3-2003

An Analysis of Information Referenced Testing as an Air Force Assessment Tool

Eric D. Larson

AN ANALYSIS OF INFORMATION REFERENCED TESTING AS AN AIR FORCE ASSESSMENT TOOL

THESIS

Eric D. Larson, First Lieutenant, USAF

AFIT/GLM/ENS/03-05

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GLM/ENS/03-05

AN ANALYSIS OF INFORMATION REFERENCED TESTING AS AN AIR FORCE ASSESSMENT TOOL

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Logistics Management

Eric D. Larson, BS

First Lieutenant, USAF

March 2003

AFIT/GLM/ENS/03-05

AN ANALYSIS OF INFORMATION REFERENCED TESTING AS AN AIR FORCE ASSESSMENT TOOL

Eric D. Larson, BS
First Lieutenant, USAF

Approved:

Stephen M. Swartz, Lt Col (USAF) (Chair)	date
Rita A. Jordan, Col (USAF) (Member)	date

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Lt. Col. Stephen Swartz, for his guidance and support throughout the course of this thesis effort. His insight and experience were certainly appreciated. I would also like to thank my sponsor, the United States Air Force Academy, and its personnel for all of their support and latitude provided to me in this endeavor.

I am especially indebted to the Air Force Academy's Management and Biology Department faculty and the students enrolled in the Fall 2002 Management 210 and Biology 331 (Botany) courses. It was only through their assistance that any of this could have been possible. Special thanks go to Lt. Col. Marie Revak and Dr. James E. Bruno. Lt. Col. Revak served as my liaison and was always available to answer my questions. Dr. Bruno provided invaluable consultation in the subject matter.

Eric D. Larson

Table of Contents

Page

List of Tables

Page

Page

Abstract

The Department of Defense (DOD) has placed a great deal of importance on training and education, throughout all areas of infrastructure development and force implementation. A more knowledgeable operating unit, in any situation, is consistently the deciding factor for success. The United States Air Force, too, has emphasized this ideal and sought to employ those persons most qualified for the required task. Yet, problems within the classroom and various training venues are always present and should be continually marked for improvement. Existing assessment techniques should provide an accurate account of the quality of information learned by DOD personnel. This is undoubtedly crucial to war and peacetime functions. Therefore, testing as an assessment tool should be challenged, and new procedures – if deemed effective – should be recognized and introduced.

This thesis looks at examination methods based on confidence-level items and two-dimensional feedback mechanisms. Information Referenced Testing (IRT) has been designed to more effectively measure and reflect the amount of knowledge attained by a student. The following research is an examination of IRT and its role in Air Education and Training Command. It will study two-dimensional items in multiple-choice examinations as a legitimate assessment tool for students, instructors, and administrators.

AN ANALYSIS OF INFORMATION REFERENCED TESTING AS AN AIR FORCE ASSESSMENT TOOL

I. Introduction

General Issue

Educational and training assessment is not an exact science. Certain problems in this field undermine the goals associated with basic, intermediate, and advanced learning. For example, little is understood about the relative strengths of the various techniques for evaluation. In addition, the question of how to apply these methods in different environments is largely unanswered. Do certain situations call for a grading algorithm based on essays, portfolio construction, or direct observation of a particular skill? Are multiple-choice (MC), fill-in-the-blank, or true-false test items more appropriate for certain students, under specific conditions? Other theories on assessment are even less conventional. Some administrators have considered self-assessment, team activities, and narrative evaluation (without grades) as viable alternatives to the traditional viewpoints that are centered on quantifiable and easily comparable measurement scores.

Another area of concern focuses on the economics behind these questions. Standardized testing and MC sets are assumed to be a strong indicator of cognitive proficiency and a predictor of general ability. However, the ease of this type of testing procedure is perhaps the leading reason behind its wide-spread use. Are there other methods that can reflect a more accurate degree of student comprehension, provide for

more quality in feedback, and still adhere to the advantages of multiple-choice questions and related formats?

Background and Overview

Two-dimensional (or confidence-level) testing is a way that may serve to more appropriately reveal the level of student understanding, when compared with traditional methods. It also hopes to remove the error or bias associated with "classical" multiple-choice tests, usually attributable to random guesses for items that reflect unlearned information. One variation on this technique, developed by Dr. James E. Bruno at the University of California at Los Angeles (UCLA), is constructed according to an Information Referenced Testing (IRT) format. Dr. Bruno earned his doctorate in Educational Administration and Systems Engineering at UCLA and is currently a professor in the UCLA honors undergraduate program and the Joint Doctoral Leadership Program at Fresno. His description of IRT is based on a measurement of "information quality" as a standard for selection and the use of output variables defined by "informed," "uninformed," "partially informed," and "misinformed" as a means for detailed feedback to teachers, students, and administrators. The advantages of this method are two-fold. IRT can be used as an accurate post-test assessment tool to assign grades and validate a progression to subsequent phases of learning. In addition, pre-testing with this instrument will allow for an initial evaluation that can more precisely denote the quality of information possessed by a student or group of students at any particular time.

In theory, this type of assessment method hopes to collect the "noise" that results from uneducated guesses and convert it into recognizable performance data. In practice, it attempts to combine the ease of objective testing with the reliability of subjective

(essay and portfolio) grading, through the careful use of confidence-level variables. This gives the student another dimensional alternative with which to respond. In the traditional multiple-choice format, the student or trainee is faced with only one correct option for a given test item. He or she makes what is hopefully an informed response; however, no measure of confidence is revealed in the process. It is impossible to differentiate between what is actually learned through an examination and what is randomly manifested through guesswork.

IRT is designed to reflect the level of the student's assurance that he or she has answered correctly, within the context of the test item. Specifically, alternatives *a*, *b*, and *c* reflect the one correct option and two appropriately wrong options, as for any standard test. Additionally, with two-dimensional testing, option *d* will suggest that *a* or *b* may be correct. This method gives the student an opportunity to exclude one of the choices, stating here that *c* is not the correct option. However, he or she cannot further differentiate between *a* and *b* and is willing to essentially choose both for a small point penalty. In this same way, option *e* will support *b* or *c* being correct, and option *f* will relate to *c* or *a* as the right alternative. Finally, option *g* will be presented as "I don't know" and will admit a total lack of comprehension and a hesitance by the student to make even an educated guess.

Problem Statement

Measuring the exact amount of information "learned" by a particular student is difficult. The level of comprehension cannot be perfectly quantified in the testing process. Examination formats should (as closely as possible) reference scores to the percent of subject matter knowledge attained by the student and provide feedback related

to the specific gaps in instruction that can be corrected in the continuing educational and training process. Today's popular testing techniques require improvements in this area.

Research Question

The research question for this experiment should challenge traditional multiple-choice and IRT-constructed examinations and stimulate an exploration of the constructs that are most actively evaluated when test items are analyzed by the student. Simply stated, is the implementation of IRT possible, in a practical sense, and how effectively can this testing method produce informative assessment results and performance feedback, when compared with other methods? By investigating this question, Air Education and Training Command (AETC) may benefit from the analysis of its conventional testing instruments and gain exposure to some viable alternatives. The goal is an easily-administered and economical program for detecting the "gaps" in student and trainee learning, before they are brought to bear on the battlefield.

Investigative Questions

In order to address the measurement problem, certain investigative questions should be answered.

- Will IRT and traditional tests show a significant difference in scoring? If so, to what factors can this be attributed?
- How will students with varying aptitudes and attributes respond to these newly-designed test variables?
- How will students and teachers view this type of examination, overall?
- What are the process issues behind the implementation of IRT in the classroom?

- Can IRT produce a more accurate reflection of the actual amount of information obtained by a particular student, when compared with traditional multiple-choice methods?

Experimentation and research will be guided by the need to answer these questions. Specifically, the noticeable effects of IRT on a group of students must be understood, and this investigation should be narrowed to account for the individual differences among test-takers. Additionally, the level of effort required to set-up and sustain this new system (and its impact on those persons involved in the actual process) should be explored and observed in a real classroom setting. Finally, the overall purpose of IRT should be attacked – yielding a greater understanding of its significance and ability to provide accurate assessment results.

Summary and Conclusion

This chapter presented a general description of IRT and proposed some possible benefits. A problem statement was given, and an overarching research question succinctly considered the main issues for future research and experimentation within the context of IRT study. Investigative questions, whose answers are intended to validate or refute hypotheses related to the given research question, were presented. Subsequent chapters will reveal the appropriate classroom and web-based experiments that will attempt to provide quantifiable and analyzable results. Chapter II will study the history of assessment, reveal some of the theories behind successful teaching and testing, and subsequently "set the stage" for IRT as it exists today.

II. Review of Literature

Background and Overview

This chapter will discuss the importance of classroom assessment and feedback, including the reasons behind the practice of testing. An exploration of the history of performance measurement will eventually classify a variety of methods, each associated with the appropriate testing tools that can be used for the most efficient results. Through this analysis, two general instruments will present themselves – constructed response and multiple-choice item formats. Most of the reasonably accepted practical assessment methods, it will be found, can be distilled into one of these two categories.

The focus will then turn to multiple-choice (MC) testing and its role in today's various classroom environments and learning applications. The inherent advantages and disadvantages of MC testing will be defined, with a special look at some of the variations used in current academia. To follow this, the relevant concepts in MC assessment will be explored, including a background report on the optimal number of test items, some sources of error, the effects of guessing, and the quality of expected feedback. Finally, complex issues associated with gender testing discrepancies and other areas of item discrimination are explored, with a look at possible solutions. This will provide the appropriate introduction for confidence-level testing and the various kinds of two-dimensional item format structures.

Information Referenced Testing (IRT) will serve as the focus for the remainder of the analysis, and the experimental portion of the study will serve to validate or refute certain aspects of its proposed efficacy. This chapter will introduce the concept of IRT,

including a brief history of its evolution, a description of its modern form, and the advantages and disadvantages that have been discovered in the research community. IRT has been shaped by the combined efforts of a collection of notably influential founders and contributors, and the feedback from their previous experiments will provide invaluable guidance. A final review of the literature will look at alternate applications for IRT, including a specific analysis of web-based instruction and testing within the virtual classroom.

Assessment: A Brief History

No summary on the history of assessment can be complete and exhaustive; however, some of the finer points, researched by George Madaus and Laura O'Dwyer, should be highlighted and presented for illumination. Performance assessment had its beginnings in Chinese culture, existing even before the Common Era. During the Sung Dynasty, candidates wishing to join the civil ranks were required to take examinations in a number of disciplines. Based on Confucian ideals, originality and composition were encouraged, and students were required to recite passages from memory, discuss literature, compose critical essays, write original poetry, debate important political conflicts, and perform readings of classical verse. Later, the emphasis on reasoning or "higher-order" thinking was eventually abandoned, "because government officials became worried that the scoring of these questions would be too subjective; thus they reverted back to questions that required more rote answers" (Madaus and O'Dwyer, 1999). Military examinations, on the other hand, were based more on the demonstration of skill. Candidates were judged on strength and aptitude with a sword and bow.

Evaluation was more objectively scored, although partial credit was awarded for less than perfect exhibitions in training environments.

In Europe, almost a thousand years later, eighth century knights and priests were examined on strict memorization and oral recital of answers to questions. "In the late 12th century, the University of Paris and the University of Bologna were the first to introduce 'examinations' as we know them" (Madaus and O'Dwyer, 1999). Again, students were required to submit oral presentations in response to questions asked about religion and literature. The lack of written tests can be attributed to a scarcity of paper and the insistence that a well-spoken individual was the mark of an educated man. Eventually, written examinations were used by 16th century Jesuit schools to test the use of Latin composition (Madaus and O'Dwyer, 1999). While testing procedures were changing, the standards of evaluation were still based primarily on qualitative appraisals and subjective assessment.

Following this period, two kinds of performance assessments persisted in Europe: "those used to certify guild members, who worked with their hands, and those used to assess 'gentlemen,' who studied the seven liberal arts (grammar, logic, music, rhetoric, arithmetic, geometry, and astronomy)" (Madaus and O'Dwyer, 1999). Early on, the concepts associated with numbers and quantitative scoring measurements represented a superstitious taboo and were avoided by most people. The age of exploration, the Crusades, expansion, and increased trade ushered in a feeling of necessity, though, for the world of time and costs. Soon, the industrial revolution began, and "this shift toward quantification intersected with the assessment of achievement in a profound way" (Madaus and O'Dwyer, 1999). Starting in the early 1800's, the transmission of specific

information was demanded in testing, and the use of schooling as a "political, administrative, and accountability technique" eventually gave rise to a form of standardization and comparison for schools, teachers, and administrators (Madaus and O'Dwyer, 1999). The 19th and 20th centuries were punctuated by overwhelming scientific achievement and efficiency in manufacturing, as well as the need for numerical ranking systems for the workers – thus the world of assessment (as we know it) was born.

The early 1900's saw the introduction of multiple-choice items. This was partly an outgrowth of Frederick Taylor's book, The Principles of Scientific Management, which required that "growing numbers of children be tested to measure a school district's efficiency" (Madaus and O'Dwyer, 1999). The early pioneer of "norm-referenced testing," Frederick Kelly, began using MC tests in 1914, while inefficient essay and oral examinations were slowly phased out. The invention of the high-speed optical scanner in 1955 "sealed the eminence of the multiple-choice item for the next 35 years," and computer-adaptive testing in the 1970's only served to enable MC exams further (Madaus and O'Dwyer, 1999).

Recently, though, MC items have begun to recede in popularity. Educators are looking at more reliable measures of proficiency and knowledge. Europe, for example, has never strayed from its adherence to essay testing as the primary form of assessment. However, in order to return to performance-based evaluations in the United States, academic systems may have to devolve, in terms of manageability, standardization, efficiency, and expense. To combat this, practitioners are looking for easily administered testing techniques which can reveal more of the student's own knowledge base, especially when compared with traditional MC exams. This has been the main cause

behind an emerging interest in confidence-level testing and evaluation initiatives that go beyond the "Right-Wrong" philosophy of currently accepted "normalized" tests.

Testing and Assessment

In order to understand the concept of "testing," one has to appreciate the fundamental nature of education and the appropriate constructs that must be measured. These constructs can be grouped into three main categories: knowledge, skill, and abilities (Haladyna, 1999). While knowledge is best defined as an attainment of facts, principles, and procedures, skill involves a mastery of some type of performance. Abilities, on the other hand, can be characterized as "complex human characteristics that grow slowly over a lifetime and consist of knowledge and skills, emotional characteristics, and the tendency to integrate these in some complex behavior toward some desired end" (Haladyna, 1999).

Different situations certainly require an exhibition of at least one of these constructs, if not a subtle combination of all three. Today's society is focused on order, balance, and a sense of fairness – as well as a measure of optimization. Appropriately, persons are placed along the educational and occupational "food chain" in an attempt to align the right person for the right position or the correct progression based on his or her level of knowledge comprehension, training level, or cognitive potential. In order to accomplish this, examination and measurement are imperative. However, the nature of the testing instrument should be "true." In other words, evaluations should produce results that can accurately reflect the level of knowledge required to complete a given task. Air Education and Training Command (AETC) should be especially vigilant in this regard, because of the obviously severe nature of its business. Therefore, administrators

should be attuned to the appropriateness of testing procedures placed in educational and training environments, to ensure an optimal level of performance.

Major Formats in Testing

Obviously, no one type of testing can sufficiently cover the spectrum of assessment goals for every learning scenario. Testing procedures that serve to measure these constructs will vary, but they might be narrowed down into the categories associated with constructed-response (CR) and MC question formats. CR items include those measurement tools designed around critiques, demonstrations, essays, experiments, interviews, oral reports, portfolios, projects, and research papers (Haladyna, 1999). MC formats can be defined by prescribed alternatives from which a student must "choose." True-false items, pictorial item sets, and matching also fall under the MC heading, but the conventional multiple-choice examination and some of its more complex derivatives are commonly used in large-scale testing programs.

The question presents itself: when should a student choose from a list of alternatives instead of writing out a detailed response? Can one method outperform (or rather out-measure) the other? The answers are elusive in most scenarios and can best be characterized as conditional for the manner of the construct in question. High-inference CR formats, for example, require "expert judgment about the trait being observed," so that "scoring guides" and "descriptive rating scales" are used to evaluate many abstract qualities (Haladyna, 1999). Low-inference formats involve simple observations. The decision to use MC will most likely depend on the expected scope and difficulty of assessment. Some educational programs and training agencies favor multiple-choice items as more efficient to construct, administer, and score. They may also prefer MC

because of the removal of grading-bias. Handwriting, presentation skills, and personal eloquence may not embody a desired measurable attribute, and standardized items presented to all examinees in a congruent and fair manner are more likely to normalize the population and reflect those with greater knowledge in a particular area. For this reason, national and statewide assessment programs, school districts, and certification and licensing companies usually elect to use MC instead of CR for the purposes of aggregate testing.

