THE IMPACT OF STUDENT ABILITY AND METHOD FOR VARYING THE POSITION OF CORRECT ANSWERS IN CLASSROOM MULTIPLE-CHOICE TESTS

By

DANE CHRISTIAN JOSEPH

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Educational Leadership & Counseling Psychology

MAY 2010

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of DANE CHRISTIAN JOSEPH find it satisfactory and recommend that it be accepted.

_____
Michael S. Trevisan, Ph.D.

_____
Jennifer Beller, Ph.D.

_____
Brian F. French, Ph.D.

Acknowledgments

I wish to thank God for the blessings he has bestowed upon me, from everyday health and wellness to the opportunity to go through the dissertation process. I would also like to thank my advisor Dr. Michael Trevisan. His knowledge and experience to guide me through the graduate process has had an enormous impact on my development as a learner, researcher, and professional. Without his patience for my many ideas and sometimes lack of discipline to bring things to fruition, this project would not have been completed.

I also wish to thank my committee members Dr. Brian French and Dr. Jennifer Beller for their advice and support. They were part of a contingent of academic peers and mentors who deserve partial credit for who I am today. Dr. French was extremely instrumental in mentoring me through many of the data analysis issues, as well as helping me to understand what it all meant. He made learning fun and enjoyable and often times put me back into the seat after I fell out. Among others, special thanks to Ms. Lynn Buckley and Ms. Priscilla Rose, the late Dr. Len Foster, Andy Boyd, Elaine Wood, Nick Sewell, and Dr. Pam Bettis. Huge thanks to Dr. Andrew Storfer for being willing to input this scholarly exercise into his class on such short notice, as well as for his generosity to work with me from a distance.

To my wife Morgan, who stuck with me through all the adversity, listened to every idea, helped prepare every manuscript, and never let me quit when I wanted to, thank you so much. She is the brains and organization behind the madness. To the Schroeders – my adopted family, thank you. Without your generosity and support, I don't know where I would be today. To the outstanding people I have met throughout my life, thanks: Sir Ellis Clarke, Dr. Steve Shoals, Bob and Peggy Arfman, Yasushi Kimura, Paul Zimmerman, Dr. Joseph Campbell, and Jason Gurran.

Finally, one person deserves the most credit, and her name is Donna Joseph – my mother. I dedicate this dissertation to you.

THE IMPACT OF STUDENT ABILITY AND METHOD FOR VARYING THE POSITION OF CORRECT ANSWERS IN CLASSROOM MULTIPLE-CHOICE TESTS

Abstract

by Dane Christian Joseph, Ph.D.
Washington State University
May 2010

Chair: Michael S. Trevisan

Multiple-choice item-writing guideline research is in its infancy. Haladyna (2004) calls for a science of item-writing guideline research. The purpose of this study is to respond to such a call. The purpose of this study was to examine the impact of student ability and method for varying the location of correct answers in classroom multiple-choice tests. Educational testing literature supports the argument that randomizing the test-key is superior to other methods because it reduces the chances that test-takers can use guessing strategies successfully. However, the scant empirical literature on the impact of test-key formats has been restricted to large-scale educational tests. For this study, three test formats were developed for one test instrument based on different answer-placement strategies discussed in the educational measurement literature: a randomized, arbitrary, and balanced format. These test formats were randomly distributed to university students that participated in this study. Students were given an option to self-report which GPA range category they fell within. Based on these reports, students were placed into high, average, and low ability groups for analysis. Factorial ANOVA was conducted on the interaction and main effects of student ability and test format on test scores. Item analyses were

also examined through indices of item difficulty, discrimination, and test reliability. Results showed no interaction effects between student ability and test format. Test scores differed across student ability as expected. Test scores were not significantly different across test formats. The argument to randomize the test-key over other methods in classroom multiple-choice tests requires more empirical attention given the evidence presented in this study.

Table of Contents

List of Tables

Chapter One

Introduction

Background of the Study

Tests permeate all levels of education where students must demonstrate proficiency of knowledge, skill, or ability (KSA). Such proficiency may be a path to admissions or advancement. Test results are efficient tools for providing important information to test users about the achievement or ability level of test takers. Test results also provide relatively good feedback on the efficacy of instruction (Haladyna, 2004). Tests are not only important in education, but in many other fields as well. Haladyna defines a test as "a measuring device intended to describe numerically the degree or amount of learning under uniform, standardized conditions" (p.4).

The use of educational tests in the United States has increased at all levels of academia, from k-12 to graduate and professional schooling (Slavin, 2006). This practice affects the lives of many people, from teachers and superintendents to students and parents (Phelps, 1998). The interpretation of test scores may have consequences for these stakeholders. Drummond and Jones (2006) argue that students' and schools' standardized test performance may be proportional to the funding these schools receive. Students' test performance could also open pathways to attend higher learning institutions and gain academic scholarships to cover personal costs (Cohn, Cohn, Balch, & Bradley Jr., 2004).

Test performance is measured by test-takers' responses on test items. A test is made up of at least one test item or set of items (Haladyna, 2004). An appropriate test or item format will allow test-takers to tap into the appropriate subject-matter knowledge or cognitive processes. Accurate and precise responses are thus elicited from the test-taker. Choosing a test or item format is a significant undertaking for test users. The multiple-choice (MC) format is highly utilized in both high and low stakes testing situations. MC test items are relatively easy to construct and administer, while responses are objective to score and interpret (Downing, 2002b; Drummond & Jones, 2006; Haladyna, 2004; Stiggins, 2001). The following table lists some of the more common MC formats as well as their advantages and limitations.

Table 1

Advantages and Limitations of Some Multiple-Choice Formats

MC Format	Advantages	Limitations
Conventional	1. Usually a straightforward format with a stem and a corresponding number of options from which to choose	1. Some sub-formats such as the incomplete stem or the use of blanks in the stem can increase test anxiety
Matching	1. Items are easy to construct 2. Presentations of items is compact, allowing for more items on a single page 3. Suited for testing understanding of concepts, principles, and procedures	1. Tendency to write as many items as there are options, so that test takers match up items to options. This invites cueing 2. Tendency to have nonhomogenous options
Extended-Matching (EM)	1. Items seem less resilient to cueing 2. Items are more resilient to guessing	1. Similar disadvantages to the MC Matching format
Alternate-Choice (AC)	1. Item writer need only come up with one working distractor 2. More items can be assigned to a testing period than compared to Conventional MC format items	1. If there are an insufficient number of items in the test, guessing can be a factor
True-False (TF)	1. Items are easy to score 2. The judgment of a proposition	1. Items tend to promote the testing of recall

	as true or false is realistic 3. Items can measure different cognitive processes	2. Concerns have arisen over differences between true TF items and false TF items
Complex MC	1. Seems well suited for testing situations where there may be more than one right answer	1. Items tend to have lower discrimination, which in turn lowers test score reliability 2. Items require more reading time, thus reducing the number of items of this type on a test 3. Format is difficult to construct and edit
Multiple True-False (MTF)	1. Avoids many of the disadvantages of the Complex MC while being able to test for more than one right answer 2. Effective format for validity and reliability 3. Format is efficient in item development, examinee reading time, and the number of questions asked in a fixed time	1. Appears limited to testing the understanding of concepts 2. Item dependence may arise
Context-Dependent Item Sets (CDIS)	1. Very effective way to measure complex thinking 2. Three different types of stimuli used (problem, pictorial, interlinear) help make the item more interesting	1. Vignettes such as reading passages may pose a problem to test takers with reading problems

