retention of data by students, the school can concentrate on those areas that receive low satisfaction scores. A more traditional method of assessment of student learning involves standardized testing of general knowledge of a subject area. ETS provides a series of major field achievement tests, which are easily administered, relatively inexpensive, and can be compared with results from other institutions. There are limitations to the use of standardized tests, as noted by Mirchandani et al. (2001). Two of the most serious concerns are that performance on a standardized exam is best predicted by performance on other standardized exams, and that exam performance may not be correlated with actual knowledge. Even with the limitations, Mirchandani et al. concluded that standardized tests are attractive vehicles for program assessment for several reasons: There is no need for administrators or professors to create or validate them, they are easy to administer, they are graded as part of the cost, they provide comparable data, and they allow testing to begin immediately.

Mirchandani et al. (2001) conducted a study at Rowan University in Glassboro, New Jersey, where the ETS Major Field Test in Business was administered to graduating seniors as part of an assessment effort. They found that two types of variables were related to performance on the ETS exam: input variables (SAT scores, transfer GPA, and gender), and process variables (grades in quantitative courses). They concluded that,
because of the nature of their institution, it was not viable to increase the entrance requirements to improve ETS exam scores. Their institution decided instead to improve their coverage of the material in the quantitative area (calculus, accounting, finance, operations management, and management information systems) even though the process variable effect sizes were smaller than those for the input variables.

While overall GPA was significantly related to both the input and process variables, the SAT score alone explained most of the variation with the ETS results for the subset of the students who took the SAT. The conclusion that Mirchandani et al. (2001) reached was that overall GPA has strong internal validity and provides a measure of student performance related to the curriculum of the school. The ETS scores provided external validity and enabled them to benchmark against national norms.

In their article on the use of standardized assessment tests, Black and Duhon (2003) reviewed how state governments and institutional boards have increasingly begun holding colleges and universities accountable for expenditures. They traced the evolution of the assessment process from the initial emphasis on structure and processes in the 1970s to the current concentration on learning outcomes and continuous improvement. Of particular interest were the changes in the requirements of the AACSB accreditation standards, which focused on evidence of learning as opposed to intended outcomes (AACSB, 2004). This shift in emphasis has created the need for effective means of measuring and improving student achievement. Standardized assessment tests, either developed locally or from a national testing service, can provide a reliable and valid method for institutions to measure the performance of their students. While locally developed tests can be tailored to specific requirements of a program, national tests
enable an institution to compare results for its students with the results for students from other institutions.

Black and Duhon (2003) administered the ETS Major Field Test in Business to 456 students enrolled in a senior-level business core class. Their sample excluded students who had at least 6 hr of business core classes remaining to be taken, and those students who had not taken the American College Testing Program (ACT) exam, resulting in a final sample of 297 students.

To develop a predictive model of determinants of ETS test performance, they gathered additional information about the students in the study. Other studies (Allen and Bycio, 1997; Mirchandani et al., 2001) showed that several variables were strongly related to ETS exam scores. These variables were GPAs (for business core, for economics, accounting, and overall), ACT score, gender, and major. In their study, Black and Duhon added age as a proxy for work experience. Black and Duhon’s (2003) regression model explained 58% of the variation in ETS scores for their student sample.

The predictor variables in their final model were business core GPA, composite ACT score, age, and dummies for male gender and management major. Their results indicated that a one-point increase in GPA was associated with a 7.49 increase in ETS exam score, while a one-point increase in ACT score was associated with a 1.51 increase. Each additional year of age implied an ETS test score .71 points higher. Controlling for the other variables, they found that men can be expected to score 3.79 points higher than women, and management majors can be expected to score 3.57 points lower than nonmanagement majors. Black and Duhon (2003, p. 94) concluded that their “regression coefficients provide credible, useful information.”
In addition, the high significance levels of R2 and the independent-variable coefficients indicate that ETS major field test in Business scores has criterion validity for our program.” From these results from their own institution, they inferred that standardized testing can be used as a method for evaluation and enhancement of programs at other business schools as well. To assist other schools, they listed 17 different uses for the standardized test scores.

They classified these uses into either learning outcomes information, which is used for program and student assessment, or continuous improvement information, which is used for program and student development, and provided several examples.