What, then, are the specific ramifications of MC testing formats, when used in place of essay versions? If time, ease, and objectivity are the main reasons behind the widespread use of MC items, what exactly is being sacrificed? A 1982 study by Donne Alverman and Ned Ratekin surveyed a group of 98 "average" seventh and eighth grade subjects, after each was asked to read a certain passage and complete a multiple-choice and essay examination (Powell, 1989). These researchers elaborated only on those results that reflected a significant difference. "They found that subjects who read to respond on an essay test 'reread' more frequently than students who read the same passage knowing they will respond to multiple-choice items" (Powell, 1989). This suggests a more concentrated effort for learning when a written or CR response is expected. In addition, the essay testers used multiple reading and comprehension strategies "nearly twice as often" as those testing with MC items (Powell, 1989). Again, it is evident that open-ended responses are more effective at eliciting a greater level of thinking and answering. Perhaps this is due to an elevated style of comprehension required in CR test items, when compared with MC items covering the same material.

With this in mind, approaches in educational reform have created unique methods for student and trainee evaluation, some of which use technological advances as a means for a more robust level of cognitive development. Oral presentations, progress interviews, computer simulations, and videotaped presentations all provide a means for challenging students in the classroom, while providing a higher quality of feedback. In this way, individual improvements are more accurately monitored. For example, computer modeling and simulation programs can be designed specifically for science lessons and experiments. Students are led through a module that is intentionally designed to test the desired objectives of the lesson, and printout reports can be collected and evaluated to ensure that an appropriate level of learning has occurred. The entire process is easy to use, economical, and accurate.

Technology is not the only driving force behind a nation-wide resolution to make curricular changes. Too often, it is generally believed, test-takers are forced to deal with artificially constructed problems, usually standardized in the form of conventional MC test items. Researchers and administrators are looking at examinations that are more adept at challenging the more applicable performance-level of the students. "Performance assessments where students read, write, and solve problems in genuine rather than contrived situations are now considered legitimate alternatives to relying only on the results of standardized tests" (Conderman, 2001). In response to this, the University of Wisconsin at Eau Claire has developed some alternative assessment activities, including student portfolios and exit interviews, to better gauge the level of proficiency gained by members in certain departments. While these types of programs can be more time consuming and require greater effort on the part of instructors, the

performance-based measures "have provided faculty members with valuable information needed for formative and summative evaluation" (Conderman, 2001).

Even less conventional tactics have been employed at some universities throughout the country. One such technique has been recently exhibited at Antioch University in Yellow Springs, Ohio. Instructors there are currently working with a grading system that is (ironically) devoid of grades. Evaluators are instead assigned to complete a brief survey form for each particular student, for each course. Each pupil is rated with respect to mastery of material, commitment to learning, group interaction, completion and quality of projects, and an awareness of diverse perspectives. These categories are marked by four levels of proficiency, ranging from inadequate to outstanding, and the instructors are then required to complete a one to two-paragraph narrative, describing the student's performance in class. These assessments become part of the student's academic record, although employers and graduate schools have expressed some concern (Malarkey, 2002). Indeed, this level of subjectivity, while beneficial in some respects, may create a level of evaluation that will not portray the most accurate level of knowledge for each student. In short, teachers may not be able to reflect individual performances that can be measured and compared in the competitive job and college marketplaces.

The MC Approach

Multiple-choice items are not ideal, but they are a necessary form of assessment, and it is important that practitioners are able to discern the specific issues governing their use. An understanding of these strengths and weaknesses (along with possible corrective methods) will allow for an appropriate evolution in the way MC is implemented.

Advantages.

To defend the quality and usefulness of any MC exam, one has to consider the requirements for reliability, validity, and efficiency. "Reliability involves the extent to which we are measuring some attribute in a systematic and therefore repeatable way" (Walsh and Betz, 1985). Reliability is predicated on the assumption that any test measurement involves a true score and some random error. A reliability coefficient therefore reflects test quality through a proportion of these two values. The weight attached to any form of error will cause an inconsistency of results, if the same testing procedure is repeated. "The most important means of increasing the reliability of a test is to improve the individual items in the test" (Miller, Williams, and Haladyna, 1978). However, low consistency in results can stem from ambiguous directions or lack of objectivity in scoring. MC tests are designed to combat these effects. More test items (derived from the speed from which multiple-choice tests are administered and scored) will dampen the errors associated with a lack of reliability. Also, MC exams are well-understood because of their common use. The instructions are easy to comprehend, and scoring is almost wholly objective. In fact, many times, grading can (and does) occur with the help of computers and machines.

Test validity refers to the "extent to which the test we're using actually measures the characteristic or dimension we intend to measure" (Walsh and Betz, 1985). In order for test items to exhibit some measure of validity, certain inferences must be accurate in relating an examinee's performance to a level of subject comprehension and understanding in real-world application. There are two type of validity that will characterize MC tests. In the classroom environment, content validity is defined as "the

correspondence between material that is taught and material that is tested" (Miller, Williams, and Haladyna, 1978). This is simply achieved by aligning test items with the desired (and hopefully covered) subject-areas from the course. Again, multiple-choice tests allow for more objectives being tested in the same amount of time, and a greater sampling of the population material can be represented on the exam. "Predictive validity represents the degree to which a test score allows you to correctly anticipate student performance on some later task" (Miller, Williams, and Haladyna, 1978). This type of validity is useful for selection of students into particular programs, schools, jobs, or award societies. As stated before, MC test items are prized for their objectivity and easily quantifiable and normalized results.

"'Efficiency' is best measured in time and cost to teacher and students" (Miller, Williams, and Haladyna, 1978). One of the disadvantages of MC testing is the time needed for construction. In opposition to any standard essay quiz (whereby students are presented with open-ended items and an opportunity to expound using a sense of personal interpretation), multiple-choice items require more effort to ensure the same degree of efficacy. However, items can be saved for subsequent testing, with some minor corrections or updates before any future use. Plus, the ease of scoring allows for timely feedback. Computers and software can assist in the grading process, thus minimizing instructor time and effort.

Disadvantages.

Despite the inherent convenience and objectivity of MC testing, some critical attention has been placed on this type of item format. Some argue, for example, that "multiple-choice tests encourage teaching and learning of isolated facts and rote

procedures at the expense of conceptual understanding and the development of problem-solving skills" (Rogers and Ndalichako, 1997). Even further, the dichotomous nature of scoring MC exams is frowned-upon, because most outputs express only the number of right and wrong responses exhibited by any particular student, and a wealth of assessment value can be obtained by looking at the specific incorrect responses that were chosen.

With respect to construct validity, some have questioned the widespread use of MC scoring models "by which a total score is simply the sum of the item scores with no regard at all as to how examinees arrive at their different total test scores" (Rogers and Ndalichako, 1997). In other words, this aggregate performance-value is a number that is perhaps wholly unrepresentative of the underlying factors that are desired for measurement. It can be assumed that students possess partial subject matter knowledge for most test items, and evaluation should not be expressed in "black-and-white" terms. To place this in a different context, it can be said that essay questions are rarely given full credit; instead, a continuous spectrum of scoring is employed, with full knowledge and explication of the correct response acting as the standard for a perfect score. MC items, on the other hand, are discretely measured on a binary scale. The student's partial knowledge is translated into a right or wrong (R-W) response, and a deeper understanding of his or her knowledge base is impossible to distinguish.

The nature of conventional MC test items is certainly assailable, especially due to its scoring philosophies, which typically ignore the incorrect answers that are chosen by the student. Conversely, selections made by the student which reflect the correct answer are equally fallible in construct measurement. Essentially, "a 'correct' response may come about because of total knowledge, partial knowledge, misinformation, or guessing"

(Rogers and Ndalichako, 2000). There is simply no way of knowing which of these levels of comprehension has been attained. Additionally, formulas to correct and account for guessing have been inadequate. "Students rarely guess randomly, thereby making invalid formula scores" (Rogers and Ndalichako, 2000). However, testing methods designed to somehow measure the degree of partial knowledge, misinformation, and full information have received some attention. The key is to somehow find a manner of examination that will guide students toward a reflection of these constructs. Otherwise, examinees will be subjected to an R-W evaluation that is typically inaccurate and provides no real feedback, especially for the lower to middle-achieving students that are prone to miss a higher number of questions and guess more frequently.

Perhaps the greatest criticism for MC testing relates back to the three elements of education (knowledge, skills, and abilities) discussed earlier. Thomas M. Haladyna (1999) describes skills as mostly performance-oriented, and, though some are derived from mental tasks, all are indirectly measured by multiple-choice exams. He specifically discusses writing and mathematical computations, both of which are widely tested using MC item formats. He asserts, "knowing how to perform a skill is not quite the same as actual performance" (Haladyna, 1999). Stated another way, obtaining knowledge and performing in an objective (MC) testing environment may not extend to the world of practical application. If one cannot use an intellectual entity in daily life, "the acquisition of knowledge and skills seems pointless" (Haladyna, 1999).

Abilities, too, are a mixture of innumerable physical, mental, and psychological traits that are ultimately assigned to specific real-world tasks. These "can be taught and learned, but there is a poor history of testing them" (Haladyna, 1999). This is especially

true if the testing vehicle is composed of multiple-choice items. Other academics point out that "some learning outcomes – such as driving a car, typing accurately, writing a convincing paragraph, or molding a clay pot – can be judged best through actual performance" (Miller, Williams, and Haladyna, 1978). However, these authors do suggest that MC tests can measure intellectual processes beyond simple knowledge of facts. The lesson is that MC has a place in the assessment environment, but practitioners should be careful to guard against incorrect causal relationships between test performance measurement and educational construct attainment.

General Structure, Philosophy, and Some Variations.

All MC questions consist of a stem, which presents the problem statement, and several alternative responses. These options available for the student should consist of a correct answer and a certain number of plausible wrong answers – known as "distracters." "A high-quality MC question should present a task that is clearly understood and be constructed so that it can be answered correctly by those who have achieved the intended learning outcome" (Hansen, 1997). Conversely, the uninformed student should not be made aware of this "approved solution." In addition, test-writers should be aware that test items that involve "all of the above" and "none of the above" responses should be used with some caution. Some practitioners argue that this may not accurately reflect the desired construct. In other words, students can choose "all of the above" or "none of the above" by merely identifying two or more choices that reflect the correct or incorrect answers, respectively. This reflects a measure of "full knowledge," when only a certain level of partial understanding may be clearly evident.

Perhaps the main reason behind the widespread use of multiple-choice testing derives from the ability of administrators to ensure the fair and equitable treatment of every test-taker. "Standardizing, in this respect, means that each student is exposed to the same or equivalent tasks, which are administered under the same conditions, in the same amount of time, and with scoring as objective as possible" (Hassmen and Hunt, 1994). Along this vein, arguments against the use of CR test items and performance measures based on open-ended questions point mainly toward their inability to standardize results and the greater amount of time spent in grading these types of examinations.

However, as discussed earlier, MC examinations are not the ideal form for student assessment. For example, these items do not always measure "higher-order" thinking and they may only require a student's simple recognition of the approved answer. Other types of evaluation, reliant on recall or production of a learned objective without cues, can better demonstrate the depth of understanding and applicability. Some critics even go so far as to say that "multiple-choice items favor the shrewd, nimble-witted, rapid reader, and penalize the subtle, creative, more profound individual" (Hassmen and Hunt, 1994). Due to the existence of these seemingly irreconcilable conflicts (efficiency and ease versus assessment value), the perfect type of testing may be impossible to create. Some variations on multiple-choice items have been developed, recently, and their goal is to combine the advantages of the MC and CR method into one, simple procedure.

The *Journal of Education for Business* supports the use of free-response testing techniques because they provide a "higher level test of student learning…" MC items, in contrast, provide no "intellectual 'tracks' or 'footprints' left by either the skilled or the unskilled student" (Wood, 1998). In an effort to somehow strike a balance between these

conflicting ideals (time versus practicality), the author discusses the linked multiple-choice format. Essentially, each item is constructed using a specific essay question and a linked MC question. The open-ended CR aspect of the item is small and quickly graded, and the MC portion is related to the same objective. Therefore, teachers can check for internal validity in their students' responses. If someone chose the correct option on the MC side of the item, but provided an incorrect explanation on the essay, some degree of comprehension was not attained, and the instructor can become aware of this fact. The author warns that this system is a compromise and not a replacement for CR testing procedures. "Linked multiple choice should be seen as a way of avoiding pure multiple choice in a large-class setting, not as a test form inherently superior to open-ended questions" (Wood, 1998).

Other MC critics point to the results from Advanced Placement examinations for high school students. Such tests provide a mixture of multiple-choice and essay questions for the given subject matter. Some administrators claim that "essays give a more accurate indication of originality, understanding, and thought processes" (Harris and Kerby, 1997). More importantly, however, is the assumption that certain types of people have a natural affinity for fixed-response (MC) test items, while others tend to rely on open-ended responses for expression. Inarguably, the additional time and money spent to use essay questions for standardized testing is beneficial, if for no other reason than the avoidance of misclassification of students entering collegiate and occupational settings. An experiment using MC and CR items concluded that multiple-choice scores alone predicted 46 percent of outstanding students in the field of economics accurately. Essay questions alone predicted 30 percent of the award-winning pupils. Clearly,

"neither score by itself tells the whole story" (Harris and Kerby, 1997), but a combination of the two may prove useful.

The literature points to some alternative methods of MC testing that may hopefully merge the speed of grading with the reliability of items focused on the testing of a student's ability to think critically. To accomplish this, formats are being developed which allow for more than one correct answer and partial credit for less-than-perfect responses. "Multiple-mark directions for large-scale objective tests direct a student to bubble all alternatives that are correct and leave blank all alternatives that are incorrect" (Pomplun and Omar, 1997). This multiple-mark format was selected for use in the Kansas assessments because of three main reasons. First, administrators believed that such items were more applicable to real-world situations, because the existence of one correct answer is rarely seen. By having students think about each response option and choose accordingly, they are requiring a more concentrated look at the subject matter. Second, this method is believed to partially combat the guessing bias. Finally, grading can still be achieved efficiently with the existing machine infrastructure. Researchers, however, point to many threats to validity as well, including other types of guessing biases. "Guessing may be a problem… because a student has a 50% probability of responding correctly to any alternative due to chance. [Also,] students appear to disproportionately leave alternatives blank rather than mark them" (Pomplun and Omar, 1997).

The idea behind "partial credit" for a student's response on an MC examination is made possible by the addition of "confidence levels." Essentially, students are required to self-assess their own responses. "Based on their confidence levels for each question,

partial, full, or extra credit can be awarded for right answers, and zero or negative credit (penalties) can be awarded for wrong answers" (Wisner and Wisner, 1997). The advantages behind this method of evaluation are numerous. For example, it rewards correct answers founded on high confidence, it penalizes guessing (especially if the student is largely unsure), and it provides a further incentive for the student to study and learn the class material, fully (Wisner and Wisner, 1997). Aside from this, the confidence level format is believed to be more fair and reflective of real knowledge obtained by a student or group of students. Therefore, feedback is more reliable and useful for further instruction. The disadvantages arise from the unique nature of the procedure. Teachers must learn to write these exams, while students must learn to take them. In both cases, a greater demand of time and effort is necessary. Finally, grading for confidence level tests will be more difficult, especially without the aid of computers. If these obstacles are overcome, though, few can deny the inherent strengths of the testing procedure. Self-assessment is inherently accurate, because the actual student is forced to systematically reveal his or her gaps in learning. While traditional MC items can only target very specific objectives and essay questions allow the test-taker to "write-around" the correct response, confidence levels pinpoint those areas that stimulate the student's own concept of what is known fully and what is still "a little shaky."

Optimal Number of Choices.

For any MC exam, test-writers are usually concerned with the most appropriate number of available options presented for the students. James E. Bruno and A. Dirkzwager took an information theoretic perspective and revealed that, "in general, three choices to a multiple-choice test item seem optimal" (Bruno and Dirkzwager, 1995).