Haladyna (2004) lists some ways in which MC tests are used in both high and low stakes testing environments. These are for example, "placement, selection, awards, certification, licensure, course credit (proficiency), grades, diagnosis of what has and has not been learned, and even employment" (p. ix). High stakes testing programs are those such as the Educational Testing Service (ETS), the American College of Testing (ACT), and the Law School Admissions Council (LSAC). These programs consist of professionals devoted to creating and validating standardized multiple-choice test items for a variety of test formats. Such tests are then administered to students nationwide. Low stakes testing is generally linked to classroom

assessment (Haladyna, 2004). In university courses, MC tests are commonly used to assess large classes (Bridgeman & Lewis, 1994; Holtzman, 2008).

MC items are usually classified as either high or low-inference (Haladyna, 2004). High-inference items target students' abilities to understand and apply abstract concepts. Low-inference items target students' recall of facts and descriptive information. They also target most mental and physical skills that can be concretely observed. Many researchers agree that MC test items are capable of assessing both high and low-inference material in classroom multiple-choice assessments (Downing & Haladyna, 2006).

The ability to write quality MC items is highly desirable (Haladyna, 2004). A quality item discriminates well among levels of student ability. Quality items ensure that high-achieving students tend to choose right answers while low-achieving students tend to choose wrong answers (distractors) (Haladyna, 2002; 2004). Haladyna and Downing (1989a, 1989b) pioneered a major line of research on the methods and effects of writing quality MC test items. They sought to establish sound item-writing principles.

Haladyna and Downing (1989a, 1989b) began with a list of 43 item-writing guidelines. Author consensus existed for some guidelines but not for others (Haladyna & Downing, 1989a). In a follow-up study they investigated the validity of these guidelines using more than 90 research studies as references (Haladyna & Downing, 1989b). Few of the guidelines were found to receive extensive empirical study. Some guidelines were investigated and supported by logical argument. But almost half of the 43 item-writing guidelines received no empirical study at all (Haladyna, 2004). A decade after the appearance of the first two studies, Haladyna et al. (2002) reprised their study of the 43 item-writing guidelines. They examined 27 new textbooks and just

as many new empirical studies since the original studies. The result was a reduced list of 31 item-writing guidelines (see Table 2).

Table 2

General Item-Writing Guidelines

Content Guidelines

1. Every item should reflect specific content and a single specific cognitive process, as called for in the test specifications (table of specifications, two-way grid, and test blueprint).
2. Base each item on important content to learn; avoid trivial content.
3. Use novel material to measure understanding and the application of knowledge and skills.
4. Keep the content of an item independent from content of other items on the test.
5. Avoid overspecific or overgeneral content.
6. Avoid opinion-based items.
7. Avoid trick items.

Style and Format Concerns

8. Format items vertically instead of horizontally.
9. Edit items for clarity.
10. Edit items for correct grammar, punctuation, capitalization, and spelling.
11. Simplify vocabulary so that reading comprehension does not interfere with testing the content intended.
12. Minimize reading time. Avoid excessive verbiage.
13. Proofread each item.

Writing the Stem

14. Make directions as clear as possible.
15. Make the stem as brief as possible.
16. Place the main idea of the item in the stem, not in the choices.
17. Avoid irrelevant information (Window Dressing).
18. Avoid negative words in the stem.

Writing Options

19. Develop as many effective options as you can, but two or three may be sufficient.
20. Vary the location of the right answer according to the number of options. Assign the position of the right answer randomly.
21. Place options in logical or numerical order.
22. Keep options independent; choices should not be overlapping.
23. Keep the options homogenous in content and grammatical structure.
24. Keep the length of options about the same.
25. None of the above should be used sparingly.
26. Avoid using all of the above.
27. Avoid negative words such as not or except.
28. Avoid options that give clues to the right answer.
29. Make distractors plausible.
30. Use typical errors of students when you write distractors.

31. Use humor if it is compatible with the teacher; avoid humor in a high-stakes test.

Note. Adapted from "Developing and validating multiple-choice test items (3rd ed.)," by T.M. Haladyna, 2004, Mahwah, NJ: Lawrence Earlbaum Associates.

Users and researchers of MC tests desire capable MC items that can elicit precise and accurate results from test-takers. The former can then make good decisions when interpreting test scores of the latter. Haladyna (2004) has advocated the need to establish a "science of item-writing". This applied science rests on the ability to validate item-writing procedures that could be useful for both test-makers and test-takers. Nevertheless, Haladyna and other MC test experts like Downing (2002a) also believe that item-writing is still as much art as it is science. Experience in writing items can also benefit item-writers.

Regardless of experience, Haladyna (2004) has urged item-writers to pay heed: "when items are written without regard for item-writing guidelines that are featured (in Haladyna, 2004, p. 99), the consequences can be negative" (p. 97). Haladyna argues that test-makers' ability to write quality MC items increases their chances to create high-quality performing multiple-choice items with good discrimination. Item-writers thus have an incentive – improved item quality – to follow Haladyna et al's (2002) item-writing guidelines. This incentive arguably leads to improved test score reliability and validity of test score interpretations.

Valid score interpretation practices are a result of having valid test items as well as a corresponding scoring rule for those test items (Haladyna, 2002; 2004). The scoring rule decides how responses to individual test items will be coded and aggregated into a total test score (Bar-Hillel et al, 2005). Test users will want to know that an individual's total test score is reliable. Thus, valid interpretations can be made from these scores, all other things being equal and accounted for such as practice effects (Drummond & Jones, 2006).

Test specialists such as educational psychologists have been studying item responses for the past half-century or more. Two psychometric variables for evaluating responses are item difficulty and discrimination (Hambleton, Swaminathan, & Rogers, 1991). Test results could be misleading if test content is too hard or too easy for the test taker. Misleading test scores increase the chance that incorrect interpretations and judgments of test scores will follow. This affects the reliability of scores and the validity of interpretations made from scores. If test content is too easy, higher ability students may become bored with the test. This may hinder their concentration on future test items (Brown & Carroll, 1984). If test content is too hard, higher and lower ability students may have systematically lower total test scores. This may be due to the difficulty of items being unreasonable or unsuitable for the specific sample of test takers (Haladyna, 2004).

Values and social consequences also have a large role to play in test interpretation (Messick, 1989). Test development, administration, scoring, and interpretation of results depend on many factors. Such factors may affect the value framework from which both the test user and test taker operate. Haladyna (2004) elaborates on this by arguing that increased accountability from administrative sources causes higher pressures on classroom instructors. In response, teachers may try to use questionable tactics to artificially raise test scores and show educational improvement (Nolen, Haladyna, and Haas, 1992).