2.3. The Effectiveness of Coaching in Improving Aptitude Tests’ Scores

Coaching has been determined to be a critical factor for attaining successful results with adult learners (Reece, 1993). Coaching programs that adhere to strict practice, drill, and feedback show a significant positive effect on math scores for adult students (Reynolds and Oberman, 1987). Hoffman reasoned that, either these questions lacked depth or, if the questions did indeed have depth, test takers could not be expected to give well-considered responses to them so quickly. Moreover, asserted Hoffman, test takers who are strong-minded, nonconformist, unusual, original, or creative are forced to suppress their impulses to conform to the norms established by the testers.

My twenty five years of experience as a testing coach confirms my belief that a successful coaching curriculum engages the mental, emotional, philosophical, and spiritual parts of students; the term "coaching", when applied to test preparation programs, involves maintaining rapport between the instructor and student through
matching the general degree of eye contact, noticing and adjusting to any unique cultural, language, etiquette or behavioral norms, noticing key facial expressions, and observing personal space perimeters (Bandler and Grinder, 1986).

Several issues surround the "coaching controversy," including those of equity and effectiveness (Messick, 1981). Equity is at issue because coaching is not uniformly available to all test takers. A situation in which some examinees can attain an advantage by attending expensive coaching programs runs counter to testing's traditional goal of promoting opportunities for the most capable regardless of economic background (Cole, 1982). Controversy also exists with respect to the effectiveness of coaching. With few exceptions (e.g., Alderman and Powers, 1980; Evans and Pike, 1973; Roberts and Oppenheim, 1966), most studies of coaching have relied on comparisons of nonequivalent control groups, instead of on groups randomly assigned to treatments (Messick, 1980).

Between-group differences in the (potentially numerous) unmeasured personal characteristics of coached and uncoached examinees have rendered the results of such studies equivocal. It has been conjectured, for example, that those who seek coaching may be highly motivated to perform well on the test. To the extent that motivation and other similar factors are related to test performance, coaching may appear to be effective even when it is not. Other unconsidered factors, such as test anxiety, which may relate positively to seeking coaching, but negatively to test performance, may result in underestimates of the effects of coaching, or even make it appear harmful. Disagreement over the effectiveness of coaching has been fueled by research studies that have sought categorical answers to a question that, as Messick (1981) has argued, is "more aptly
posed as one of degree: It is not a question of whether coaching works or not, but of how much student time and effort devoted to what kinds of coaching experiences yield what level of score improvements" (p. 11). In assessing the relationship between time and coaching effects, Messick and Jungeblut (1981) found substantial correlations between effects on Scholastic Aptitude Test (SAT) scores and the level of student contact time entailed in coaching programs. Their analyses revealed logarithmic relationships, as is frequently the case with diminishing returns, suggesting that each increase in SAT scores may require geometrically increasing amounts of student contact time (and of all the factors that time may be proxy for).

Messick and Jungeblut cautioned, however, that because these correlations were between average values for groups of examinees, they were predictably higher than comparable correlations based on individuals within groups. In an experimental investigation of the susceptibility of three analytical item types to special test preparation, Powers and Swinton (1984) also found strong relationships between the average time devoted to test preparation and actual performance on two of the three analytical item types. In a more general educational context there is also considerable evidence for the role of time in student achievement, either as an indication of individual effort or of program length (Fredrick and Walberg, 1980; Husen, 1972; Keith, 1982; Polachek, Kniesner, and Harwood, 1978; Stallings, 1980). Much of the controversy has focused on the SAT, a test used to facilitate admission decisions at the undergraduate level. Because of differences in the characteristics of SAT and Graduate Record Examination (GRE) candidates, and in the kinds of test items used in these tests, the studies of the SAT do not necessarily generalize to the GRE Aptitude Test. As a result of these studies (Powers and
of a study of within-test practice effects (Swinton, Wild, and Wallmark, 1983), two item types were removed from the analytical portion of the 1981-82 version of the test. The only other study of the GRE Aptitude Test of which I am aware is one in which Evans (1977) investigated the susceptibility to coaching of the quantitative section of the test, but severe attrition in the sample rendered his results inconclusive.

To date, however, coaching companies have, to our knowledge, documented their claims only by surveying previous customers to ascertain score changes after coaching. Although sometimes verified by prestigious accounting firms, these survey results do not constitute scientific studies. At a minimum, it is necessary to compare these score changes with those exhibited by uncoached test takers, who, for a variety of reasons (test practice, regression effects, and real growth in the abilities measured by the test, for instance) also show improvements upon retesting.