Their fundamental belief is that the amount of information extracted from a test (in the form of observable knowledge exhibited by the test-taker) will increase with the number of offered choices. However, this is not a perfectly linear relationship. In other words, the "mean information per alternative… has a maximum" (Bruno and Dirkzwager, 1995), because too many alternatives for each item will obviously present a certain level of "noise." Therefore, at a particular point, the addition of another single option will have diminishing marginal returns. It is then necessary to find the optimal number of alternatives. Bruno and Dirkzwager derived a formula, expressing the amount of mean information per alternative received by the examiner as a function of the number of options, "k." This is given below in Equation 1.

$$F(k) = \left(\frac{1}{k}\right) \cdot \frac{\ln(k)}{\ln(2)} \tag{1}$$

By differentiating the function with respect to "k," the first derivative is obtained in Equation 2.

$$F'(k) = \left(\frac{-1}{k\text{^}2}\right) \cdot \left(\frac{\ln(k)}{\ln(2)}\right) + \left(\frac{1}{(k\text{^}2)\cdot(\ln(2))}\right) \tag{2}$$

Setting the resulting value equal to zero will find the maximum (Equation 3).

$$\left(\frac{-1}{k\text{^}2}\right) \cdot \left(\frac{\ln(k)}{\ln(2)}\right) + \left(\frac{1}{(k\text{^}2)\cdot(\ln(2))}\right) = 0 \longrightarrow k = e\text{^}1 \tag{3}$$

It can be seen that the optimal number of choices is approximately 2.718.

$$F(2) = 0.500 \tag{4}$$

$$F(3) = 0.528 \tag{5}$$

Through substitution of the integer values of 2 and 3 into Equation 1, three options are found to produce the most information per alternative, which is ideal (Bruno and Dirkzwager, 1995).

Another study evaluated the effects of three and four-choice MC items on high school students in Alberta, Canada. The teachers were asked to comment on whether they supported the use of three or four alternatives for each question, and their responses were overwhelmingly in favor of three-option items, citing a difficulty in "identifying a third… functional distracter for all items" (Rogers and Harley, 1999). Additionally, the results of the experiment, using both types of MC items on a mathematical examination, revealed that three-option questions were "at least equivalent" to four-option tests, with respect to internal consistency score reliability. Students were also observed to spend the same amount of time on each test, though the requirement for mathematical problem-solving (instead of recall) may have caused this phenomenon (Rogers and Harley, 1999).

The authors also point-out that three-option tests may be less susceptible to the effects of testwiseness (a student's ability to use the test or test-taking situation to receive a higher score than deserved). The proof behind this is tentative, but the presence of absurd distracters will add little value to the legitimacy of any test question. Three-option MC exams are more reflective of true, learned information if the student has minimal exposure to implausible response alternatives. Despite the number of given alternatives, test-writers should focus on MC items with the correct answer and only non-testwiseness distracters included.

Sources of Error.

Though the magnitude may be unknown, sources of error in multiple-choice testing can be identified and repaired. Bruce Walsh and Nancy Betz cite five major error sources: "time influence, test content, the test examiner or scorer, the situation in which testing occurs, and the examinee himself/herself" (Walsh and Betz, 1985). Time

influence error may stem from a mechanical (and thus unlearned) remembrance of responses given on a previous exam. This is magnified in MC testing, especially if item stems and alternatives are repeated, word-for-word. Students are more likely to recall the correct answer in this situation, without a specific cognitive review of the tested material.

Test content denotes a random sampling of items that is either unrepresentative of the student's targeted areas of study or poorly constructed for measuring the attainment of those objectives within the scope of the course material. As stated earlier, MC tests may serve to attenuate this error source, because these exams allow for a greater number of total items and a larger sampling from the population of possible questions. Proper item composition, however, is more difficult for MC exams when compared with essay formats, but specific guidelines can accommodate an appropriate construction.

The test examiner can easily be at fault, either through inappropriate proctoring or faulty grading. In addition, the testing environment itself may be different across the entire population of students. MC formats are strongly recommended as a cure for these problems. These tests are easy standardized, despite the expected disparity between different classrooms, schools, or training environments. Also, multiple-choice items will serve to remove the errors caused by subjectivity of test administration or bias among human evaluators.

Finally, the examinee has the ability to produce significant error in the process. Sickness, lack of motivation, or the desire to misrepresent oneself in the testing procedure will cause certain unknown levels of "noise" in the performance metrics. These are difficult to assess, for any type of examination. However, the student's propensity to guess is a major factor in evaluation. MC is almost exclusively victimized by this form

of bias. New and unique versions of multiple-choice tests are hoping to significantly block the effects of errors attributed to student "guesswork."

The "Guessing" Factor.

It should be evident that answering a test item correctly does not always result from direct knowledge of the tested objective. Random and educated guesswork is a major aspect of assessment, "particularly if the test item is multiple choice (Weitzman, 1996). The Rasch model is used as a means for quantifying the effects of a student's guesses. It attempts to measure the difference between the probability that a person will know the correct answer to an MC problem and the probability of getting the same item right on a given exam, attributing this expected variance to guesswork. This model cites a functional relationship between the number of test items known and answered correctly on an examination, based on the quantity of test items and the number of available options for each question.

Assuming that a student accurately reflects all of those tested objectives that are known, he or she will be forced to guess on the remaining items. Depending on the number of options from which to choose, he or she will be able to inflate the measured score by adding this source of error or bias. Obviously, the number known approaches the number of items answered correctly as the number of options increases to infinity. As more options are presented to the "unknowing" test-taker, the probability of guessing correctly diminishes. However, this model does show a significant distinction between comprehension and the applicable assessment value, lending some credibility to the advantageous nature of guesswork. In other words, as long as the number of options is reasonably low, random guesses for unknown subject areas will always increase scoring

on multiple-choice tests. It should also be recognized that the quality of available options will degrade as the quantity increases. It is very difficult to devise a great number of plausible responses from which the student should choose. Clearly, the number of options cannot approach infinity; although, a careful balance should be established which combines reasonable ease for the practitioners with a precise measurement of the student's knowledge-level. Additionally, the test taker's magnitude of confusion should be minimized, in order to obtain the true depth of subject material understanding.

Quality of Feedback.

The actual outputs of MC testing are largely unremarkable and can seldom be used as a tool for improvement within the classroom. Formative evaluation, which serves to enhance the learning environment by locating areas of poor comprehension for each student and directing further instruction, must be supported by the actual assessment feedback that testing provides. MC examination, with its R-W philosophy, reflects very little about the items answered correctly, and it reveals even less about the items that were answered incorrectly. New developments in the MC format, if they are viewed as effective instruments, must try to eliminate the need to guess while also providing more information about the resulting score, in general.

In addition, the nature of traditional objective test formats may also possess an inherent degree of non-uniform feedback, across the spectrum of attribute-groups. Certain types of students are more likely to succeed with "cut-and-dried" knowledge areas presented neatly in R-W form. MC exams are typically criticized for penalizing clever students who can see ambiguities that may bypass their duller colleagues. Also, those test-takers with a more commanding comprehension of the material may be

frustrated by the discrete and packaged nature of MC examination. Instead of presenting questions that cover a blanketed area of information, multiple-choice items have a tendency to choose certain specific (and maybe inappropriate) objectives with which to base an assessment measurement.

Finally, there have been some accusations claiming that MC exams, especially standardized tests, can be gender-specific. Males and females are known to learn and express knowledge differently. This can have especially disastrous results on the existing social structure, as college-entrance and occupational qualification exams are composed mainly of MC items. This unbalanced feedback may also spread across cultural and environmental boundaries. Obviously, any type of testing procedure must ensure that the overall quality of its reports for both formative evaluation and summative evaluation (used for institutional selection indices) must be accurate and uniform across the entire range of student-types.

Gender Discrepancies.

Evidence in support of the gender difference stems from the results of the Scholastic Aptitude Test (SAT), which is at least partly designed as a predictor of the student's freshman grade point average. Reportedly, the SAT predicts less well for women, because "even though females in general receive higher grades both in high school and college, their average score on the SAT is lower" (Hassmen and Hunt, 1994). Item bias is believed to be one of the contributing factors of this syndrome. Reportedly, questions that favor a correct response from males are more likely to occur with MC items, though conclusive evidence in this field has not been found. Teachers will usually cover and test those concepts related to their personal affinities. Because of the

specificity of multiple-choice items, students are at the mercy of the material covered on the exam, with no option to "show what they know" and expound on their own knowledge base. If the instructors is male, female students could be disadvantaged by their biased selections. (Walstad, 1997).

Another possible explanation lies in the fact that females are usually considered to be more expressive and creative and could benefit from more open-ended questions or essay items. Success in multiple-choice tests is reliant on memorization of facts or procedures that are systematically replicated. Critical thinking is not challenged in these contexts, which could explain the disparity among the genders.

Finally, the way in which females take a MC exam is shown to be different from their male counterparts. Males are less likely to change their responses, which can be advantageous with a timed exam, and it has been proposed that females are less likely to develop "test-wiseness." The ability to guess or infer the correct answer, without actual knowledge, is believed to be a "cue-specific ability that tends to develop as students pass through the grades and share information on test-taking skills" (Hassmen and Hunt, 1994). Apparently, this is more easily developed by boys.

Differential Item Functioning and Item Response Theory.

In order to improve multiple-choice testing, the way in which tests are constructed should be revised, or entirely different formats should be introduced. In support of the former idea, researchers agree that test items that may be biased in any way should be eliminated from testing environments. Differential Item Functioning (DIF) describes this bias in a less pejorative manner. "[It] suggests that items may work for different groups in positive and negative ways across the ability spectrum" (Walstad, 1997). This may

apply to any type of student or groups of students; effective DIF measurement should allow practitioners to revise MC tests in a way that is more fairly balanced for all.

Item Response Theory is used to identify DIF items, by calculating the relationship between student performance and the traits and abilities that immeasurably underlie scores on an exam. This function can be plotted by matching correct responses with abilities, characteristics, or aptitudes. The resulting graphs can then be compared among different attribute-groups, and "unbiased, or non-DIF, items will have [curves] for the two groups that substantially overlap and have the same basic shape across the ability spectrum" (Walstad, 1997). For example, consider two groups (males and females) on a given testing instrument. Item Response Theory creates a frequency chart for males, representing the number of correct responses on a particular question, across the range of grade point averages. This same chart is produced for females. DIF is only evident if the two patterns are distinguishable (they do not overlap, fully). This exhibits a tendency for one group to "scatter" correct responses differently than the other. An unbiased item would have congruent or similar frequency graphs.

Confidence-Level Examinations

As previously noted, actual format changes to traditional multiple-choice exams may provide a better means of assessment and begin to lessen the cultural and social bias associated with standardized instruments. To do this, the respondents may need extra dimensions with which to react, creating additional levels of assessment. Instead of a conventional R-W multiple-choice item, one correct answer could exist and several other options could be attached to varying levels of "incorrectness." For example, a math examination could consist of items with several options, each reflecting a certain

procedure for developing the problem, and some of these analytical paths would be more correct than others, with an appropriate scoring algorithm. This allowance for partial credit is a more evolved look at MC items; it is more reminiscent of constructed-response or essay questions, which seek finer shades of meaning within the student's response – thereby understanding more about his or her level of comprehension. Partial credit is also the basis for confidence-level examination, which relies on two dimensions of item analysis – i.e., what does the student believe is the right answer and how confident is he or she with that particular choice?

There are three main reasons for applying this self-assessment method in MC examinations. First, it will allow the testing instrument a higher level of accuracy and scope in measuring the knowledge of the test taker. Second, students will be rewarded for elevated levels of comprehension and assurance – guesswork will not be masked. Finally, the quality of feedback is much higher. Students and teachers will be able to point to specific areas or objectives that were not completely learned, and those areas that were confidently missed can be more effectively identified and fixed.

Early proponents to this format advocated a system whereby the student provided his or her own level of confidence (based on a percentage scale). Scoring was therefore achieved by adding the probabilities, which reflected correct answers, and subtracting the probabilities associated with those items that were answered incorrectly. Hopefully, correct answers were matched with high probabilities (the student was confident with the right answer) and incorrect responses were given low probabilities (the student was wrong, but he or she knew it). Guessing was eliminated, because it was no longer rewarded, and instructors had a better idea of the level of attainment for each of the

desired objectives. Also, teachers were able to see if some parts of the material were learned incorrectly – high confidence for wrong answers. However, certain questions in this area must be covered. First, "is one not measuring two different factors: knowledge and (self) confidence? And are subjects able to report their probabilities correctly?" (Dirkzwager, 1996). For younger students unversed on the principles of percentages and elementary statistics, these questions are definitely valid. Also, this type of exam must find ways to guard against the "confidence-bias." Students with higher (and perhaps artificial or unmerited) levels of assurance should not perform differently than those with more realistic viewpoints. Dr. Bruno's concept of Information Referenced Testing (IRT) is designed to "sidestep" both of these pitfalls and combat all of the previously discussed problems with MC exams.

Information Referenced Testing

Dr. James E. Bruno of the University of California at Los Angeles (UCLA) devised the IRT concept in multiple-choice exams as a way to develop and enhance school assessment techniques in three main areas. First, a two-dimensional response format allows the student to indicate partial knowledge in a particular area. Conventional tests reveal zero or full comprehension on a given MC item, and this can produce significant errors. Second, the level of feedback is comparably robust. The result is a more valuable type of formative evaluation (FE), which identifies the gaps in classroom learning and contributes to the educational process as the student progresses through the material. Summative evaluation, in contrast, merely provides a final grade or score, with no opportunity for correction or improvement. Finally, this FE is based on an "information referenced metric," which cites additional constructs, to include

"misinformed," "uninformed," "partially informed," and "fully informed." The goal is to produce an assessment measurement that will transcend traditional views on "Right-Wrong" evaluation and provide outputs that are much more meaningful to all of the interested parties.

IRT supports policy issues in education that are dedicated to measuring the effectiveness of the schooling process and examining ways to use feedback to provide better classroom instruction. One of the more traditional alternatives, norm-referenced testing, scores a student's exam and normalizes the results in a standard curve, relative to peer performance. Another conventional method, criterion referenced testing, bases a given score on the number of right and wrong responses. Information Referenced Testing, as the name implies, attempts to produce a score that is more indicative of the percentage of information learned, which is the ultimate educational goal. Also, each graded item reveals more about what the student has comprehended and how confident he or she is with the material. In this way, the purposes behind testing are completely shifted from summative evaluation (sorting, selection, inclusion, and exclusion) to formative evaluation (identification and improvement).

Dr. Bruno believes that regular MC test items can be ambiguous and will sometimes promote guessing. In addition, the feedback provided for continuous learning is limited, and the outputs do not indicate whether a student is misinformed, partially informed, uninformed, or fully informed. To combat this, his IRT formats allow for response variables that permit a student to garner some credit, even if he or she is not completely sure or doesn't know at all. For example, if a student is so inclined, he or she can respond with "I don't know," resulting in no points (and no penalty). If a student

would like to narrow the MC choices (from three to two), partial credit will be given for indicating partially correct knowledge. Full credit is given for full knowledge (the student chooses one response from all of the choices given). Finally, a loss of points will occur if confidence is shown with an incorrect response.

Dr. Bruno's assertion is that guessing on a traditional exam will overcorrect for middle and low-achieving students, placing these scores unfairly in-line with the more-informed students. "The R-W procedure encourages guessing behavior that typically results in an over-assessment of subject matter mastery" (Bruno, 1988). IRT is designed to remove the guessing bias. In essence, guessing on these confidence-level exams will be detrimental to the maximization of scoring, because some positive credit is given for partially correct responses and no point penalties are assessed for an admission of "I don't know."

"If you work out the mathematics, the expected values should indicate that the best overall score is when you don't overvalue your information or guess" (Bruno, 2002). In other words, students should honestly respond to the variables. If, for example, a test-taker can narrow three MC options to two, removing one as a definite incorrect response, yet he or she cannot confidently choose between the remaining alternatives, it would not be beneficial to guess. The ramifications for an incorrect answer (confidently asserted, but wrong) would be more severe than to choose a partial-credit variable.

"There is considerable research evidence that when students guess, they are likely to have at least some information (partial knowledge) that allows them to eliminate some alternatives as incorrect, thus improving their chances of guessing correctly" (Bruno, 1986). This means that a typical MC exam, with this guessing-syndrome in place, will

hide those variables which were systematically (and correctly) eliminated. Essentially, if this same student went on to make an incorrect response, a R-W evaluation would belie the fact that the student had accomplished a correct omission in the process of answering (partial knowledge). Conversely, there may have been no guesswork involved at all, and the student's wrong choice was confidently answered incorrectly (misinformation). Lastly, a guess may have resulted in a correct response, but the student had not learned anything related to the objective (no information). These masked factors of learning are supposedly uncovered in an IRT format, and the nature of the scoring algorithm should motivate students to use all of the confidence-level variables to the best of their ability.