One tactic is to train and encourage students to use test-wise strategies when responding to items on a test. Such test-wise strategies become useful in situations where the test-taker must guess (Gibb, 1964; Roediger & Marsh, 2005; Supon, 2004). When guessing helps test-takers in answering an item, the difficulty of that item is lowered. Inaccurate results and interpretations of the test-takers' subject KSA thus follow (Haladyna & Downing, 2006). Regarding the importance of having well-performing items, test-makers will want their MC test items to

discriminate well (Haladyna, 2004). This is the case in both k-12 and higher-education classroom assessments where validating test items prior to test administration is likely unfeasible. If precise and accurate scores are desired, classroom instructors must be able to write quality MC test items for classroom assessments. The use of validated guidelines will therefore help test-makers to write higher quality MC test items.

Statement of the Problem

The multiple-choice item-writing guideline to vary the location of the correct answer according to the number of options seems logical. This logical structure is indicative of good face validity. Face validity is the degree to which a given measure appears to assess what it is supposed to assess (Slavin, 2007). Face validity can be established by providing a sound argument to support the underlying logic of a construct. The face validity of this guideline seems sound when considering test-taker responses.

Studies have shown that guessing test-takers may be apt to choose middle options in a systematically biased fashion. Test-makers also exhibit tendencies to systematically place the correct answer to an MC item in middle positions as well (Bar-Hillel & Attali, 2002; Bar-Hillel, Budescu, & Attali, 2005). Guesses are part of the construct-irrelevant-variance (CIV) of test scores (Downing, 2002a). CIV is "the degree to which test scores are affected by processes that are extraneous to its intended construct" (AERA, APA, & NCME, 1999, p.10). Any test score is some aggregate of correct (and incorrect) item scores. A subset of correct scores that are the result of correct guesses will reduce the reliability of the total test score.

Research on the impact of edge aversion and middle bias concludes that test-takers can take advantage of answer-keys (Attali & Bar-Hillel, 2003). These types of answer-keys are the

result of test-maker bias. Edge aversion is the tendency of respondents to avoid edges when a series of options are presented in a linear fashion (Rubinstein, Tversky, & Heller, 1996). Middle bias is simply the tendency to choose central options from linear formats without regard for avoiding the edges (Attali & Bar-Hillel, 2003). Attali & Bar-Hillel surmise that test-takers exhibit the same biases in guessing the location of answers as test-makers do in varying their location. Students can benefit, however slightly, from guessing middle options, especially when they know that at least one edge option is incorrect.

MC item-writers may therefore benefit from varying the location of correct answers according to the number of options. In other words, correct answers are assigned to both edge options and middle options. The benefit comes in the form of identifying test-takers who are prone to systematically choosing middle options under uncertainty. The resulting answer-key is thus more balanced.

A recent study by Bar-Hillel and Attali (2002) examined the use of a particular strategy known as the 'Underdog Strategy'. The underdog strategy can be used by test-takers to guess the location of an answer in a balanced key. Bar-Hillel and Attali give a description of how to use the underdog strategy in a paper-and-pencil test (p. 301):

a. Answer all the questions in a section you can
b. Count the frequency of each position among your answers
c. Select the position with the lowest frequency – the "underdog" position (in case of a tie, any one of them will do)
d. Give the underdog position as the answer to all as-yet-unanswered positions

A Monte Carlo simulation to compute the benefit of the underdog strategy with 10,000 SAT test-takers was used. Bar-Hillel and Attali (2002) demonstrated that higher ability students

can take advantage of the SAT using the underdog strategy to 'guess' the location of correct answers. Bar-Hillel and Attali thus advocated that test-takers randomize the position of correct answers in answer-keys. They showed that because randomization balances over time, a randomized key can very likely produce a balanced key. However, there is a unique advantage of a randomized process as opposed to a balanced one. Specifically, test-takers are unable to detect the pattern in randomized keys. They can thus do no better than randomize when guessing.

Test-takers would thus be better off focusing on answering each item as best they can. Furthermore, a randomized key has a (smaller) chance of turning out severely unbalanced. This procedure produces long-run sequences or may neglect options. Guessing test-takers will still be unable to take advantage using a particular strategy – other than randomizing. Every answer-key will more likely than not produce a different answer pattern over the long run (Bar-Hillel & Attali, 2002). In other words, the process of randomizing answer-keys would depend on chance. Thus, it could very well produce a severely unbalanced key one day but an overly balanced key on another.

Research on key balancing and randomizing seems circumstantially based. The recommendation to randomize answer-keys is built from an argument against the use of key balancing. But this argument was applied to MC tests in large-scale testing programs (Bar-Hillel & Attali, 2002; Bar-Hillel et al, 2005). Key balancing assigns the position of correct answers in a key. This results in an approximately (but not necessarily) equal frequency of options available in the test. Having equal representation among the choices is insufficient. Restrictions on long runs and cyclic patterns may lead to test-takers identifying patterns and thus increase CIV. Large testing programs employ arduous policies when balancing answer-keys to avoid runs and cyclic patterns (Bar-Hillel & Attali, 2002).

In classroom assessments however, it is unlikely that professors will go through the trouble of 'perfectly' balancing their answer keys. Bar-Hillel and Attali (2002, p. 299) presented a version of a large testing program's rules of thumb characterized in its balanced-key policy. These rules not only aim to balance the key but they also prevent unnecessary patterns from occurring. These unnecessary patterns could be detected by test-takers. Arguably, CIV increases if test-takers can guess correctly based on the patterns. The NITE's (Israel's National Institute for Testing and Evaluation) balancing key policy for their Psychometric Entrance Examination (PET) is as follows (Bar-Hillel & Attali, 2002, p.299):

1. No position should appear in the section key more often than nine times, or less often than four
2. Correct answers should never be placed over three times in a row in the same position
3. A sequence of about half-the-length of the section should not lack one of the four (option) positions

   Additional rules of thumb characterized were:

4. Do not exclude runs altogether (e.g., have some short ones, at least one run of three and two runs of two)
5. Avoid overly patterned sequences, such as obvious symmetries or repeated cycles

Given time and logistic constraints, a more likely possibility is that course instructors may use visual inspection to arbitrarily balance the key. This arbitrary approach can still be unbalanced (middle-biased). Such answer-keys can be taken advantage of by some students. Employing rigid key balancing procedures is also unlikely if teachers lack the necessary testing software. Large-scale testing programs can use software capable of inputting the balancing policy in the answer-keys to make the job efficient.