As documented elsewhere (e.g., Powers, 1993), studies published in scholarly journals simply do not support current claims about the effectiveness of coaching for the SAT: the average reported effect for the Princeton Review and for the Kaplan Educational Centers, for example, is 25 to 40 points on the verbal and math portions of the SAT—less if only the best-designed studies are used to gauge impact. All of this evidence, however, was collected for the pre-1994 version of the SAT, not the revision that was introduced in April, 1994. Therefore, as pointed out in a special report on the new SAT (College Board, 1994), until there is carefully controlled research no one will know for certain whether or not the new test is more or less coachable than the old one.
Generally the score increases attributed to special preparations were less than 15 points on the SAT scale of 200-800 points. Despite the consistency of their results, most earlier studies had some methodological weakness which could detract from the strength of their conclusions. These weaknesses arise primarily from the lack of an adequate control group and include: (a) absence of any control group for reflecting usual patterns of score changes (Coffman and Parry, 1967; Pallone, 1961); (b) control groups taken from schools other than those that offered the special preparation, confounding treatment and school effects (Dyer, 1953; French and Dear, 1959); and (c) control groups formed by matching an experimental group on selected variables, permitting initial differences between the groups on other, unmeasured factors (Frankel, 1960; Whitla, 1962).

Those studies with random assignment to treatment and control groups addressed specific topics, such as the effects of intensive individual tutoring (French and Dear, 1959) or the impact of special instruction among student of very low ability (Roberts and Oppenheim, 1966). In their study of the impact of special instruction on item formats under consideration for the mathematics section of the SAT, Evans and Pike (1973) showed significant improvements for certain types of test items when students received over 20 hours of instruction on an item format.

Besides the skills or abilities that they are intended to measure, to some degree standardized tests may also reflect individual differences in prior test-taking experience. While many test candidates are extremely knowledgeable about the appropriate procedures for taking particular tests, other test takers may possess considerably less familiarity with tests and with the proper techniques for taking them. The concerns of those who develop tests then are twofold: (a) to prevent sophisticated examinees from
"beating the test" and (b) to keep less experienced test takers from "being beaten" by it. Professional test developers usually take two approaches to ensure that neither of these outcomes is likely. The first strategy is to construct tests that contain no extraneous clues or other such features that might either inadvertently help test-wise examinees or disadvantage less experienced test takers. This strategy is implemented by developing tests with directions that are easy to read and follow and question formats that are relatively uncomplicated.

The second approach entails reducing the variation among examinees' test-taking skills, i.e., ensuring that all prospective test takers are well versed in the procedures required for a particular test. This approach is often accomplished by imparting the principles of good test taking and conveying, usually in test information bulletins or in test directions, appropriate test-taking strategems ("Budget your time,. Don't spend too much time on troublesome questions. Guess if you can eliminate one or more choices."). In the same publications, test developers also often relay their intentions or their beliefs about appropriate preparation for the test ("Only knowledge of elementary algebra and geometry is presumed. Exposure to instruction in formal logic is not necessary.").

In addition, major test publishers often provide all examinees with examples of question types and response procedures and, increasingly, with full length practice tests or other extensive test familiarization materials. The objective of these efforts is to put all prospective test takers on as nearly equal grounds as possible with respect to their knowledge of a particular test and of the procedures required to take it. In addition to the materials provided by test sponsors, test preparation books are available from numerous commercial publishers. Despite the various claims that are often made for such offerings,
little documentation exists regarding the effects of these books on either test performance or other test-taking behaviors. Some research has shown that programmed texts can be an efficient means of increasing test-taking skills (Petty and Harrell, 1977; Slakter, Koehler, and Hampton, 1970), and short test orientation sessions, such as those employed by Wahlstrom and Boersman (1968), have been cited as effective ways to equalize test sophistication (Anastasi, 1981).

Messick (1980) has critically reviewed the research on the effects of coaching for the Scholastic Aptitude Test (SAT), finding that the length of such programs may be related to the size of the effects on SAT scores. Messick (1980), however, did not consider any studies of examinees' self-study of test familiarization materials. Halpern and Sasajima (1965) did survey PSAT and SAT candidates to obtain their reactions to a candidate information bulletin, but they did not assess the effects of using this bulletin on test performance or other test-taking behaviors. As a result of increasing requests from secondary schools and students for more pre-examination information about the SAT, in the Fall of 1978 the College Board introduced a new test familiarization booklet called Taking the SAT (College Entrance Examination Board, 1978). This includes a full-length sample test and answers to questions, and provides directions for estimating scores on the 200-800 SAT score scale.

This review has been helpful in formulating the hypotheses and methodology with reference to the objectives of this research. However, apart from multiple regression an attempt to innovate the relevant methodology in naïve forms such as residuals’ analysis, forecasting and control.