Of course, student testing in this area has its problems. The uniqueness of the format will require sufficient instruction and training, and grading by hand can be time and effort-intensive. "The biggest problem is that it is different and requires a computer to score. Students need to become familiar with the process. Once they are oriented… the feedback reports can be an excellent way to support instruction" (Bruno, 2003).

The Mechanics of IRT.

The system that was used in the subsequent experimental design attempts to use the IRT concept in a simple and easy-to-understand manner, while still providing those two-dimensional variables necessary to reflect the level of student confidence. The detailed features of this format are presented here.

Specifically, each MC item consists of a question (stem) and three alternatives, only one of which is correct. These *a*, *b*, and *c* options are labeled as the "first column" choices. Additionally, options *d*, *e*, and *f* give the student an opportunity to exclude one of the original three alternatives by choosing from "*a* or *b*," "*a* or *c*," and "*b* or *c*,"

respectively. The point values associated with a correct response from these "second column" choices is worth one-half of the item's full value. Because the student is not forced to randomly choose between the two remaining alternatives, he or she is conceding 50 percent of the credit. Finally the "third column" consists of *g* or "I don't know" and admits a total lack of knowledge or comprehension. Credit here should reflect the expected value of a guess between three options, resulting in one-third of the item's original worth. Admittedly, scoring for these variables is subjective and dependent on the instructor or department's own viewpoint. Granting point values in excess of the expected values for random guesswork may provide the needed incentive to persuade students to use the second and third-column alternatives. This is advantageous to the students, because scores will be higher, and the instructors will benefit from the strength of the assessment feedback.

The primary use of these confidence-level variables is to gain an understanding of four main constructs. If the student chooses the correct option, with full confidence, he or she is "fully informed." If however, the correct option is selected, but some level of doubt is exhibited, he or she is "partially informed." The "I don't know" option, if chosen by the test-taker, denotes a total lack of confidence, and this reflects an "uninformed" state of comprehension. Finally, an incorrect response suggests that the student is "misinformed." These constructs, if accurately calculated and presented to students and instructors, can reveal the gaps in classroom learning and point-out the necessary steps for extra instruction in critically deficient areas.

While classical objective test formats uncover student attributes related to being "correct" and "incorrect," the IRT model expressed here provides for the more precise

characterizations of "informed" and "uninformed," while also deriving two additional metrics for assessment – "partially informed" and "misinformed." The following table summarizes the factors associated with these concepts:

Table 1. Summary of IRT Experimental Model

Student Action	Root Cause	Observable Effect	Required Follow-up
Chooses correct option from first column; receives full credit.	Student confidently comprehended the objective.	Student is "fully informed."	None.
Chooses correct option from second column, with 2-D variables; receives partial credit.	Student is not confident or comprehends part of the objective.	Student is "partially informed."	Cover the material again, fully and gain confidence.
Chooses incorrect option from first or second column; receives zero credit.	Student is confident, but wrong.	Student is "misinformed."	Re-evaluate learning; use alternative methods of instruction to correct the problem.
Chooses "I don't know;" receives minimal credit.	Student cannot answer the test item.	Student is "uninformed."	Adjust the scope of instruction and study to "fill in the gaps."

Despite its proposed strengths, this unique format may have the propensity to confuse and intimidate those persons without at least a basic level of exposure. Indeed, successful implementation in the classroom and training environment hinges on a complete understanding of the methodology and scoring techniques. If, however, the student is well-versed on the directions and the procedures for grading, he or she should not be handicapped by time or general confusion. Indeed, quite the opposite is true; the test-taker has every opportunity to exhibit his or her knowledge level in a practical sense, while still maximizing the scoring potential. The key, as with any introductory

assessment model, relies on proper training and guidance in the beginning stages of implementation.

Using the "I don't know" variable can also be applied to those examinations with true-false and fill-in-the-blank items. Conventional true-false questions allow for the greatest fundamental error in assessment, as students and trainees are subjected to minimal risk of answering incorrectly, even if random guesses are employed. By using this third option of "I don't know," the uninformed test-taker will hopefully be able to show a lack of knowledge in the particular objective tested, while still earning a comparable score. Presumably, this score will not penalize the student, and feedback will be enhanced. In addition, an instructor might give the student some freedom on a constructed-response item by allowing unfilled blanks or an admission of "I don't know" for some partial credit. Both of these techniques, along with the multiple-choice format described above, can help differentiate an "uninformed" student from a "misinformed" student. This distinction can be crucial in many cases, especially within the training environment.

IRT on the Web.

Dr. Michael W. Klymkowsky, in his essay entitled "The Evolution of Biology Teaching and the Web," addresses the need for changes in the way that educational material is taught. His ideas still adhere to those long-standing approaches to instruction, such as the Socratic method, "i.e., working directly with the material to be learned, with access to an open, encouraging and competent instructor, either in a one-to-one setting or within a small group" (2002). However, he believes that other methods are predominantly used today, usually centering around monologue-type lectures and little

student involvement. His answer to this Aristotelian method of learning is a careful combination of web- and classroom-based teaching. Stated simply, if a student is given an opportunity to take part in an interactive computer module, with the upcoming lesson's notes and investigative questions presented before the start of class, classroom time will be better spent. Of course, motivation for this type of pre-learning should be governed (and enforced) by quizzes incorporated within the online program. This is perhaps a more practical answer to pre-class and pop-quiz examinations, because the instruments are graded immediately, and the instructor is allowed complete and accurate feedback, before he or she begins instruction.

To facilitate this program in his own educational venue, Dr. Klymkowsky has employed "Knowledge Factor, Inc.," a company designed to provide web-based platforms for universities and corporations focused on developing technology-based programs for educational and training assessment. "Knowledge Factor" has worked with Dr. Bruno and UCLA, and it has gained expertise in the area of confidence-level testing. IRT can, in fact, be fully developed through this company's systems, and the feedback reports can be customized to instantaneously target the gaps in learning, as students and trainees are exposed to their programs of study. "The traditional approach in organizations is to train and then assess or evaluate the training. ['Knowledge Factor's'] approach is to assess first [and] tailor the training to specifically address information gaps in the learner" (Goel, 2003).

Identification of areas of "misinformation" is perhaps most critical, especially in the corporate operational environment. Obviously, personnel that exhibit confidence in the wrong principles of doing business can have "serious implications with regard to

safety, security, liability, and the bottom line" (Goel, 2003). IRT, in this respect, is extremely beneficial, because it can distinguish those areas which contribute to partial information, misinformation, and a total lack of any discernible knowledge.

Dr. Klymkowsky is himself a professor of Cellular and Developmental Biology at the University of Colorado, Boulder, and his experience has yielded an understanding in the student's psychology. In a sense, each classroom attendee is involved in a type of game theory, fully aware that "coming to class 'unprepared' is not fatal, and often not even mildly unpleasant. Indeed, there are rarely any negative repercussions, [and] teachers have adopted their teaching style to this 'fact of life.'" This lack of readiness in the classroom (or anywhere) is indeed a naturally resulting phenomenon. It is an expected survival technique that is employed by those in the learning occupation – they will prioritize their time and effort to reach a maximum benefit. Dr. Klymkowsky hopes to use the web as a supporting instructional tool to avoid a "cold" classroom, partial learning, and "cramming" for important examinations. His approach has many advantages. For example, the instructor is aware of students who have not completed a module's set of questions. And, in the event that a student fails the online quiz, it must be retaken, and different items (reflecting the same material) can be substituted. Most importantly, however, "the tracking system generates a report for the instructor as to which modules and which questions were the most difficult for students [and] can then tailor the in-class instruction to deal with those concepts that are most difficult for the majority of the class" (Klymkowsky, 2002).

As for the implementation of confidence-variables, Dr. Klymkowsky uses IRT-related multiple-choice and true-false questions to quiz the students. For each item, they

must provide an answer and a level of assurance, in the form of "absolutely sure," "kinda sure," or "just guessing." Scores are appropriately matched with these selections, giving the instructor a better view of the class's overall knowledge in specific areas.

Summary and Conclusion

This chapter focused on the principles of assessment, beginning with a brief history and proceeding toward a discussion of the various techniques for student evaluation. The benefits and drawbacks of MC examination, specifically, were analyzed, along with the major issues associated with this format. Confidence-level items were offered as a means of correcting some of the inherent problems of multiple-choice questions, and the concept of IRT was eventually introduced. Dr. Bruno's testing philosophy was dissected and applied to the "Information-Referenced" model. The mechanics of IRT were presented, with a cursory look at the construction and proposed efficacy of these items and their component-parts. Appendices D and E provide samples of the MC and true-false questions used in the experiment. Students within the Management 210 and Biology 331 courses at the United States Air Force Academy (USAFA) were exposed to these specific problems (along with others) in a real examination-environment.

Subsequent chapters will turn the attention of the thesis to the actual research conducted at USAFA. With the background behind IRT already presented, it is necessary and appropriate to subject this assessment model to Air Force students in a true educational setting. The methodology behind the experiment is discussed next, with results, analysis, and a final discussion to follow.

III. Methodology

Background and Overview

In light of the theoretical benefits of Information Referenced Testing (IRT), the practical advantages are not immediately evident and will have to be explored through experimentation. This extension of the IRT ideal in a "laboratory" environment should provide some added visibility within the given problem statement and research question, by resolving the appropriate investigative issues through data collection, surveys, interviews, and specific hypothesis testing.

The following chapter will outline the procedures used in the experimental method, including a view of the academic setting, subjects, facilitators, instruments, tools for analysis, assumptions, and constraints. This section will then document the specific processes and analyses that will hopefully answer the five investigative questions posed in Chapter I and reveal the evidence necessary to refute or support the existing hypotheses on the efficacy of IRT.

Experimental Design

For the purposes of this research study, experimentation was conducted solely within the academic confines of the United States Air Force Academy (USAFA), in Colorado. Students enrolled in USAFA's core management course, Management 210, were used as the subjects for the main experimental method. The quantitative and qualitative data harvested from these "cadets," as well as their instructors and administrators, will provide the primary basis for answering the investigative questions and resolving the research problem.

The management students were exposed to two types of testing procedures. One examination format consisted of traditional MC items, and IRT items were used exclusively on another exam. Both test types were supplemented with the same "constructed response" or essay questions. Two major examinations, known as graded reviews (GR's), were administered in this experiment. For the first GR, the "experimental group" was tested with the IRT variables, while the "control group" was exposed to conventional MC test items. For the second graded review, both groups were tested with an identical exam, constructed solely of traditional MC items and appropriately matched essay questions. In this manner, the experimental group students were tested using IRT on GR #1 and conventional items on GR #2, while the control group students were treated with standard MC items for both examinations.

Table 2. Experimental Design

Group	GR #1	GR #2
Experimental	X	O
Control	O	O

Table 2 shows a visual representation of the experimental process: an "X" denotes a testing treatment using IRT variables, while the "O's" represent traditional exams with standard MC items. This design allowed for two different comparisons. The relative performances on GR #1 could be observed, with two groups of students – each using different exam formats. Additionally, a difference in scores could be measured for the same group of students over time, as each was allowed to move from GR #1 to GR #2. For the experimental group, this testing migration revealed a comparison between IRT and traditional testing formats, for the same set of students. For the control group, scores for the two graded reviews provided a normalized standard that could be compared

with the experimental sections. These directly observable phenomena were then used to shed some light on the real-world ramifications of IRT implementation.

Subjects

The experimental method was conducted using USAFA students, all of whom were enrolled in Management 210 for the fall semester, 2002. 235 students were assigned to the experimental group, and 241 students were assigned to the control group, for a total of 476 cadets sampled out of a population of approximately 4,000. Fourteen sections were randomly classified into the control group, and fourteen sections were (again) randomly defined as the experimental group.

Before experimentation began, it was assumed that the two groups consisted of similar "types" of students, each sample sharing equivalent attributes and aptitudes, when taken in the aggregate. USAFA databases were able to produce characteristics for each student, for the purposes of comparison between the two groups. With respect to student attributes, each subject was identified and sorted into specific categories under the broader headings of gender, race, departmental major, and class year. Additionally, those students classified as a Management major were counted and compared between the control and experimental groups. Of course, both genders were represented, and all of the categorized races (limited to Asian, Black, Caucasian, and Hispanic) constituted about 97% of the entire sample. Departmental majors included Engineering, Humanities, Social Science, and Basic Science disciplines. And class year was narrowed to the Class of 2004 (Juniors) and the Class of 2005 (Sophomores). Again, this represented the overwhelming majority (in excess of 99 percent) of sampled cadets. Below is a tabular breakdown of these attributes for both groups.

Table 3. Control Group Attributes

Attribute	Population Proportion	Population Size
Female	0.151	239
Male	0.849	239
Asian	0.065	232
Black	0.043	232
Caucasian	0.802	232
Hispanic	0.091	232
Engineering Major	0.308	234
Humanities Major	0.094	234
Social Science Major	0.457	234
Basic Science Major	0.141	234
Management Major	0.222	234
Class of 2004	0.129	241
Class of 2005	0.863	241

Table 4. Experimental Group Attributes

Attribute	Population Proportion	Population Size
Female	0.188	234
Male	0.812	234
Asian	0.035	228
Black	0.070	228
Caucasian	0.855	228
Hispanic	0.039	228
Engineering Major	0.329	231
Humanities Major	0.117	231
Social Science Major	0.390	231
Basic Science Major	0.165	231
Management Major	0.190	231
Class of 2004	0.166	235
Class of 2005	0.826	235

Given the above values, it is appropriate to look at the proportion differences for each group and see if a significant difference is discernible. Because the sampled

population sizes were all large (greater than 40), a large-sample test procedure for differences between population proportions was used (Devore, 2000). Hypothesis testing with a two-tailed rejection region was employed as a means to find the significant differences in proportions, if any, with an alpha value of 0.05. The calculated P-values and results of the test are given below.

Table 5. Attribute Comparisons for Control and Experimental Groups

Attribute	Proportion Difference (Absolute)	P-Value (Two-Tail)	Significant Difference? ($\alpha = 0.05$)
Female	0.037	0.28327	No
Male	0.037	0.28327	No
Asian	0.030	0.13994	No
Black	0.027	0.20984	No
Caucasian	0.054	0.13181	No
Hispanic	0.051	0.02388	Yes
Engineering Major	0.021	0.62697	No
Humanities Major	0.023	0.41909	No
Social Science Major	0.068	0.14374	No
Basic Science Major	0.023	0.47214	No
Management Major	0.032	0.39421	No
Class of 2004	0.037	0.25470	No
Class of 2005	0.038	0.26528	No

Here, it was evident that no significant difference in attribute proportions existed between the two groups, with the exception of the number of represented Hispanics. This is largely due to the relatively small number of documented Hispanics in each sample (21 in the control group and 9 in the experimental group), and it can be assumed that this had no detrimental effects on the experimental results and analysis.

Student aptitudes were also considered. Information was extracted from USAFA databases, revealing each cadet's individual assessment in academics and military bearing. Cumulative grade point averages (GPA's) acted as a measure of each student's

undergraduate performance in course work offered only at USAFA. An Academic Composite Average (ACA) revealed the combined score (calculated using an unknown function) of college entrance exams taken by each cadet before admittance into the Academy. Military Point Averages (MPA's) were roughly indicative of individual, "professional" performance within the cadet wing. Though not strictly defined as an "aptitude," age was included within this table, because its value is averaged on a continuous scale, in line with GPA, MPA, and ACA Score. The tables below give the mean, standard deviation, and sample (population) sizes for the control and experimental groups, respectively.

Table 6. Control Group Aptitudes

Aptitude	Mean	Standard Deviation	Population Size
Cum GPA	2.66	0.52	241
Cum MPA	2.737	0.315	239
ACA Score	3168	310	238
Age	19.78	1.01	238

Table 7. Experimental Group Aptitudes

Aptitude	Mean	Standard Deviation	Population Size
Cum GPA	2.77	0.51	235
Cum MPA	2.775	0.351	234
ACA Score	3193	322	233
Age	19.70	0.85	233

It appeared that all of the aptitudes were approximately equal for the two groups, with the exception of cumulative GPA, which exhibited a much higher average for the experimental group. The table below is a check for significant differences. Designed in the same manner as Table 5, it uses a large-sample test statistic and a standard Z-curve to

obtain a two-tailed P-value and a test for significant difference, with an alpha level of 0.05 (Devore, 2000).