Nevertheless, the item-writing guideline to vary the location according to the number of options is good advice for instructors. Varying the location through certain methods can help prohibit edge aversion or middle bias from test-takers. Varying the location arbitrarily could be attempted by visual inspection in an attempt to "roughly" balance the key. It also helps to avoid unnecessary sequences such as runs and cycles that may result in randomized (chance) processes. But the arbitrary approach is still not foolproof. Other patterns may result that can be detected and taken advantage of by test-takers. Randomizing the key seems to be an easier process than balancing policies such as NITE's, and just as simple to do as with an arbitrary approach. To randomize, one just needs a pure randomizing device such as a deck of cards, coins, or computer program like EXCEL. Which approach is conclusively better remains to be determined empirically.

Nonetheless, the process of varying the location of the correct answer is good testing practice. It is part of the process of ensuring that precise and accurate test scores are derived from test-takers. These processes in turn help reduce CIV by eliminating test-takers' ability to detect or use key patterns to make correct guesses. Trained test-wise students will be more apt to find key patterns if they know what to expect (Attali & Bar-Hillel, 2003). In (more perfectly) balanced keys, higher ability students will be able to use the underdog strategy to greater effect. Theoretically, they will benefit from obtaining more correct answers before even using the strategy to guess the remaining ones. This decreases the number of items that require guesses. But it increases the probability of success that the underdog strategy will work when they guess. A study is thus warranted to detect whether any differences exist for randomizing, arbitrarily balancing, and perfectly balancing MC test answer-keys on test scores.

Purpose of the Study

The primary purpose of this study is to determine the impact of student ability and method for varying the location of correct answers in university classroom multiple-choice tests. MC tests are commonly used in large university undergraduate courses (Haladyna, 2004). Thus, such a course will be the setting for this study. Item analysis through difficulty and discrimination will also be considered as it may play a significant role in the ability of students to answer or guess correctly.

There will be three MC test versions in this study. Each version will represent a particular type of multiple-choice test answer-key variation method. One version will be randomized, another arbitrarily-balanced (visual inspection only) and the third perfectly balanced. Note that these three versions are just some of the different ways that a key can be varied. The efficacy of any guessing strategy might result from student knowledge of test-wise strategies (Roediger & Marsh, 2005). However, no instruction on test-taking strategies will be given. This is because students in the randomized group will be at a disadvantage to students in the other groups. Underdog might work for those in the latter but no strategy besides a randomized one will work for the former group. Since students cannot internally randomize (Bar-Hillel & Wagenaar, 1991), they will not be able to guess using any particular feasible strategy. This disadvantage will not be permitted in this study.

Research Questions

This study investigates the impact of student ability and method for varying the location of correct answers in classroom multiple-choice test answer-keys. The one manipulated condition investigated in this study will be the method of assigning answer positions among item

options. The three methods of assignment will be randomization, arbitrary-balancing, and perfect-balancing. Student ability will also be used as an independent factor in the analysis of results. Student ability to understand subject-matter is thought to have a significant effect on both the number of correct responses and guesses made (Bar-Hillel & Attali, 2002). The student ability factor cannot be manipulated. Item analysis will also be examined in the analysis of results. This provides a better understanding of students' ability to answer items correctly by method of assignment given varying levels of item difficulty and discrimination. To understand these issues, the following research hypotheses will be investigated:

1. To what extent does the method of answer-key assignment impact students' total test scores in a general biology course?
2. To what extent does the interaction of method of answer-key assignment and student ability impact students' test scores in a general biology course?

Significance of the Study

MC item-writing guidelines require further empirical support (Haladyna et al., 2002). Haladyna et al. call for increased attention to validating these guidelines. Downing (2002a) states that guideline validation should not be limited to one subject matter. This study responds to this call for validity research on item-writing guidelines. The results will contribute to developing a "science of item-writing". It will provide empirical evidence for the impact of employing a particular guideline in a MC test. Empirically generated evidence contributes to establishing sound conclusions of the impact between theory and practice (Mohr, 1995; Shadish et al, 2002). Although no specific number of studies can be assumed to indicate conclusive evidence for or against a particular guideline's validity, there is a significant lack of significant research on some

item-writing guidelines. The guideline to vary the position of correct answers in MC tests by a particular method – especially at the classroom level – has scant and inconclusive empirical support in the educational measurement literature.

Haladyna (2004) and Haladyna et al (2002) showed that some item-writing guidelines have received much more attention than others. Others are in need of serious attention immediately. The recommendation to vary the location of the correct answer, as well as methods to do so, seems to have good face validity. But research results providing empirical data for support are lacking. This research study attempts to provide such support.

Impact studies are common in the social sciences, including educational research (Shadish et al, 2002). The increased focus on our nation's educational standards has led to a substantial increase and focus on accountability (Slavin, 2006). Experts in the field of educational and psychological measurement have been studying multiple-choice item-writing for decades now (Haladyna & Downing, 2004). Item-writing guidelines are thought to improve test scores by reducing item-writing flaws. Such flaws may systematically affect item difficulty and discrimination, among other psychometric properties (Haladyna, 2004).

Guidelines have been validated through research in numerous academic disciplines. One example centers on the National Board of Medical Examiners (NBME). These medical examiners have been studying the effects of various MC item-writing guidelines as well as MC formats (Downing, 2002a, 2002b). Empirical evidence then contributes to making the validity argument for or against the use of specific MC item-writing guidelines and formats in medical examinations.

Haladyna (2004) and Haladyna and Downing (2006) advocate the "judicious" use of MC item-writing guidelines. Since the validity of some guidelines is still in question, item-writers

must be cautious when applying these rules. The art of item-writing is thus still important. But a science of item-writing is needed now more than ever. Accountability in education has grown significantly (Slavin, 2007). Test scores at the classroom level are just as much important in their own way as test scores from large-scale examinations (Haladyna, 2004). The item-writing guideline to randomize answer-keys has been shown to have a (slightly) statistically significant effect in a large-scale test (Attali & Bar-Hillel, 2003). This study will seek to confirm whether the same is true in classroom assessments. Course instructors will therefore benefit by having validity evidence that supports or refutes the need to randomize answer-keys. This is opposed to the approach to balance keys, either arbitrarily or according to some perfect rule.

Assumptions

The course professor responsible for creating the test items will have the knowledge and ability to create items reflective of the test content learned up to the point of testing. The arbitrary-balancing process will occur via visual inspection only. The perfect-balancing process will occur via a rule schema. This schema will be explained in Chapter 3's 'Instrumentation' section of the methodology. Students will be expected to give an honest effort during the testing session. An honest effort will produce more representative scores of the student's true ability to understand the test subject-matter.

Chapter Two

Review of Literature

Introduction

This dissertation investigated the impact of student ability and methods to vary the location of the correct answer in classroom multiple-choice tests. This chapter provides the conceptual foundation and empirical support that warrants the undertaking of such a study. The chapter is divided into three sections. In the first section, an overview of general item-writing guidelines is presented from the work of Haladyna (2004) and other test specialists. This section specifically focuses on the validity of some of these guidelines, as discussed in Haladyna et al (2002). The second section reviews literature on the guideline to vary answer location in multiple-choice tests according to the number of options. A logical argument for the various methods to employ the guideline is provided. Limitations for this argument are discussed. In the final section, I review some literature on the role of student ability in item-writing guideline validity studies.

The Validity of Multiple-Choice Item-Writing Guidelines

Developing multiple-choice tests and items.

The development of multiple-choice (MC) tests and items is a significant undertaking requiring item-writers to have a proficient level of subject-matter knowledge (Haladyna, 2004). Test-makers must decide whether an MC test format will suit their needs for assessing students over other formats (such as the constructed-response format). The decision is based on the

intended target of measurement – knowledge, skill, or ability defined by the level of learning being measured (abstract or concrete). Other considerations are the costs and benefits of competitive formats, and the consequences of using a particular format (Haladyna, 2004).

If a decision is made to use an MC format, item-writers must develop a stem, one single correct choice, and one or more distractors as part of an MC item. A stem may be complete or incomplete and consists of a stimulus for response. A correct choice is the one and only right answer that can be given as a response to the stem in question. Distractors are incorrect options that are plausible to some test-takers. Such test-takers do not possess adequate levels of subject-matter KSA to answer the stem (Haladyna, 2004). Good items are the result of test-makers' understanding of test (subject-matter) content and the type of mental or behavioral processes to be examined. Test-makers must therefore know how to choose an item format and how to write test items (Haladyna, 2004).

Haladyna and Downing (1993) argue that the development of plausible distractors is the most difficult part of item-writing. They found most items to have had only one or two working distractors. They concluded that three options (a single correct answer and two working distractors) were natural. Besides the true-false (TF), multiple-true-false (MTF), and alternative-choice (AC) formats, most other formats consist of three to five options (Haladyna, 2004). Test-makers must then make a decision as to where to assign the correct answer for an item. For example, the correct answer for a four-choice conventional MC format can go in any one of four positions. These positions are typically represented as option-A, option-B, option-C, and option-D.

Varying the location of answers reduces test-takers' ability to identify patterns in the answer-key and benefit from guesses (Haladyna, 2004). The method to vary the location might depend on the preference of the item-writer. The purpose of assessment is to gain an accurate and precise description of test-takers' KSAs. Assuming that item-writers use assessments for this purpose, they may prefer to use validated methods when varying answers to test keys.

Item-writing guidelines for multiple-choice tests.

Research on item-writing guidelines has been ongoing for the past few decades (Haladyna, 2004). The number of items that survive after all item-writing activities, checks, and reviews may be only around 50% (for testing programs) (Haladyna, 2004; Holtzman et al, 2002). Haladyna et al (2002) concluded that the lack of empirical studies on item-writing has led to the belief that some guidelines are already established due to their prominent occurrence and support in academic texts and scholarly opinion.

Haladyna and Downing (1989a, 1989b) examined 43 prominently occurring item-writing guidelines in the scholarly literature. They reviewed 96 theoretical and empirical studies and formulated a guideline taxonomy based on their data. These studies were pulled from the available educational testing and measurement literature. Approximately half of these guidelines had no empirical support. They were widely regarded by scholars as feasible and useable. Many scholars believed that these guidelines are valued but were not well assessed empirically. A revised taxonomy of guidelines resulted in a minimized version (Table 2). Repeated guidelines or those capable of being merged with others were ameliorated to facilitate conciseness in the taxonomy.

The lack of empirical results to support the validity of various guidelines stresses the need for continued item-writing guideline research (Haladyna, 2004). Some guidelines are admittedly well-researched or supported. These guidelines include: the number of options for an item (#19), the placement of the central idea in the stem (#16), the avoidance of clues (#28), and the need to have plausible distractors (#29). One of the few guidelines to have received scant empirical attention is Guideline #20 (Table 2). This guideline reads: "Vary the location of the correct answer according to the number of options. Assign the position of the right answer randomly".

In discussing the answer-varying guideline, Haladyna (2004) argues that response set – the tendency to mark in the same response category – causes biased keys to arise. But arbitrary-key balancing – or approximately equal distribution of correct answers across all options – creates a slight bias due to edge aversion (Attali & Bar-Hillel, 2003). Because edge aversion affects item difficulty and discrimination, test-makers should randomize option positions to avoid such negative psychometric consequences. Perfect-balancing is also problematic for psychometric reasons.

Validity research on item-writing guidelines.

A major research study in item-writing validity was accomplished through Haladyna and Downing's (1989b) study. This follow-up to their (1989a) study sought to validate their taxonomy of item-writing guidelines. Academic resources were drawn from and yielded 96 studies. Haladyna et al's (2002) revised taxonomy included an additional 27 educational test and measurement textbooks as well as 27 empirical studies. Haladyna et al reported that studies were collected from: conference proceedings, review of electronic databases of educational and

psychological articles, and references provided by each article. Studies included tests given to: medical, dental and nursing students; selection-test for entry-level police officers; undergraduate psychology and communications courses; the ACT; a biology test; and a science test for middle-school students.

The majority setting for these studies occurred in the classroom. This was also the setting for the measurement texts featured in the study. The validity evidence for their study was primarily "intended for teachers and others who write test items to measure learning" (Haladyna et al., 2002). Nevertheless, implications exist for "large-scale assessment programs involved with promotion, graduation, certification, licensure, training, or program evaluation" (p.311).

The distinction between classroom assessment and large-scale assessment item-writing rules is noteworthy. It is conceivable that both forms of assessment share common ground for item-writing rules. However, some rules may not apply or may be inappropriate in one but not the other. For example, the final item-writing rule concerning humor will only be practically conceivable when addressing classroom teachers. When students take large-scale tests, they respond to a selective set of items developed on a nationally standardized scale. Items are developed by large-scale assessment program item writers. The chances of the item-writer knowing the test-taker in this case are very little. Using humor is thus nullified when it is based on anonymity of the teacher's personality and classroom environment.

The difference between the level of assessment and its relation to item-writing rules also impacts this study. Haladyna's (2004) suggestion to randomize answer-keys is in response to Attali and Bar-Hillel's (2003) study that looks at two large-scale educational tests. Haladyna et al's (2002) study makes no such suggestion to randomize the key. This may be due to a lack of

validity evidence of the method to do so. Since their sole focus was on classroom assessments, they may have believed that such a suggestion was not feasible at the time. Or, they may have overlooked the psychometric effects of methods to vary the location because they were unaware of the consequences. Whichever way, Bar-Hillel and Attali's (2002) study calls into question the psychometric effects on manipulating answer-keys. The concern relates to both classroom and large-scale assessments. The lack of specificity for a method, call for more validity evidence and a "science of item-writing" are thus psychometric concerns. These concerns warrant an investigation of the impact of answer-placement strategy in classroom assessments.

In the validity procedure, Haladyna et al (2002) evaluated and classified authors' treatments of each guideline as cited, supported, or not cited. The least cited guideline (#5 – avoid overspecific and overgeneral content) received about 15% for. The most cited guideline (#15 – place the central idea in the stem) received 100% support for its use. The guideline to vary the location of the correct answer received 52% votes for, 48% votes uncited, and 0% of votes against its use.