Table 8. Aptitude Comparisons for Control and Experimental Groups

Aptitude:	Mean Difference (Absolute)	P-Value (Two-Tail)	Significant Difference? ($\alpha = 0.05$)
Cum GPA	0.11	0.01981	Yes
Cum MPA	0.038	0.21572	No
ACA Score	25	0.39089	No
Age	0.08	0.35185	No

Disregarding cumulative GPA, it appeared that aptitudes between the groups were extremely similar. The relative difference in college grade point averages, however, was apparent. The subsequent presentation and analysis of experimental results will have to somehow account for this discrepancy.

Facilitators

At the time of the research study, all of the instructors in the Management 210 course were assigned to the Air Force Academy as full-time professors, acting without the support of graduate assistants. The level of experience for each instructor ranged from Captain to Lieutenant Colonel and also included non-military personnel. All of these facilitators were acting under the guidance of a single course director and a standardized lesson plan. Each instructor personally administered the graded reviews for his or her own sections. All of the MC items (IRT and traditional) for both groups and both GR's were graded, scored, and recorded by a single person – the researcher. Each individual instructor was responsible for grading essay questions for his or her own students, though a course-wide solution key and grading scale were employed, to ensure standardization.

For each of the samples, however, there was a minor discrepancy in the facilitation of classroom instruction. In other words, the control group did not have the exact same teachers as the experimental group. Although some of the instructors taught within both samples, the Management 210 faculty was randomly assigned to all of the 28 sections – some instructors taught only within the experimental group and some taught solely within the control group. The table below diagrams the specific instructors within each group.

Table 9. Instructors for Control Group

Instructor	Average Cumulative GPA	Standard Deviation	Number of Sections	Population Size
Instructor A	2.64	0.56	1	21
Instructor B	2.52	0.53	4	68
Instructor C	2.50	0.45	2	32
Instructor D	2.83	0.52	1	17
Instructor E	2.76	0.41	1	17
Instructor F	2.85	0.54	3	53
Instructor G	2.67	0.48	2	33

Table 10. Instructors for Experimental Group

Instructor	Average Cumulative GPA	Standard Deviation	Number of Sections	Population Size
Instructor C	2.82	0.56	2	33
Instructor D	2.84	0.55	2	39
Instructor F	2.80	0.46	1	15
Instructor G	2.63	0.56	1	13
Instructor H	2.70	0.45	4	72
Instructor I	2.80	0.53	4	63

The control and experimental groups did possess four instructors in common (C, D, F, and G). Part of any future analysis will have to involve an examination of student

performance, strictly isolated to the cadets taught by these individuals. It is important to note that cadet GPA was relatively similar for three of the four common instructors. Instructor C had a large disparity in mean GPA between his control and experimental students (2.50 versus 2.82, respectively). Instructors D, F, and G each had student samples with a negligible difference in grade point averages, across the two groups.

Experimental Instruments

The instruments used for the purpose of this experiment were the testing formats administered to the USAFA students. For GR #1, the IRT and traditional exams were administered at the same time, to ensure against the risk of sample contamination. Aside from the confidence variables presented to the experimental group, GR #1 was identical for the two groups. The MC and essay questions were stated in exactly the same manner and covered the same course objectives. The available options were uniform for each version of the test. However, as stated, the confidence-level variables were appropriately included in the IRT exam. Please review Appendix D for examples of the "Information-Referenced" items used in this portion of the experiment. GR #2 consisted of the same testing instrument for both groups, as indicated in the experimental design section, above.

The control group was tested using a traditional multiple-choice test, with three options per test item: *a*, *b*, and *c*. The students were only allowed to choose from these alternatives, and guessing was encouraged, as an unanswered item was given zero credit. The experimental group was given a confidence-level test on the first GR. It was composed of the same test items and available options – *a*, *b*, and *c* were identical to the items given to the control group. However, options *d* (*a* or *b*), *e* (*a* or *c*), *f* (*b* or *c*) and *g* ("I don't know") were available to facilitate two-dimensional assessment. If a student

chose an option that reflected two possible alternatives and one of the two alternatives was the correct option, the student received half-credit. If a student chose option *g* ("I don't know"), he or she was automatically given one-fifth credit. Of course, a correct option chosen from *a*, *b*, or *c* was rewarded with full credit, and any type of incorrect response was awarded with no points. The students were briefed on the specifics of the test procedure and scoring methods.

Tools for Analysis

Results from the experiment were used for comparisons between the control and experimental groups, for both graded reviews. Additionally, linear regression and a correlation analysis linked test performance with certain student aptitudes and attributes. The objective was to build a model that related a dependent variable to more than one independent variables. In this case, the dependent variable was defined as test performance (traditional and confidence- level scores), and the goal was to recognize each independent variable as a viable predictor. Student aptitudes and attributes included in the model have already been mentioned above. Statistical software helped determine the extent to which a fit model actually existed.

Assumptions and Limitations

Due to the strict control governed over the procedures of this experiment, assumptions dealt primarily with human behavior and psychology in the test-taking process and other administrative areas. First, it was assumed that teaching styles were relatively uniform across the two groups. This is supported by the fact that all of the lesson plans were standardized, with set objectives that were established before the start of the semester. Also, the tests were all created by the same author, and the format was

identical between both groups, except for the presence of confidence level variables provided for the experimental students on GR #1. As a way to strengthen this assumption through the analysis, some comparisons were made for those students instructed by professors that were common for both sections.

Other assumptions must be considered, on behalf of the students, themselves. It should be understood that the cadets were not interested in skewing the results of the study. In other words, maximization of individual performance was desired by each subject. This is evidently true, because the graded reviews represented major mid-term examinations, contributing to a large portion of the overall grade in the course. Additionally, in order for the experiment to be valid, students had to possess complete understanding of the test directions. For the standard MC exams, this was not a problem. However, the procedures in effect for Information Referenced Testing could have been slightly more complicated, especially considering the unique quality of the task. Researchers had to assume that students and instructors had complete understanding of the directions and apportionment of points, in order to ensure completely valid results. The instructions were, in fact, well-designed, with an applicable example provided for the cadets.

Scoring, on the other hand, could have created some relevant classroom issues that may have inhibited a sound analysis. Considering the three-option MC items administered to all of the students, the "I don't know" alternative should have reflected a credit of one-third (or higher) of the particular question's total value, in order to provide fairness in scoring and give the cadets an incentive to choose this option. This experiment, however, only provided one-fifth credit for a selection of *g* ("I don't know").

Test administrators and researchers must assume that the particular students tested had an awareness of probabilities – thereby minimizing the selection of "I don't know," because the awarded point values were less than the expected value of answering correctly through random guesswork. The point values for the remaining confidence variables, however, were in line with the appropriate probabilities.

It should be realistically understood that any student's actual level of comprehension is immeasurable and unknown, even to the test-taker. The role of an examination is to provide an instrument for reflecting this unknown value into a quantifiable result, and this effect will never be exactly true. A myriad of constructs are under scrutiny. For example, the general health and state-of-mind of the test-taker is measured – mental and physiological conditions are manifested on most exams, as well as personal comfort, seriousness of distractions, pressures to perform, and stresses outside of the class, to name only a few. Simply stated, a test-procedure is not a vacuum, and an innumerable amount of factors will contaminate the results. Perhaps the greatest assumption was that this real-world academic classroom conformed to "laboratory" experimental rules, despite the subjective nature of testing in an atmosphere full of "unknowns." Regardless of all the controls exhibited in the research project, this limitation was always present.

Threats to Validity

For GR #1, the IRT and traditional examinations were administered simultaneously. This allowed for internal validity with respect to history, maturation, and mortality, because no time was allowed to elapse between test offerings. Also, the exams used for both groups on the first graded review were identical in every aspect, with the

exception of the presence of confidence variables; therefore, testing validity was controlled. As for regression and selection, these threats to internal validity were combated by a random assignment of student sections to the control and experimental groups. Finally, interaction between the groups, for the first exam, was obviously prohibited and subsequently guarded in the proctoring.

Instrumentation may have been a factor in this comparison, because different instructors were present in each group. This creates a possible lack of standardization in teaching styles, presentation of material, assistance in learning, and instructions during the exam. As a means for controlling this, the professors were surveyed on their respective teaching techniques and the quality and quantity of information presented to the students in preparation for the GR. Scoring for the MC questions was uniformly conducted, as only one person was allowed to grade all of the sections. Essays were scored by the individual teachers, with the help of course-approved solutions.

The second comparison occurred as the experimental group was tested over both GR's. This required a more "perilous" examination of internal validity. History, maturation, and instrumentation may have been facilitated by the influence of time and events between the graded reviews, which would have caused some unexplained "noise." Experimental mortality was a possibility, as well, due to the inevitable occurrence of students dropping a course after the first exam. The contamination of samples through subject interaction was likely to have occurred, as well, because students will often discuss an exam, after the fact. Finally, and perhaps most importantly, testing itself was a huge factor for consideration, because GR #1 and GR #2 were completely different

examinations. Regression and selection were not subject to internal validity, again owing to the random appropriation of students for the two groups.

In order to control all of these perceived threats, it was imperative to observe the control students as they were administered both graded reviews. Their results provided a measurable and expected change in performance that was compared with the performance of the experimental students. Only through the use of this control section "barometer" could any noticeable effects among the experimental section (as it was exposed to both formats) be considered valid. This allowed for a stronger interpretation of results. Table 11 is a summary of these threats, as they were applied to the experiment. A plus-sign indicates that the source of invalidity was successfully controlled, while a question mark means that a possibility for concern remained.

Table 11. Internal Sources of Invalidity

Threats	Comparison between Control and Experimental Groups for GR #1	Comparison between GR #1 and GR #2 for Experimental Group
History	+	+
Maturation	+	+
Testing	+	+
Instrumentation	?	?
Regression	+	+
Selection	+	+
Mortality	+	?
Interaction	+	?

External validity is governed by three principles: the existence of a real-life setting within the context of the experiment, the representative nature of the sample, and the repeatability of the study in a different environment. For this scenario, reality was not a major problem. In fact, the subjects were reliably genuine, and the testing procedure

was evaluated under the control of an operational university course curriculum. The instructors and administrators "experimented" with IRT in real-life classrooms under standard conditions, and the assessed results of the examinations were contributory toward student grades. The sample, however, could not have been considered representative of the world's population of students and trainees. The USAFA cadets occupied a very narrow spectrum of human characteristics. The ages of the subjects were relatively confined to the range of 19 to 21, all of the students were standout high school graduates, and a diversity of gender and race was not present. United States Air Force (USAF) personnel were not represented by this sample. The third element of external validity (repeatability) was maintained throughout the experiment. The details of this project were uncomplicated and could easily be replicated in almost any environment, regardless of existing computer infrastructure or technical support.

Methodology Behind Investigative Question #1

The first investigative question is concerned with a difference in scoring performance that may be present, if IRT variables are introduced to students. To follow this, it is important to know the specific factors leading to a significant contrast, if it is found that confidence-level exams had some noticeable effect. The following procedures were used to check for (and analyze) this possible phenomenon.

In order to test for a significant difference in scoring, it was necessary to look at the Management 210 students and compare overall performances for the entire control and experimental groups. For GR #1, a large-sample test based on two samples revealed the presence or absence of a statistical difference. This was measured through the use of a test statistic Z, the standard normal distribution, sample means, and standard deviations

derived from both groups. The null hypothesis stated that the difference in population means was zero. The alternate hypothesis asserted that a difference in population means did, in fact, exist. The null hypothesis was rejected at any reasonable alpha value (Devore, 2000). P-values for the two-tailed test gave a more definite resolution for the satisfaction of this investigative question.

Assuming that a difference did exist, the next step was an investigation of the individual student parameters that may have caused the contrast in scores. To complete this test, the cadets in both groups were broken down into specific categories, based on certain traits and relative academic strengths. The mean scores and standard deviations were compiled for all of the given attribute- and aptitude-groups (for both GR's), and a check for statistical difference in performance between the two testing methods was made for each "type" of student. The use of small-sample tests was needed for some of these comparisons. To check those groups with sample sizes of 40 or less, the t-distribution sufficed – based on degrees of freedom, ν, estimated from the data (Devore, 2000). The results of these evaluations documented those specific student aptitudes and attributes which caused the significant difference in overall scores. Also, those human characteristics not marked by a recognizable disparity in performance were noticed and assumed to be unaffected by IRT variables.

Methodology Behind Investigative Question #2

The second question asks for some explanation of how students with varying aptitudes and attributes will respond to these newly-designed (IRT) test questions. Obviously, certain types of students will exhibit noticeably higher or lower scores, if asked to respond with confidence-levels. It is important to find those specific traits

which dictate performance on IRT examinations and compare the results with traditional tests administered in the same type of environment.

A correlation analysis was appropriate here, whereby a model was built, linking a dependent variable (test performance) with all of the given aptitudes and attributes used in the previous analyses. This regression technique sought the strongest relationship possible through the "step-wise" removal of independent variables. Those characteristics that were found to show the most robust correlation were retained by the model and judged to have a comparatively significant effect on IRT scores. This process was enacted upon control group performance on traditional MC exams, and the results from both studies were subsequently compared.

Methodology Behind Investigative Question #3

Aside from the quantitative ramifications of IRT, it is obviously imperative to gauge the level of acceptance that students and teachers attach to this new method. Information Referenced Testing was designed to enhance the educational relationship between instructors and pupils and provide for more accurate assessment metrics to benefit both parties. Therefore, opinions taken from the cadets and Management 210 professors were necessary to add value to the study.

This question was answered most appropriately by interviews and personal testimonies from the teachers and administrators involved in the processes. In addition, a type of pilot study was completed by USAFA's Biology Department. Biology 331 (Botany) students were given two graded reviews, using confidence-level items on a small portion (roughly 36%) of the MC and true-false items. The students were then surveyed and asked to respond critically about the IRT-type questions. Data was

collected and compiled to see if the cadets believed the two-dimensional questions to be more effective as an assessment format, relative to standard test questions. They were also queried with respect to the inherent difficulty and cumbersome nature of the IRT portion of the exam. The Botany students were divided into control and experimental sections (as with the Management students), though sample sizes were small, and surveys were conducted for all of the cadets, after each exam. The groups were switched between the first graded review and the second, to ensure fairness and the reliability of results.

Methodology Behind Investigative Question #4

The kinds of processes that are occurring "behind the scenes" for any type of academic exercise are always an appropriate consideration. The level of human effort and expense required to administer tests to the students, score the results, and provide feedback cannot be ignored. IRT, with its unique construction and added dimensions, should be assessed and evaluated by the types of process issues (and the needed infrastructure) that will accompany its implementation in the classroom.

Again, administrators, instructors, and students provided some input toward answering this question. The Management 210 Course Director was interviewed extensively, as he was directly involved in the entire process, and a transcript of his testimony was included in the results. The researcher himself, was given a venue for personal observations, as he was present during the first graded review for both the Management and Biology sections. Experts in the field were also tasked for individual comments. The Director of Academic Assessment for the Air Force Academy's "Center

for Educational Excellence," provided valuable suggestions for improved exercise and control of IRT in future endeavors.

Methodology Behind Investigative Question #5

Finally, the investigation would not be complete without seeking the "Holy Grail" of assessment – can IRT accurately reflect the percentage of information learned? The answer to this question is the overriding purpose behind the experimental study. In an attempt to find the solution, an item-by-item analysis was conducted for the Management 210 test results on GR #1. The 15 multiple-choice questions were individually scored, for each of the 28 sections within the control and the experimental groups. This laborious process was followed-up with another instructor-survey. The Management professors in both groups were asked to determine the amount (and quality) of classroom lecture used to emphasize each of the objectives that was tested in the MC portion of GR #1.

Given these two metrics (mean item performance and average item-objective coverage in the classroom and other learning laboratories), the results were subjected to a correlation analysis. Assuming that the increased quality of instruction for a certain learning objective will result in higher recognition of that correct objective on an MC exam, the IRT and control models were both checked for a positive linear relationship. A more observable, direct correlation between the two methods provided a determination of which format was more successful at reflecting the amount of information actually "learned."

Summary and Conclusion

This chapter outlined the specific procedures involved in the experimental study, including a look at the research environment, the subjects, the facilitators, the

assumptions, and the controlled checks on possible threats to validity. In addition, the means for attacking the five investigative questions were outlined. The next chapter will present the results of this methodology, after the data and observational effects of the experiment have been analyzed sufficiently. This will hopefully open the doors to a sound resolution of the research question and an informed and factual perspective on the benefits and drawbacks of IRT.