Research on Varying the Location of the Correct Answer

The validity of any guideline may change depending on a new, compelling, logical argument and the collective old and new evidence bearing on this argument (Haladyna et al., 2002, p. 313). After the publication of Haladyna et al's revised taxonomy, two studies came out bearing results with the potential to significantly affect the validity of guideline #20. Research identified significant limitations in the methods used for varying the location of correct answers (Attali & Bar-Hillel, 2003; Bar-Hillel & Attali, 2002). The argument was to keep the recommendation to vary the location of correct answers, but to change the established means of

doing so. The recommended change was to transition from a more balanced process to a randomized one.

The thesis underlying the argument for randomization centered on elements of test-takers' cognitive ability to identify systematic answer-key patterns. Identification could result naturally or through test-wise training of test-taking strategies. Guesses based on particular test-wise strategies allow students to take advantage of these patterns. An element of guessing exists with the use of MC test items (Haladyna, 2004). Test-takers may guess when they have partial knowledge by eliminating implausible distractors in the absence of any or full knowledge (Haladyna, 2004).

For example, a student may be unable to choose the right answer from among four possible options if she lacks the necessary KSAs to identify the right answer. This insufficiency might prompt her to eliminate known distractors. The field of possible choices is thus narrowed down to more likely ones. She might then make a guess between the remaining options. If she has no knowledge whatsoever of the distractors, she might nevertheless choose to guess, especially if omitted responses are penalized.

The type of strategy used when partial knowledge of distractors is present is called an elimination strategy (Attali & Bar-Hillel, 2003). This guessing strategy is part of the broader framework of test-wise strategies that exist. Test-wise strategies are passed on to test-takers from test administrators or instructors during test preparation training. They may also be habitual elements of test-takers' response-set in the absence of knowledge needed to make a choice (Gibb, 1964; Roediger & Marsh, 2005; Supon, 2004).

Attali and Bar-Hillel (2003) investigated systematic tendencies exhibited by test-takers and test-makers in MC items. Both groups were found to prefer middle options to edge positions

up to 3 or 4 to 1 in isolated questions. 55% of option choices were favored for middle options. To carry out their study, Attali & Bar-Hillel investigated within-item answer position from five perspectives. The first two sections could be examined together. The first focuses on how test-makers might go about positioning the answer to an item. The second provides empirical evidence of this facet. The third section deals with test-takers' tendencies to favor middle options. The final section is an argument that this middle bias may really be edge aversion. From this, they examine the effects of edge aversion from a psychometric standpoint.

Several hypotheses are put forth to explain how test-makers might go about positioning answers in a key. Attali and Bar-Hillel (2003) examined several prior studies on answer-placing strategies. In all of these studies, subjects were asked to position an answer in one of four choices. In a four-choice MC test, only 50% of options would be expected by chance ($p<.0001$) to occupy middle positions. Instead, all of the studies reported at least a 70% occurrence of middle-positioning. In two studies, 80% of the answers were found in central options.

To investigate where people seek the position of correct answers, Attali and Bar-Hillel (2003) simulated a guessing scenario that elicited responses with blank options. The questions in order were:

What is the capital of Norway?

A B C D

What is the capital of The Netherlands?

A B C D

Respondents were either asked to respond to both questions or respond to the second question if the first was already pre-answered. The percentage of instances each position was chosen over the 196 total choices made in both questions were: A-15%, B-38%, C-39%, and D-

8%. Again, an almost identical finding to the test-maker answer-positioning studies was found with approximately 80%middle choices selected for the two questions ($p<.0001$).

Given these results, Attali and Bar-Hillel (2003) revisit an older notion of edge aversion with experiments in psychology and consumerism (Ayton & Falk, 1995; Christenfeld, 1995; Falk, 1975; Rubinstein et al, 1996). They investigated people's placement and selection choices in various situations. These studies adamantly supported the notion that subjects tend to avoid the edges rather than choose the middle. This notion of edge aversion was supported in a study of SAT five-choice items (Claman, 1997). Options A and E were the least popular, but option C – the exact middle option – was not as popular as B and D. In this case, extreme middle bias was not confirmed, but edge aversion was.

Attali and Bar-Hillel (2003) summarize this occurrence in several other answer keys. The results are shown in Table 3.

Table 3

Percentage of Correct Answers by Positions in Various Answer Keys

Number of Choices	Test	Number of questions	A	B	C	D	E	% in middle
4	PET pilot (1997-1998)	8905	25	26	25	24		51*
	10 operational PET tests	1640	25	24	23	27		48*
	Yoel (1999)	2312	24	28	27	21		55*
	Offir & Dinari (1998)	256	20	27	29	24		56*
	Kiddum (1995)	1091	24	26	26	24		52
	Open University (1998)	258	27	27	25	21		52
	Gibb (1964)	70	24	34	21	20		55*
	Trivia (1999)	150	23	27	27	23		53
	SAT (Claman, 1997)	150	29	23	23	25		47
5	SAT (Claman, 1997)	1130	19	20	22	21	19	63*
	MPT (1988-1999)	1440	18	22	21	21	18	64*
	INEPE (1998)	432	18	25	21	19	18	64*
	GMAT (GMAC, 1992)	402	17	19	23	22	19	64

Note. Significantly different than expected (50% in 4-choice tests; 60% in 5-choice tests, $p<.05$). Adapted from "Guess where: The position of correct answers in multiple-choice test items as a psychometric variable," by Y. Attali and M. Bar-Hillel, 2003, *Journal of Educational Measurement, 40 (2)*, 109-128.

These results clearly support that central options are occupied more often than not with the correct answers. Attali and Bar-Hillel (2003) then examined some psychometric consequences of edge aversion. In switching middle answers to edge positions, the effect of a position on percentage correct responses was larger than effects on incorrect responses (3.3%). But the magnitude of a position effect would depend on the difficulty of an item. This in turn affected items' discrimination.

Bar-Hillel and Attali (2002) advocated the use of randomized keys to minimize the effects of test-takers' use of strategies to guess the position of correct answers. Randomizing the key means that on average a balanced key is expected to turn out. The main difference between the two key versions is that randomizing devices can produce any one element of a possible set of elements. That is, in four-option items, each option has an equal chance a priori of occurring. Some forms of balancing can help to prohibit the final key from having certain sequences such as cycles, palindromes, and runs of certain lengths. Randomizing the key does not do this. In theory, a randomized key could eventually hold a severely skewed distribution of options – such as 60% As, 14% Bs, 20% Cs, and 6% Ds. As the randomized procedure continues in this instance, we would expect fewer As to occur after more trials. More Bs, Cs, and Ds will then occur until the frequencies average out at approximately 25% for each option.