IV. Results and Analysis

Background and Overview

The purpose of this chapter is to systematically analyze each of the investigative questions proposed in Chapter I. The results of the experiment will provide a quantifiable and objective viewpoint for comprehending IRT, as it pertains to those issues under study. In addition, personal testimonies from students, teachers, and administrators directly involved in the process will shed some light on the real-world ramifications of confidence-level use in the classroom. The goal is to obtain a deeper understanding of Information Referenced Testing and perhaps construct a prescriptive formula for a more efficient use of its principles in future educational and training scenarios.

Investigative Question #1

The first pertinent issue in the study focuses on the observable results of the experiment and an investigation into the related factors causing these known effects. The scores obtained from the control and experimental students, throughout the course of both exams, should help characterize the nature of Information Referenced Testing. In essence, it is important to know if IRT and traditional MC formats reveal a significant difference in scores. And, assuming that a contrast does exist, the contributing factors must be isolated and eventually identified.

For GR #1, the control and experimental groups were tested under traditional and IRT variables, respectively. For this first exam, a comparison between the sections showed that the control students appeared to score better on the MC portion. As stated previously, the test questions were identical for both groups; however, the experimental

subjects were provided confidence-level options and an appropriately-matched scoring algorithm. Therefore, the results (provided on the table below) should reflect no difference in instrumentation, aside from the desired construct under measure.

Table 12. Group Performance on Graded Review #1

Group	Mean Score	Standard Deviation	Population Size
Control	70.07%	12.84%	241
Experimental	66.09%	11.62%	235

Based only on the information presented above, it is not obvious that a significant difference exists between the two groups. Although the mean score was higher for the control students, the difference may have been a chance occurrence. A two-sample hypothesis testing procedure was proposed to discover if, indeed, there was a statistical relevance in the results. The results are given below. The p-value is the product of a two-tailed test, designed to see if an absolute difference in scores is apparent, for a reasonable alpha level.

Table 13. Group Performance Comparison for GR #1

Group Comparison	Mean Difference (Absolute)	P-Value (Two-Tail)	Significant Difference? ($\alpha = 0.05$)
Control vs. Experimental	3.98%	0.00039	Yes

The p-value for this analysis is extremely low, suggesting a significantly higher score for the control group. Also, the sample sizes are considerably large – adding to the legitimacy of these results. It can be safely assumed that the students exhibited a difference in scores for the first GR, although it is not clear if the IRT testing procedure acted alone in producing a relatively lower average performance.

The next step involves a comparative look at control and experimental performances on the second graded review. In this case, both sections were subjected to the same testing procedures – traditional items were used across the board. In order for the study to safely assume that the nature of the IRT items (by itself) caused a lowering of scores for the Management 210 students, GR #2 must result in equal performances for both groups. In other words, if it is found that the same disparity in scores exists on this "controlled" instrument, the results of GR #1 can likely be attributed to factors outside of the desired construct – IRT variables cannot be cited as the sole cause for the observed difference. The results for the second exam are presented in Table 14.

Table 14. Group Performance on Graded Review #2

Group	Mean Score	Standard Deviation	Population Size
Control	78.35%	11.40%	235
Experimental	78.41%	10.93%	218

It is evident that student scores were very similar, for both groups. The experimental group tested slightly higher, but a hypothesis test for significance, shown below, omits the possibility that these two samples showed any variance in performance.

Table 15. Group Performance Comparison for GR #2

Group Comparison	Mean Difference (Absolute)	P-Value (Two-Tail)	Significant Difference? ($\alpha = 0.05$)
Experimental vs. Control	0.06%	0.95439	No

It can therefore be assumed that the two groups were similarly dispersed into the control and experimental sections, with respect to MC testing abilities. The relative congruence of this aptitude is important, because it points back to the resulting contrast in IRT versus

traditional scores (for GR #1) as an important experimental discovery. Why would the experimental and control groups, which are otherwise equally adept at answering MC items, show such a startling disparity in performance – when IRT items were given to one group and not the other? Obviously, an observable difference in scores is evident, but the specific factors involved in this phenomenon are, as of yet, unknown.

Given that a difference does exist, the next phase of investigation must extend into an analysis of the specific factors involved in the performance results. The question can be simply stated: why did the experimental (IRT) group produce lower scores for the first GR? Are there any particular student characteristics that can be identified as the cause? In order to respond to this, an analysis of the samples' component parts is necessary.

The control and experimental groups were composed of students with varying backgrounds, to include some level of gender and ethnic diversity. Most of the cadets were juniors or seniors (Classes of 2004 and 2005, respectively) and ages were tightly grouped around the 19 to 21 year range. Departmental major, too, was of some significance in the study. Because Management 210 is a core course, required for graduation, both student groups encompassed every possible area of academic interest offered at the Air Force Academy. Four of the department's instructors, too, were shared by both the control and experimental groups. With all of this in mind, it was appropriate to take the control and experimental sections and break each one down into smaller samples – isolating those attributes mentioned above. Table 16 references the control group's performance on the MC portion of graded review #1, for each of the applicable categories of students (gender, ethnicity, departmental major – with a special look at

Management majors – age, class year, and common instructor). Standard deviation and population sizes are also given.

Table 16. GR #1 MC Attribute Performance for Control Group

Attribute	Mean	Standard Deviation	Population Size
Female	69.26%	13.36%	36
Male	70.21%	12.80%	203
Asian	72.00%	11.60%	15
Black	62.00%	8.92%	10
Caucasian	70.65%	12.81%	186
Hispanic	65.08%	13.32%	21
Engineering Major	71.30%	12.64%	72
Humanities Major	71.21%	12.91%	22
Social Science Major	68.91%	12.92%	107
Basic Science Major	70.51%	13.54%	33
Management Major	69.49%	11.83%	52
Age: 19	72.10%	12.20%	103
Age: 20	68.61%	13.19%	89
Age: 21	66.67%	12.07%	26
Class of 2004	76.56%	12.40%	31
Class of 2005	69.07%	12.64%	208
Instructor C	67.71%	12.80%	32
Instructor D	70.20%	12.94%	17
Instructor F	76.10%	11.28%	53
Instructor G	66.87%	13.99%	33

The purpose here is to get a better idea of how each attribute-group performed on the MC portion of the first graded review. By comparing these scores with the same divisions within the experimental group, it can be seen which of the student characteristics contributed toward the significant difference between the two test formats. To accomplish this, the experimental group was divided into exactly the same sample-

types and matched with mean performance, standard deviation, and population size. Table 17 summarizes this information:

Table 17. GR #1 MC Attribute Performance for Experimental Group

Attribute	Mean	Standard Deviation	Population Size
Female	67.11%	11.77%	44
Male	65.89%	11.63%	190
Asian	71.83%	9.34%	8
Black	61.92%	12.29%	16
Caucasian	66.37%	11.48%	195
Hispanic	62.07%	13.64%	9
Engineering Major	64.74%	11.10%	76
Humanities Major	69.14%	14.88%	27
Social Science Major	66.64%	11.09%	90
Basic Science Major	66.25%	11.31%	38
Management Major	63.53%	9.98%	44
Age: 19	66.82%	11.52%	111
Age: 20	66.27%	10.25%	88
Age: 21	63.31%	14.72%	26
Class of 2004	69.38%	11.72%	39
Class of 2005	65.57%	11.49%	194
Instructor C	70.42%	10.61%	33
Instructor D	60.87%	12.52%	39
Instructor F	71.78%	9.16%	15
Instructor G	57.38%	12.04%	13

As stated earlier, it was clearly established in the first investigative question that a significant difference in sample scores did exist, and it appears that the control group's performance was higher than that of the experimental group. It cannot be assumed, however, that IRT variables will cause a lowering of scores for every "type" of student. The purpose of this analysis is to find the difference in scores between control and experimental students, for all of the attributes listed above. For each category, an upper-

tailed hypothesis testing procedure was used to see if the control group had a significantly higher score. Depending on the particular population sizes for each of the groups in the study, large and small sample tests (with corresponding Z and t-distributions) were used to obtain a p-value.

Table 18. GR #1 MC Attribute Performance – Group Comparison

Attribute	Mean Difference (Control – Experimental)	P-Value (Upper Tail)	Significantly Higher? (a = 0.05)
Female	2.15%	0.22639	No
Male	4.32%	0.00023	Yes
Asian	0.17%	0.48507	No
Black	0.08%	0.49250	No
Caucasian	4.28%	0.00030	Yes
Hispanic	3.01%	0.29283	No
Engineering	6.56%	0.00041	Yes
Humanities	2.07%	0.30243	No
Social Science	2.27%	0.09225	No
Basic Science	4.26%	0.07944	No
Management	5.96%	0.00371	Yes
Age: 19	5.28%	0.00058	Yes
Age: 20	2.34%	0.09358	No
Age: 21	3.36%	0.18631	No
Class of 2004	7.18%	0.00824	Yes
Class of 2005	3.50%	0.00182	Yes
Instructor C	-2.71%	0.82144	No
Instructor D	9.33%	0.00906	Yes
Instructor F	4.32%	0.06907	No
Instructor G	9.49%	0.01517	Yes

According to the information displayed above, certain student attributes did exhibit a significant difference in performance (between IRT and traditional items), while others showed no considerable amount of change in scores. Those cadets encapsulated within the headings of "Male," "Caucasian," "Engineering Major," "Management

Major," "Age: 19," "Class of 2004," "Class of 2005," and students within the classes of "Instructor D" and "Instructor G" did score statistically lower with the IRT items, when compared with regular multiple-choice methods. All of the other tested attribute-groups did not show the same level of disparity in relative scores. In fact, members of Instructor C's sections actually performed *better* on the IRT examination.

What does this reveal about the nature of confidence-level variables? Certainly, within the constraints of this experiment, it can be shown that females, along with some ethnic groups, older students (ages 20 and 21), non-engineering and non-management majors, and certain instructor-based classrooms had relatively no problem with the IRT options. Scores were not diminished for these types of students, while the remaining groups constituted the bulk of disparity between the different MC formats.

The next stage of this concentrated look requires an analysis of different aptitudes (or academic strengths) within the groups. In the table below, students within the control group were segmented into various categories. First, all 241 cadets were ranked by cumulative college grade point average (GPA). The top 50 students, the bottom 50 students, and the 50 students located directly in the middle were then pulled out of the greater sample, and mean performances on the MC section of GR #1 were recorded, along with the standard deviation. This analysis was repeated for cumulative Military Point Average, ACA (a measure of each student's performance on the SAT and ACT), and essay performances on GR #1. The results may provide some insight into how various levels of overall student performance react to the IRT variables. Please reference Tables 19 and 20, below.

Table 19. GR #1 MC Aptitude Performance for Control Group

Aptitude	Ranking	Mean	Standard Deviation
GPA:	**Top 50**	78.67%	11.02%
	Middle 50	68.40%	12.76%
	Bottom 50	63.87%	11.28%
MPA:	**Top 50**	74.13%	14.55%
	Middle 50	69.47%	11.28%
	Bottom 50	69.33%	11.51%
ACA:	**Top 50**	75.33%	12.51%
	Middle 50	68.67%	11.99%
	Bottom 50	68.80%	12.38%
Essay:	**Top 50**	73.73%	12.59%
	Middle 50	65.60%	12.59%
	Bottom 50	68.40%	13.92%

Table 20. GR #1 MC Aptitude Performance for Experimental Group

Aptitude	Ranking	Mean	Standard Deviation
GPA:	**Top 50**	69.33%	11.93%
	Middle 50	64.88%	11.99%
	Bottom 50	61.13%	9.98%
MPA:	**Top 50**	66.97%	13.44%
	Middle 50	66.13%	10.00%
	Bottom 50	63.40%	9.48%
ACA:	**Top 50**	70.92%	10.97%
	Middle 50	67.00%	10.33%
	Bottom 50	64.52%	10.31%
Essay:	**Top 50**	68.03%	11.43%
	Middle 50	65.79%	12.70%
	Bottom 50	64.52%	9.25%

The goal here is to once again look at the difference between these selected sections and see if a notable distinction is evidenced. Control group scores, within these aptitude-groups were checked for areas of significantly higher performance in order to

explain the overall average discrepancy. The large sample test was used, because population sizes were all above 40, and the p-value was calculated. The results are given below:

Table 21. GR #1 MC Aptitude Performance – Group Comparison

Aptitude	Ranking	Mean Difference (Control – Experimental)	P-Value (Upper Tail)	Significantly Higher? (a = 0.05)
GPA:	**Top 50**	9.34%	0.00002	Yes
	Middle 50	3.52%	0.07751	No
	Bottom 50	2.74%	0.09922	No
MPA:	**Top 50**	7.16%	0.00529	Yes
	Middle 50	3.34%	0.05856	No
	Bottom 50	5.93%	0.00246	Yes
ACA:	**Top 50**	4.41%	0.03047	Yes
	Middle 50	1.67%	0.22783	No
	Bottom 50	4.28%	0.03019	Yes
Essay:	**Top 50**	5.70%	0.00889	Yes
	Middle 50	-0.19%	0.52989	No
	Bottom 50	3.88%	0.05029	No

The results suggest that the top 50 performers in each category scored significantly lower on the IRT examination, when compared with the control students. Also, the bottom 50 students, with respect to MPA and ACA also showed a contrast in performance. Middle performers in all areas were unaffected by the presence of confidence-levels in the testing procedure. It is difficult to assume if these finding are aligned well with those of the attribute-comparisons.

As a final check on these observations, it is necessary to compare performances on GR #2, whereby both groups tested under the same MC formats. Those student attributes and aptitudes that were proven to have a significant difference in scores

(between the two groups) for GR #1 should be analyzed in the same way for the second graded review. If GR #2 does not reveal a statistical change in performance within these groups, it would certainly help point to the IRT variables as the lone cause for the original performance discrepancy found on the first graded review.

Table # 22 isolates those nine attributes and six aptitudes shown to cause a significant difference in scores for GR #1. The same hypothesis testing procedures were conducted again – brought to bear on the MC results of GR #2, and a check for statistical variance was repeated as before.

Table 22. GR #2 MC Performance – Group Comparison

Attribute or Aptitude	Mean Difference (Control – Experimental)	P-Value (Upper Tail)	Significantly Higher? (a = 0.05)
Male	-0.61%	0.70678	No
Caucasian	-0.22%	0.57691	No
Engineering	-0.83%	0.68225	No
Management	-1.11%	0.69986	No
Age: 19	1.71%	0.12924	No
Class of 2004	3.93%	0.05130	No
Class of 2005	-0.69%	0.73171	No
Instructor D	-0.01%	0.50159	No
Instructor G	2.08%	0.27373	No
Top 50 GPA	-0.63%	0.62930	No
Top 50 MPA	1.40%	0.25175	No
Bottom 50 MPA	-1.72%	0.77518	No
Top 50 ACA	-2.23%	0.88179	No
Bottom 50 ACA	-2.39%	0.87097	No
Top 50 Essay	1.13%	0.30503	No

Interestingly, none of the selected parameters yielded a significantly higher MC score for the control group over the experimental group. In many cases, the experimental students actually outperformed their counterparts. This makes perfect sense, as the mean

score on GR #2 was higher for the experimental group (and control students were shown to have a lower average GPA – documented in Chapter III). Consequently, it can be logically stated that the higher control group performance on the first graded review (isolated into the above attribute and aptitude groups) was a direct result of the IRT influence. In other words, those "types" of students denoted in Table 22 exhibited lower MC scores because of the presence of confidence-level items. Conversely, no single student group performed significantly better with IRT, on average.

Investigative Question #2

Experimentally, the concept of IRT has introduced some notable effects that may have some bearing on future classroom instruction. Aside from the extra level of assessment that is hoped to result from this "information theoretic" approach, it appears that confidence-level test scores are influenced by certain student traits. The next step of the analysis will focus on the evaluation of IRT performance as a function of these traits. The purpose is to research how students with varying aptitudes and attributes are predicted to respond to these newly-designed variables.

Using the same student characteristics as those studied in the first investigative question, MC scores for both the control and experimental groups were matched with cadet statistics in a linear correlation model. Considering all four multiple-choice examinations (two graded reviews for each group), MC test results acted as the dependent variable and were measured against the specific characteristics of each individual – to include: instructor, section, essay score, GPA, MPA, ACA score, sex, race, departmental major (with a special look at Management majors), age, and class year. A model of best fit was constructed in all four scenarios, and the R-squared value (coefficient of

determination) was recorded – revealing the proportion of test score variation that can be explained by the simple linear regression model. The higher the R-squared value, the more approximate is the linear relationship between the dependent variable and its independent predictors. Table 23, below, diagrams the coefficient of determination for each graded MC section – GR #1 (experimental) reflects the examination with IRT items.