A follow-up study confirmed the argument that randomized keys are superior and indeed, more rational to use than their balanced counterparts (Bar-Hillel Budescu, & Attali, 2005). Yet for all their sophisticated arguments, the validity of this guideline recommendation is lacking sufficient empirical support. Attali and Bar-Hillel (2003) run Monte Carlo simulations involving a large-stakes test – the SAT verbal section. This neglects the application of their specific methodology to the guideline in low-stakes environments such as classroom assessments. More

validity research is needed in this latter environment. The empirical evidence will support or refute the arguments put forth by Attali and Bar-Hillel, as well as Bar-Hillel and Attali (2002). Studies of randomizing items (Marks & Cronje, 2008) have supported the argument against the use of these procedures. Students are negatively affected when more difficult items get relocated to the front of the test. It is time for the educational measurement community to study the effects of the randomized answer-key argument put forth by Attali and Bar-Hillel.

The Place of Student Ability in Item-Writing Guideline Validity Studies

Logically speaking, higher ability students will on average get more correct scores on item responses. As student ability increases, we would also expect higher ability students to guess more items correctly. Ability and successful-guessing are therefore theoretically proportional. This relationship increases if the answer-key is developed in a certain way. For example, in balanced keys, we would expect higher ability students to answer more correct items as usual. Using the underdog strategy suggested by Bar-Hillel and Attali (2002), higher ability students will be better able than lower ability students to identify less frequent choices. This is because they are able to count the frequency of options that they have already selected. Assuming their answer selections that are not guesses will more likely result in correct responses, they can benefit from the strategy to guess correctly.

Studies in psychology and decision sciences have found that people are quite adept at locating patterns (Attali & Bar-Hillel, 2003; Haladyna, 2004; Rubinstein et al, 1996). For these other reasons, Attali and Bar-Hillel (2003) again recommend use of randomized answer-keys. A randomized key theoretically limits subjects' ability to guess the correct answer using any strategy other than a randomized one. But randomization processes will balance outcomes over

long runs (Bar-Hillel & Attali, 2002). If this happens to be the case for a particular test-key, then higher ability students might still be able to identify the key's pattern after getting more correct answers. The same is the case if a randomized key were to yield a greater number of correct answers in more central options. Ability should thus be included in this dissertation study.

Student ability has been assessed in item-writing guideline validation studies (Green, Sax, & Michael, 1982; Trevisan, Sax, & Michael, 1991). Ability has been assessed through the use of proxy variables such as Grade Point Averages (GPAs). GPAs allow for the comparison of students at various achievement levels and correlate highly with intelligence test scores and academic ability (Gardner, 1986; Sax, 1989; Slavin, 2006; Trevisan et al, 1991). Haladyna (2004) devotes an entire chapter section on the role of student ability in assessing item responses from MC tests. With so much focus on the role of student ability in analyzing test scores, and given the link between student ability and successful guessing, this study employs the use of GPAs as a proxy to student ability to evaluate the impact of different methods for varying the location of correct answers.

Cronbach and Snow (1977) cite student ability grouping was being used in education since the early 1920s. On the basis of ability, various educational interventions could be assigned to schools or classrooms in need. Cronbach and Snow also discuss the interaction of aptitude and treatment. A common problem in educational research is to investigate interactions of individual differences among learners going through various instructional treatments. "Aptitude is defined as any characteristics of a person that forecast his probability of success under a given treatment" (Cronbach & Snow, 1977, p.2). Hence, the aptitude x treatment interaction is a natural one to investigate. The present study investigates student GPAs as the main aptitude variable. Cronbach and Snow discuss how a treatment may be any type of manipulated variable. Since the

manipulated variable in this study is the answer-key version, this study investigates the interaction of GPA with answer-key version.

To test the significance of interaction, an aptitude measure is investigated through random assignment to treatments from which outcome measures will be collected (Cronbach & Snow, 1977). These outcomes can be test scores, for example. ANOVA can then be used to test for the significance of the interaction by blocking students on aptitude. They can then be grouped according to various levels of ability. Trevisan et al (1991) used one such particular methodology by blocking treatment and ability level. This variability then reduced the error term and made the ANOVA procedure more powerful.

A large sample size is required to utilize such a methodology (Cronbach & Snow, 1977). Cronbach and Snow's recommendation is to have 100 students assigned at random to each treatment group. This reduces the chances of type-II errors since they become highly probable under sample sizes of 40. In other words, no interaction effect could be accepted when indeed one was the case (Cronbach & Snow, 1977, p.46). Neither Attali and Bar-Hillel (2003) nor Bar-Hillel and Attali (2002) empirically investigate the interaction impact of student ability with different answer-key versions.

The form of student ability investigated will vary according to the study. Gibb (1964) examined students' test-wiseness. This test-wiseness was categorized as a complex higher-order construct. The construct of test-wiseness represents a student's ability to use test-wise strategies when taking tests. Multiple-choice test formats are included in this boundary. Haladyna (2004) discusses the role of clues that test-takers gather from MC tests. These clues can prompt the test-taker to use test-wise strategies to increase the likelihood of successful guesses. For example, the assumption of local item independency in a MC test section attempts to keep the answers to

one item independent from the answers to other items in that section. When this assumption is broken, test-takers can guess the location of the correct answer. Such guesses are contrary to their knowledge of the subject-matter itself.

In this case the student is no longer exhibiting his or her ability or mastery of the subject-matter. Rather, the student is exhibiting his or her ability to identify the clues in the test – perhaps subconsciously identifying local dependency. The correct item responses as a result of this ability are more representative of the students' test-wiseness and less so of their subject-matter KSAs. Note that this is a matter of degree since it is still possible that a correct guess could be nothing more than pure luck, as a result of a blind choice.

Chapter Three

Methodology

Introduction

The call for more validity support for multiple-choice test item-writing guidelines has been made (Haladyna et al, 2002). Good item-writing practices reduce flaws in items and minimize test-taker confusion. The ability of students to focus on answering the item is increased. When students can give their undivided attention, their test scores will be a more accurate and precise representation of their true ability. This dissertation provides empirical support for the multiple-choice item-writing guideline to vary the position of correct answers according to the number of options by randomizing the answer-key. The guideline concerned recommends that specific methods be carried out to perform this action. These methods inhibit test-takers from taking advantage of the answer-key when guessing. Student ability also has a role to play in the extent to which students can take advantage of the key. The limited research conducted on this guideline has been mostly examined in large-scale educational testing programs. Studies of the guideline's impact in classroom assessments are needed. Thus, the impact of student ability and method for varying the location of correct answers in classroom multiple-choice tests was examined.

Research Design

A true experiment was conducted with a between-subjects factorial design employed. Univariate ANOVA was chosen because it has the ability to evaluate the significance of mean differences of a dependent variable (DV) between two or more groups or levels of the

independent variable or factor (Mertler & Vannatta, 2005). The manipulated factor in this study was the method for varying the location of correct answers. Students were assigned to conditions and their ability taken into account in the analysis. Student ability was divided into three levels: high ability, average ability, and low ability. The research questions provided in chapter 1 that were examined are again as follows:

1. To what extent does method of answer-key assignment impact students' total test scores in a general biology course?
2. To what extent does the interaction of method of answer-key assignment and student ability impact students' test scores in a general biology course?