Table 23. Overall R-squared Values for Multiple Regression Models

Group	GR #1 (MC)	GR #2 (MC)
Experimental	0.295978	0.281158
Control	0.314153	0.290119

It seems that all four of the models were relatively similarly explained by the chosen variables. In other words, the student characteristics listed above and measured within the check for correlation seem to have at least a moderately equal effect on both traditional and IRT MC test scores. This may be misleading however, as some student traits that were strongly correlated for one model may have been latent within another. To combat this, a "step-wise" regression, whereby a statistical software package identifies the independent factors that can most aptly define the y-axis variation, was then used to pin-point those variables with the greatest contributions toward linearity. For the first graded review, the experimental and control group were subjected to this analysis.

Table 24. GR #1 – Strongest Estimators of MC Performance

Experimental Group	Control Group
Section	Section
GPA	GPA
Departmental Major	Race
Class	Management Major (Y/N)
	Age

It should be noted that some of the traits shown in Table 24 are common for both groups. "Section" and "GPA" were both highly explanatory of MC test performance (IRT and traditional), which should be expected. Student sections are defined by physical classrooms, and it can be safely implied that every class is confounded by innumerable factors that will either augment or handicap student learning. GPA, too, is an index of each cadet's relative success in college courses, and performance on any test should be (at least) partially aligned to the student's comprehensive grade point average. Therefore, the presence of these two factors in both models is encouraging. However, the differences between IRT and conventional exams may be defined by those characteristics that are uniquely manifested within the experimental group's model – shown here as "departmental major" and "class year." These two traits are not expressed within the control group's linear regression. Instead, "race," "age," and the "management major" consideration were believed to explain traditional MC test performance.

In order to fully understand the consequences of these models, it is necessary to look at the results of graded review #2, whereby the experimental and control groups were still composed of the same students; however, all of the subjects were given a uniform exam, with regular MC-items. The results of this "step-wise" regression should provide a baseline comparison as the control and experimental students were allowed to progress from GR #1 to GR #2. It is understood that time had elapsed between the administration of these exams, but the instruments were identical for the second graded review. It can then be assumed that changes occurring during the elapsed interval were standardized for both groups. In general, this should provide some additional insight into the characterization of the students in the study. The results are given below.

Table 25. GR #2 – Strongest Estimators of MC Performance

Experimental Group	Control Group
Section	Section
GPA	GPA
Class	Class
Gender	ACA
	Age
	Race
	Essay Performance

Section and GPA were again measured as strong predictors in these last two models. The experimental group showed that gender had an effect on performance, but departmental major was not seen on the second GR, following a strong presence on the first graded review. This proves that a student's particular major or general area of interest in school may dictate his or her performance on a confidence-level exam, while it seems to have no effect on traditional formats (it was not seen in any of the other three models as a significant correlative trait). The "class year" attribute was not shared by both groups on GR #1 – only the experimental (IRT) students saw it as an overwhelming presence in the model. It was shown to exist as a considerable factor on GR #2; it was seen in both of the groups. This shared trait may have been a factor of the testing instrument and can be equalized as a non-player. Therefore, the ability of "class year" to make its voice known in the experimental group only, for the first exam, is a matter of some legitimacy. The two student attributes of "departmental major" and "class year" were then isolated as two factors which may uniquely dictate performance on an IRT exam. To see the actual effects of these characteristics, mean scores for the IRT instrument were assessed – for all of the subsets within these two traits.

Table 26. Class Year MC Performance on IRT Examination

Class Performance	Mean Score	Standard Deviation	Population Size
Class of 2004 (Juniors)	69.38%	11.72%	39
Class of 2005 (Sophomores)	65.57%	11.49%	194

Table 27. Departmental Major MC Performance on IRT Examination

Departmental Major Performance	Mean Score	Standard Deviation	Population Size
Engineering	64.74%	11.10%	76
Humanities	69.14%	14.88%	27
Social Sciences	66.64%	11.09%	90
Basic Sciences	66.25%	11.31%	38

From the given data, it seems plausible that those members of the Class of 2004 were more likely to succeed on the IRT examination. Also, those cadets classified as "Humanities" majors seemed to outperform the other areas, and Engineering majors scored the worst of all. Hypothesis testing for all of these comparisons yielded the following information about the magnitude of observed differences on the IRT examination – see Table 28.

Table 28. Class Year and Departmental Major Comparisons on IRT Exam

Attribute Comparison	Mean Difference (Absolute)	P-Value (Two-Tail)	Significant Difference? ($\alpha = 0.05$)
Juniors vs. Sophomores	3.81%	0.06858	No
Engineering vs. Humanities	4.40%	0.16889	No
Engineering vs. Social Sciences	1.90%	0.27177	No
Engineering vs. Basic Sciences	1.51%	0.50121	No
Humanities vs. Social Sciences	2.50%	0.42455	No
Humanities vs. Basic Sciences	2.89%	0.39973	No
Social Sciences vs. Basic Sciences	0.39%	0.85847	No

While none of the comparisons appeared to show a significant contrast in scores for the IRT portion of GR #1, a look at the p-values might allow researchers to believe that some of the differences were more pronounced than others. For example, the shift between junior and sophomore scores may suggest that the more academically experienced cadets had an advantage on the IRT exam. Similarly, the distinction between "Humanities" and "Engineering" scores was somewhat considerable. This presents an interesting perspective on those students majoring in History, Philosophy, Fine Arts, English, or Foreign Area Studies as they were shown to test better than the cadets professing engineering abilities. Are students with supposed analytical strengths more likely to test poorly on an IRT examination, when compared with other students – especially those in Humanities fields? While these results seem enlightening, it is not known if they possess a universality that can be applied to all students in the educational and training world.

Investigative Question #3

As a means for evaluating the level of cadet and instructor acceptance of the IRT model, all of the professors involved in the Management 210 and Biology 331 experimental projects were interviewed. Additionally, the Botany students were surveyed on their interpretation of assessment-"fairness" that these items represented, as well as the understandability of the required procedures. The results of the survey are presented in Table 29. In summary, it appeared that student responses favored the IRT items, suggesting (as a whole) that the test items were perceived as fair estimators of knowledge. They also felt that the way in which the questions were constructed was not

overly complicated, when compared with the traditional (control) items. The percentages shown below represent the level of "agreeability" shown on the survey instrument, averaged for the control and experimental groups (on both graded reviews). Results higher than 50% show a positive response (test items were fair and easy to understand). These two constructs were measured using multiple survey-items, and those students with overwhelmingly contradictory responses were removed from the analysis.

Table 29. Biology 331 Survey Results

Surveyed Opinions	Control Group	Experimental Group
Test Items were Fair	54.58%	61.25%
Test Items were Easy to Understand	67.08%	71.67%

As for the Management 210 instructor interviews, opinions were mixed. Respondent A believed that the students disliked the confidence-level testing, but he "liked" it. "On a regular test, I have no idea if the student actually knows the material or made a good guess." Another professor (Respondent B) observed that the cadets were initially "thrilled to be given the opportunity for partial credit. That feeling evaporated very quickly when they discovered that they lost points for responses they really knew, but chose a somewhat safer, half-way response." Respondent B went on to suggest that the confidence-level approach not only encourages partial credit, but "partial effort," as well. "The students that do successfully hedge responses they are unclear of, do not learn the correct response. They only knew they were half right. The value of this for knowledge acquisition I feel is limited at best." Respondent C disagreed with the methodology of the experiment, feeling that "we really needed to run the complete test – only half of the students got to try the experimental version." Indeed, this would have

been ideal, but process issues precluded the extension of IRT implementation on the second graded review. The reasons for this are explained in the examination of process issues, in Investigative Question # 4.

The instructor involved with the Botany classroom study (Respondent D) also provided some valuable inputs. His feelings, overall, were that students exhibited some hesitance in using the confidence-level variables. "In my posttest discussion, students told me they avoided them because the point values assigned were below a simple guess." In other words, the cadets did not feel that venturing into the two-dimensional alternatives was as advantageous as randomly guessing – the level of credit was not worth it – "they wanted to take their chances rather than bite for the 'I don't know' option." However, this instructor had some success through his allowance of IRT-type responses on the fill-in-the-blank portion of the exams. "When I gave them a total recall, lab practical exam, one student asked if they could write 'I don't know' for partial credit, and I spontaneously said 'yes.' They used this more often… rather than guessing from a larger universe of possible responses."

Respondent D, himself, felt that the IRT model was a good "one-time deal to learn about the students and understand their thinking while taking the exams." He believes that such procedures, though, would have a more practical use in training environments, because one "could quantify the growth of confidence over time as training repetitions ensue." In short, he didn't see IRT as an important way to measure content mastery (summative evaluation) as much as a method for helping students to develop better study skills (formative evaluation). The erosion of error through confidence, over time, would be more beneficial for the learning process.

Investigative Question #4

Before the start of experimentation, it was wondered if significant process issues would plague the IRT procedures and possibly skew the results. This concern was actualized in the Management 210 portion of the study. The Botany sections, in contrast, were not overly burdened by the administration and scoring of the confidence-level variables, though this was probably attributable to the smaller sample sizes – less than 25 students took part in this supplementary experiment.

Despite the decreased scope of testing with the Biology students, Respondent D expressed some frustration with the "scanning" process of scoring the exams. Indeed, these items were assuredly more difficult to deal with, mainly due to the existence of multiple acceptable responses and non-integer point values. He insists that the scoring device must be appropriately programmed to handle these problems – continuing with these types of questions would not be possible without a way to ensure quick and accurate feedback through automation.

The Management classes experienced far more challenging obstacles in the experiment. The course director felt that the students, though initially welcoming the opportunity for partial credit, were disappointed by the lower scores. Instructors, too, were "skeptical" overall. In fact, after the completion of the first graded review, the use of IRT variables in the study was discontinued, due partly to lower scores and a lack of faculty belief in the system. However, a large part of the decision to abandon confidence-level testing altogether was based on process issues, most of which could be avoided in the future with increased communication and understanding.

The Management 210 Course Director summed-up this problem: "We (instructors) didn't fully understand the instructions on what we could or could not say to our students when administering the exam; therefore, students received different instructions based on who was administering the test." This may have created some level of confusion amongst the various sections. Also, "due to the size of the course, we used our computer help desk to barcode our answer sheets… but [their personnel] had no idea how to score this version." Apparently, it took several meetings and discussions to rectify the problem and produce accurate results – causing the Management Department staff to characterize the affair as a general "disaster." In order to ensure that the results of the experiment were salvaged, the researcher was forced to "hand-grade" each of the IRT and traditional tests on GR #1. As a method of standardization, this practice was extended into GR #2, resulting in the uniform scoring and recording of all (929) examinations. This was certainly a process issue and should not have to be repeated, by anyone, in any type of scenario.

When asked if he would be willing to work with IRT again, the Management 210 Course Director answered "yes," with some qualifications:

> First, the semester prior to administering IRT, the course director, researcher, and computer help-desk folks would have to develop a method to score the exams, and the method would have to be tested to my satisfaction before I would agree to this, again. Second, the course director and the researcher should coordinate during the semester prior, as well, so that the course director would fully understand all of the issues involved in administering the test. Also, the researcher should meet with the instructors who would be teaching the course to answer their questions about IRT and to advise them on how to properly facilitate the test. Third, the semester prior, we would need to determine the correct length of the test, to ensure that the students have enough time to complete the MC questions and the short answer questions, as well.

Some of his other concerns focused on the experience-level of the instructors handling the examination: "Possibly, for future experiments, researchers should use a course that has experienced instructors, instead of masters-level inexperienced instructors teaching undergraduates." In summary, he was initially excited to take part in the study, but was quickly let-down by the afore-mentioned problems. "I think this approach has merit, but these process and administration issues need to be addressed before I would participate, again."

The United States Air Force Academy's Director of Academic Assessment was also involved in the experiment. Her comments were in general agreement with the course director's, stating that the system is a "bit complicated… and no matter how much scientific evidence is presented, there's still quite a bit of resistance because of the traditional 4- or 5-option multiple choice test that everyone is used to taking and giving." If any type of merit can be attached to this testing format, it would still be a challenge to see it gain any wide-spread use. "Old habits die hard."

Investigative Question #5

Perhaps the most important issue to be resolved relates to the claim that IRT exams can more accurately reflect the percentage of information learned, as the name implies. If this can somehow be proven, it would be a monumental discovery in the world of educational and training assessment. It should be understood that no testing method can measure with 100% assurance the amount of comprehension within a particular student or classroom, but the question here addresses the comparative difference between IRT and traditional MC testing. It was necessary to use the Management 210 data in order to see if one method was more effective than the other.

One way of testing the legitimacy of IRT, as it was applied to this experiment, was to evaluate the GR #1 MC scores (with confidence level and traditional items) as a predictor for success in five areas: essay scores on GR #1, MC performance on GR #2, essay performance on GR #2, overall success on GR #2, and final exam scores for the entire course. To accomplish this, a linear regression model was set-up – using all of the individual scores for the control and experimental groups on the first graded review, appropriately matched to performances on the above-mentioned exam constructs. It was to be assumed that the "better" testing procedure would be more adept at predicting these five levels of Management 210 fluency. However, it should be noted that an experimental confound was manifested through this modeling technique. IRT test items were asked to predict future (traditional MC) test performance, resulting in the comparison between two radically different exam variables (shaded). Therefore, the given correlations with essay scores (not shaded) were stronger indicators of comparable predictive ability. Table 30 documents the coefficient of determination (level of explainable correlation) for all of these tests.

Table 30. IRT R-squared Values for Regression Models

Group (GR #1)	GR #1 Essay	GR #2 MC	GR #2 Essay	GR #2 Total	Final Exam
Experimental MC (IRT)	0.01819	0.05473	0.00225	0.02541	0.05456
Control MC	0.03184	0.08919	0.05359	0.12876	0.11986

Those columns shaded in gray represent predictions for regular multiple-choice tests. More appropriate comparisons can be made with the essay sections on GR #1 and GR #1. As can be seen, none of these predictive models were very strong, but the control

group seemed to act superior in every category. A more comprehensive investigation is, of course, necessary.

The central methodological format involved in answering this question was based on an item-by-item analysis. Data collected through experimentation revealed the control and experimental groups' performance on each question (1-15). It was assumed that the experiment was sound, because: the tests were exactly the same for both groups, the samples were comparatively equal (and large), and the examinations were given simultaneously. The only confounding variable existed in the fact that some different instructors taught the various sections involved in the study.

By surveying the instructors, with respect to the amount of coverage given to each of the tested objectives, the researcher was able to neutralize this contaminating influence and provide a means for resolving the given issue. Professors within the Management 210 course were asked to quantify the amount of time, effort, and instructional resources used to teach each of the objectives used in the 15 MC test items, based on a scale of one to ten. For the control and experimental groups, a weighted average was then applied to each question, reflecting the overall level of rigor applied by the department in "teaching" the given objective.

Assuming that the samples were indeed equal, surveying the instructors in this manner allowed for the investigation of the usefulness of IRT. A simple correlation was then performed – comparing the mean scores for each test question with the numerical representation (given by the instructors) of classroom attention. This was done separately for both groups. It can be assumed, on some level, that the more appropriate testing

method created a higher coefficient of determination, because performance should react directly with the level of teaching assistance given to the students.

This is of course based on a number of key assumptions. Can it be proven that classroom instruction is the only predictor of test performance? No, but the literature has pointed toward this variable as a main contributor, and to attest to the opposite would undermine the value of teachers in the classroom. Indeed, though, some of the students may have had prior knowledge of the material, studied the objectives on their own time, or simply exhibited more intelligence in the testing process. This was uncontrollable. But, it was evident in the analysis of Investigative Question #2 that each particular student section was extremely predictive of MC test performance; different teachers and teaching styles will perhaps affect their students more than anything. Therefore, the results of this correlation analysis were isolated for those "unique" and "common" instructors in the two groups, as well as for "all" of the professors represented in the course. For each of these three comparisons, it was assumed that the testing method with the stronger correlation exhibited less unexplained variation and perhaps more accurately reflected the percentage of information actually "learned" by the student.

Table 31. Item Analysis R-squared Values for Regression Models

Group (GR #1)	Unique Instructors	Common Instructors	All Instructors
Experimental MC (IRT)	0.17015	0.20801	0.23747
Control MC	0.36004	0.35292	0.48426

Again, it appeared that IRT came-up short, showing a lesser degree of correlation in every category. The most important comparison perhaps relies within the "common" instructors, because instrumentation would not be manifested in this check. The control

group exhibited an R-squared value of 35.3 percent, while the experimental (IRT) group showed 20.8 percent. This seemed to be a significant difference. The "unique" and "all" instructor categories appeared to mirror this observation, strengthening the overall conclusion: IRT was not proven to surpass classical multiple-choice examination as an information-referenced assessment tool.

Summary and Conclusion

The preceding chapter dealt with each of the investigative questions, in turn, relying on quantitative data extracted from the experiment and observations from those persons directly and objectively involved in the process. The results were allowed to speak for themselves. The final chapter will center on an increased level of interpretation and subjective analysis. The usefulness of IRT will be explored in other areas of academic and training assessment, and a final recommendation, on behalf of the research personnel, will be given to AETC and other interested parties.

V. Discussion

Background and Overview

The purpose of this final chapter is to summarize the effects of Information Referenced Testing (IRT) and attempt to answer the research question, as it applied to the cadets in the experiment. Recommendations concerning the use of IRT in future educational and training fields will also be presented as a means for better equipping its implementation in those selected areas. Finally, questions for additional research endeavors will be raised. The answers will help direct a full understanding of IRT and provide further information for Air Education and Training Command (AETC), as it considers the employment of confidence-level examination as a viable assessment tool.

Research Summary

The first part of the research question asks if IRT can be implemented in a practical learning environment. For the given experiment, it was evident that the administration of confidence-level formats is possible. However, some aspects were not successfully developed. Students and teachers within both departments (in general) could not fully accept the principles behind this assessment technique. IRT scores were significantly lower than those measured in the "control" sections, though this does not necessary indicate a negative effect. The allocation of points for the "partially sure" and "unsure" option variables did not always provide fair summative evaluation for the Management and Biology students involved in the study. Scoring procedures were also problematic. The Air Force Academy's automated system for grading MC exams was

programmed for traditional testing formats, and IRT was not accommodated by the existing infrastructure.

The second part of the research question dealt with the efficacy of IRT as an accurate instrument for assessment and performance feedback. IRT, as a summative assessment instrument, was handicapped by a lack of integration and communication. This resulted in poor faculty understanding and facilitation. Student comprehension of the IRT process was not fully acquired, yielding a somewhat disrupted "snapshot" of material mastery. With respect to the formative assessment strength of IRT, this experiment was inconclusive. Confidence-level results were not used in future phases of instruction to fill in the "holes" of student learning. Despite this, there were methodological designs used to indirectly gauge the degree of learned information reflected in both IRT and traditional tests. Traditional exams seemed to more accurately predict future performance, and IRT items were less reflective of the instructors' assessment of each objective's in-class coverage. In short, there was no evidence that Information Referenced Testing supported an enhanced measurement of information learned. This is definitely an area for future research.

It was apparent that some student attribute and aptitude groups were sensitive to the two-dimensional variables. Specifically, Engineering and Management majors, along with less-experienced students were labeled with "poor" IRT success. Also, high historical performers in other areas (GPA, ACA, MPA, and essay items) seemed to have more trouble with the confidence-levels on the exam. And, a number of instructor-based classrooms exhibited significantly lower scores, with IRT, than those given traditional testing formats. It can therefore be argued that IRT may have detrimental effects for

some "types" of students, while others are seemingly immune. Success may be dependent on the following factors: instructor's teaching style, student learning philosophy, experience level, and general academic success. While the first three are perhaps self-explanatory, it was interesting to see that high GPA, MPA, ACA, and essay performers were "hurt" by the IRT variables, when compared with the same aptitude-groups in the control section. This may be indicative of a problem within the conventional framework of education and testing. It would not be outrageous to suggest that some of the perceived "gifted" students are better test-takers, knowing how to manipulate traditional MC exams and receive inflated scores. These "test-savvy" cadets, when given a unique format, performed worse than the same "controlled" students because guesswork was masked and they were essentially challenged to think more about what was actually "known."

Recommendations

It is the opinion of this researcher that, despite some of the experimental results, IRT possesses fundamental strengths that will perhaps provide invaluable formative evaluation and constructive feedback in various Air Force training venues. Dr. Bruno's four constructs, which are founded on logical principles of assessment, are vastly superior to traditional right-wrong analyses. IRT outputs, if automated and categorically presented, can provide instantaneous classroom data, with respect to the type of information that is "fully" known, "partially" known, "unknown," and "misunderstood." The reliability of this two-dimensional construction could be profoundly useful as a means for pre-testing trainees and directing the course curriculum for a more focused attack on areas of weakness and confusion. Again, this seems most appropriate for

military training areas, where the difference between an "uninformed" and "misinformed" pupil could be disastrous.

Major recommendations for future testing of IRT should focus on the process issues that were experienced in this case study. All of the major players, including the students, should realize a complete understanding of the mechanics of IRT. They should be made fully aware of the general format, rules, and scoring procedures involved in the process.

The distribution of points for each of the confidence variables should also be considered. The researcher and teachers should agree on a formulated system that will motivate students to select the two-dimensional options, if they are unsure. This will more effectively take "guessing" out of the assessment picture and reward actual, full knowledge where it has been appropriately manifested. Areas of fractional confidence or no confidence can then be supplemented by re-education and additional assessment exercises. Complete confidence will eventually develop.

Finally, the manner in which these testing methods are used should be investigated and improved. Paper-and-pencil examination, using IRT, can be laboriously tedious and confusing for students and administrators. The use of computer modules and Internet platforms should be looked upon as a definite alternative. Clear examples, with practice questions, can be more clearly laid out in this environment. And, obviously, feedback will be quickly and succinctly provided, within the boundaries of the four desired constructs for assessment.

Questions for Future Investigation

How has IRT fared in other applications? Information Referenced Testing is currently in use at academic institutions and industry firms around the country. Dr. Michael Klymkowsky has initiated IRT as a pre-quiz instrument in his "Biofundamentals" learning laboratory at the University of Colorado, Boulder. Knowledge Factor has also set-up computer training modules, with IRT, for corporate training. An important part of any future experiment should focus on the experimental and practical success of these operations.

Should IRT be considered for use in other Air Force educational and training venues? AETC is composed of a diverse collection of academic institutions and technical training schools that fuel the mission-oriented needs of the United States Air Force. The Air Force Academy, along with Reserve Officer Training Corps programs and the Air Force Institute of Technology, provide the facilities and personnel for college and graduate-level degree acquisition. The bulk of Air Force assessment is accomplished in training courses, technical schools, and career-field proficiency programs in place at Air Force units, worldwide. All of these settings are dependent on MC exams. If, after further research, derivations of IRT are found useful, this exam format should be considered for careful implementation. Again, there is much more to be learned about this unique testing method, and scholarly research in the field should be exhausted before final acceptance and administration. It should perhaps be used guardedly in training scenarios and valued for its formative strengths and feedback report system for continuous learning.

How could future experiments be improved? Any additional experimental success is most definitely predicated on greater levels of understanding by the teachers and students involved. The full effects of IRT can only be realized by this level of comprehension, integrated facilitation, and comfortable acceptance of the confidence-level principles. Of course, as mentioned earlier, the use of computer programs to administer, grade, and report the tests would provide immeasurable ease and effectiveness to the entire process.

If implemented, what are the major considerations? Point-value assignment is the most important aspect of IRT – students must not be deceived by unfair appropriations of credit. The confidence levels are more reflective of "learned" information if the two-dimensional variables are worth more than the calculated "risk" of guessing. And, finally, teachers must understand the basic purpose of the testing variables and provide instruction for the students. This will ensure a psychological benefit, because the test-takers would not be inhibited by unnecessary stresses or ambiguity attached to the actual examination model.

Conclusion

This paper briefly summarized the history and current application of testing procedures, focusing on MC items and the major issues governing their use. Confidence-level exams, with a special look at Dr. Bruno's IRT method, were introduced to the reader, and an experimental method for testing IRT was presented. The results of the study were given, relating data to a specific research question and five appropriate investigative questions. An analysis and interpretation of the results followed, hoping to shed light on the assessment benefits of this examination procedure. Finally,

recommendations were made and questions were outlined for future experimenters to attack. AETC should continue to study the complexities of educational and training assessment and attempt to resolve the issues uncovered in this report (and others), in order to gain greater assurance that existing testing methods are accurate and indicative of student and trainee learning. Complacency in this area would perhaps adversely affect the fundamental execution of mission-essential operations.

Bibliography

Bruno, James E. and A. Dirkzwager. "Determining the Optimal Number of Alternatives to a Multiple-Choice Test Item: An Information Theoretic Perspective," *Educational and Psychological Measurement*, 55: 959-66 (December 1995).

Bruno, James E., Judith R. Holland, and Joseph W. Ward. "Enhancing Academic Support Services for Special Action Students: An Application of Information Referenced Testing," *Measurement and Evaluation in Counseling and Development*, Volume 21: 5-13 (April 1988).

Bruno, James E. "Assessing the Knowledge Base of Students: An Information Theoretic Approach to Testing," *Measurement and Evaluation in Counseling and Development*, Volume 3: 116-30 (October 1986).

-----. Professor, Graduate School of Education, University of California, Los Angeles. Personal Correspondence. October 2002.

-----. Professor, Graduate School of Education, University of California, Los Angeles. Personal Correspondence. 22 January 2003.

Campbell, Donald T. and Julian C. Stanley. Experimental and Quasi-Experimental Designs for Research. Boston: Houghton Mifflin Company, 1963.

Conderman, Greg. "Program Evaluation: Using Multiple Assessment Methods to Promote Authentic Student Learning and Curricular Change," *Teacher Education and Special Education*, Volume 24, No. 4: 391-94 (2001).

Cycyota, Cynthia. Management Professor, United States Air Force Academy, CO. Personal Correspondence. 9 December 2002.

Devore, Jay L. Probability and Statistics: For Engineering and the Sciences (5th Edition). Pacific Grove CA: Duxbury Thomson Learning, 2000.

Dirkzwager, A. "Testing with Personal Probabilities: 11-Year-Olds can Correctly Estimate Their Personal Probabilities," *Educational and Psychological Measurement*, 56: 957-71 (December 1996).

Giacomini, Andrew. Management Professor, United States Air Force Academy, CO. Personal Correspondence. 11 February 2003.

Goel, Pramod. Vice President of Business Development, Knowledge Factor, Inc., Lafayette CO. Personal Correspondence. January 2003.

Haladyna, Thomas M. Developing and Validating Multiple-Choice Test Items. Mahwah NJ: Lawrence Erlbaum Associates, Publishers, 1999.

Hansen, James D. "Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks," *Journal of Education for Business*, 73: 94-7 (November/December 1997).

Harris, Robert B. and William C. Kerby. "Statewide Performance Assessment as a Complement to Multiple-Choice Testing in High School Economics," *The Journal of Economic Education*, 28: 122-34 (Spring 1997).

Hassmen, Peter and Darwin P. Hunt. "Human Self-Assessment in Multiple-Choice Testing," *Journal of Educational Measurement*, Volume 31: Number 2: 149-60 (Summer 1994).

Klymkowsky, Michael. "The Evolution of Biology Teaching and the Web." Unpublished Report on MCDB1111/Biofundamentals Experiment. University of Colorado, Boulder, 2002.

-----. Biology Professor, University of Colorado, Boulder. Personal Correspondence. 21 January 2003.

Levy, David. Management Professor, United States Air Force Academy, CO. Personal Correspondence. 18 December 2002.

Madaus, George F. and Laura M. O'Dwyer. "A Short History of Performance Assessment: Lessons Learned," *Phi Delta Kappan*, Volume 80, No. 9: 688-95 (May 1999).

Malarkey, James. "Assessment Without Grades." Address to Southwestern Ohio Council for Higher Education. Edison Community College, Piqua OH. 06 November 2002.

Miller, Harry G., Reed G. Williams, and Thomas M. Haladyna. Beyond Facts: Objective Ways To Measure Thinking. Englewood Cliffs NJ: Educational Technology Publications, 1978.

Noyd, Robert. Biology Professor, United States Air Force Academy, CO. Personal Correspondence. January 2003.

Pomplun, Mark and M. D. Hafidz Omar. "Multiple-Mark Items: An Alternative Objective Item Format?" *Educational and Psychological Measurement*, 57: 949-62 (December 1997).

Powell, Janet L. "How Well do Tests Measure Real Reading?" *ERIC Digest*, ED306552: n. pag. (1989).

Revak, Marie. Director of Academic Assessment, United States Air Force Academy, CO. Personal Correspondence. 21 January 2003.

Rogers, W. Todd and Dwight Harley. "An Empirical Comparison of Three- and Four-Choice Items and Tests: Susceptibility to Testwiseness and Internal Consistency Reliability," *Educational and Psychological Measurement*, Volume 59, Issue 2: 234-47 (April 1999).

Rogers, W. Todd and Joyce Ndalichako. "Comparison of Finite State Score Theory, Classical Test Theory, and Item Response Theory in Scoring Multiple-Choice Items," *Educational And Psychological Measurement*, 57: 580-9 (August 1997).

-----. "Number-Right, Item-Response, and Finite-States Scoring: Robustness with Respect to Lack of Equally Classifiable Options and Item Option Independence," *Educational and Psychological Measurement*, Volume 60, Issue 1: 5-19 (February 2000).

Sacks, John. Management Professor, United States Air Force Academy, CO. Personal Correspondence. 9 December 2002.

Walsh, W. Bruce and Nancy E. Betz. Tests & Assessment. Englewood Cliffs NJ: Prentice-Hall, Inc., 1985.

Walstad, William B. and Denise Robson. "Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics," *The Journal of Economic Education*, 28: 155-71 (Spring 1997).

Webb, Robert R. Management Professor, United States Air Force Academy, CO. Personal Correspondence. 20 December 2002.

Weitzman, R. A. "The Rasch Model Plus Guessing," *Educational and Psychological Measurement*, 56: 779-90 (October 1996).

Wisner, Joel D. and Robert J. Wisner. "A Confidence-Building Multiple-Choice Testing Procedure," *Business Education Forum*, April 1997: 28-31.

Wood, William C. "Linked Multiple-Choice Questions: The Tradeoff Between Measurement Accuracy and Grading Time," *Journal of Education for Business*, Volume 74, No. 2: 83-6 (November/December 1998).

Vita

First Lieutenant Eric D. Larson graduated from Mountain Home High School in Mountain Home, Arkansas in 1995. He entered undergraduate studies at the United States Air Force Academy in Colorado Springs, Colorado where he graduated with a Bachelor of Science degree in English Literature in June 1999.

His first assignment was at Moody AFB, Georgia as a supply officer with the 347th Supply Squadron. In August 2001, he entered the Graduate School of Engineering and Management, Air Force Institute of Technology (AFIT). Upon graduation, he will remain at AFIT as an instructor in the Logistics School.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
25-03-2003	**Master's Thesis**	Jun 2002 - Mar 2003

4. TITLE AND SUBTITLE
AN ANALYSIS OF INFORMATION REFERENCED TESTING AS AN AIR FORCE ASSESSMENT TOOL

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Larson, Eric, D., First Lieutenant, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 P Street, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GLM/ENS/03-05

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The Department of Defense (DOD) has placed a great deal of importance on training and education, throughout all areas of infrastructure development and force implementation. A more knowledgeable operating unit, in any situation, is consistently the deciding factor for success. The United States Air Force, too, has emphasized this ideal and sought to employ those persons most qualified for the required task. Yet, problems within the classroom and various training venues are always present and should be continually marked for improvement. Existing assessment techniques should provide an accurate account of the quality of information learned by DOD personnel. This is undoubtedly crucial to war and peacetime functions. Therefore, testing as an assessment tool should be challenged, and new procedures – if deemed effective – should be recognized and introduced. This thesis looks at examination methods based on confidence-level items and two-dimensional feedback mechanisms. Information Referenced Testing (IRT) has been designed to more effectively measure and reflect the amount of knowledge attained by a student. The following research is an examination of IRT and its role in Air Education and Training Command. It will study two-dimensional items in multiple-choice examinations as a legitimate assessment tool for students, instructors, and administrators.

15. SUBJECT TERMS
Assessment, Evaluation, Testing, Training, Education, Knowledge, Information, Management, United States Air Force Academy

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Stephen M. Swartz, Lt. Col., USAF (ENS)
U	U	U	UU	110	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-6565, ext 4285; e-mail: Stephen.Swartz@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18