Study Participants

Participants came from a large land-grant and research based university located in the Pacific Northwest region of the United States. This institution had an approximate annual enrollment of 20,000 undergraduate and graduate students. Because it was a 100-level course, the majority of participants were freshman and sophomore class students. There was some junior and senior class students enrolled. The course was a general biology course. This course regularly employs MC testing as an assessment tool since it is an efficient means of testing large groups of students (Haladyna, 2004).

The course professor had previously taught all sections of this course. Sample size was a total of 540 (15 sections of 36) students. 369 students agreed to have their test scores analyzed in this study. This sample size was necessary to have adequate power for the analysis of differences between the groups.

Instrumentation

The test instrument was a multiple-choice general biology exam. This test instrument was designed to assess student knowledge from course lectures and required course readings. In the cognitive tradition of assessing KSAs at an undergraduate introductory level, students were required to demonstrate basic knowledge, including facts, principles, and relationships in general biology. The course instructor developed the items based on his experience in writing and using classroom multiple-choice tests, as well as his expertise in the subject matter.

In creating the test, the course instructor was asked to vary the location of the correct answer arbitrarily. A conventional MC format was used since it is the most widely used format in MC testing. Incoming students to this research institution were required to take entrance exams containing conventional MC items (for example, SAT, GRE, etc.). This significantly increased the likelihood that the study participants had previous encounters with MC item formats.

Empirical research has been conducted on the number of MC items required in a testing period. The rule of thumb is to use as many items as there are minutes for the testing period – i.e., one minute per item (Burton, 2006). For a 50-minute lecture period, a 50-item test instrument was developed. One extra-point item was administered in this exam. This item was not included in the analysis of results, however. Each item was worth one-point each for a total possible test score of 50 points. Each test-item had five options, A-E.

After the researcher collected the developed instrument, he then informally proofread and screened the items. This double-check helped ensure that generally recommended item-writing content guidelines were followed, including style and format concerns surrounding item clarity, grammar, punctuation, and spelling. Following this, each of the three test versions was created.

Each version had the same exact items. The versions represented three different answer-varying strategies for positioning the correct answer in the key.

Test format A – randomized.

Test version 1, the randomized key, was developed in the following manner. In an EXCEL spreadsheet, a list of all possible multiple-choice options was placed into column A. Because this study took into account 5-choice MC items, column A's first five rows were occupied by the letters A, B, C, D, and E respectively. Because there will be fifty items, the next 50 cells below were occupied by the formula below. Note that the formula was only typed once, and then copied and pasted into the subsequent 49 rows. Each of the fifty cells was in a 1:1 correspondence with the order of MC items as they appeared on the test.

=INDEX(A1:A5,RANDBETWEEN(1,COUNTA(A1:A5)),1)

This formula automatically and randomly selected one of the five options (A, B, C, D or E) to be placed into the cell. These options represented the locations of the correct answer for each of the fifty MC test items.

Test format B – arbitrary.

Test version 2, the arbitrary-balanced key, was developed by the course instructor by choosing an option position for each item's answer location. An assumption was that the instructor would give his best to balance the key in mind. No retrospection or changes were made to his selection. In other words, the instructor wrote one question and then chose the location of the correct answer while writing out the possible options. This was the final location of the answer for the arbitrary format.

Test format C – balanced.

The arbitrary method was opposed to perfectly balancing the key. To accomplish this for test version 3, the researcher assigned the correct answer ten times to each possible option. None of the NITE's balancing rules mentioned in chapter 1 (Bar-Hillel & Attali, 2002, p.299) were heeded. Instead, there were two columns in an open EXCEL spreadsheet. In the left column, there were numbered items 1-50. In the right column, there was a place to put one of five options, A-E. The researcher then had an accomplice assist in placing the location of the correct answer by saying "Start" and "Stop". When the accomplice said "Start", the researcher began moving the mouse in the second column down through each cell. On the accomplice's "Stop" command, the researcher then placed an "A" in that cell (corresponding to the item number). This was repeated for 10-As, then 10-Bs, etc. If a cell that the accomplice had stopped on was already occupied by an option, then the next available option space further down the column was chosen.

Data Collection Procedures

The test was administered on a day and time at which the course usually met. Each student randomly received one of the three test versions – an arbitrarily-balanced answer-key, perfectly-balanced answer-key, or a randomized answer-key. Randomizing the test by student rather than by section was intended to reduce systematic error and increase the power of the design to detect treatment effects (Trevisan & Sax, 1990, p. 9). Following Haladyna's (2004) discussion that guessing is inversely proportional to total test score, subjects were also informed that unnecessary guessing would tend to lower their total test score. Haladyna examined the

probability of obtaining correct scores in a ten item test when examinees guessed for 0-10 of the items. The more guesses examinees made, the less likely they were to get items correct.

Descriptive data on subject demographics was collected to examine whether the subject sample in this study were representative of the general university population. These statistics were gender, ethnicity, and age. Collecting demographics of this type is commonplace in educational measurement studies (Slavin, 2006). Multiple-choice item-writing guideline studies have also analyzed the distribution of such demographic data (Downing & Haladyna, 2006).

Data Analysis Procedures

A between-between analysis of variance design was used. Ability level was discerned by requesting that students respond to a GPA-estimate item located at the end of the test versions. Student ability levels were analyzed according to the following GPA cut-score categories on a 4.0 scale: High ability (3.7-4.0); Average ability (3.0-3.4); Low ability (2.0-2.7). Students' predicted and self-reported GPAs were collected and used as a proxy measure in the analysis (Trevisan, Sax, & Michael, 1991). The noncontiguous design excluded students with GPAs between these ability cutoffs in the analysis. No verification of the students' GPA estimates was done.

Using these noncontiguous ability groupings increased the power of the design by narrowing within-group variability (Trevisan, Sax, & Michael, 1991). A noncontiguous group is established by discrete cut-off points at either end of the group range. Noncontiguous designs were recommended by Cronbach and Snow (1977). The attempt was to increase the power in their designs when using continuous variables such as intelligence (Trevisan, 1990). Trevisan explains that since the distance between the ability groups on the range of GPA scores are spread

out, the standard deviation within the groups is decreased. The reduction of within-group variability increases the power of the overall design. Trevisan states that although GPAs are a measure of achievement rather than intelligence, achievement measures such as GPA correlate highly with intelligence test scores (p.38).

Item difficulty and discrimination was assessed. Item statistics such as difficulty and discrimination are used to assess the nature of items and the quality of item responses (Hambleton, Swaminathan, & Rogers, 1991). In this study, item difficulty was assessed through the use of p-values. As opposed to more advanced item response theory (IRT) techniques, p-values are a classical test theory (CTT) approach to evaluating difficulty. P-values are relatively easy, reliable, and feasible to use (Haladyna, 2004).

Item discrimination in the form of point biserial correlation coefficients was computed. Point biserials provide a correlation between a dichotomous variable and a continuous one. In this study's case, items were scored dichotomously as 0 or 1 – 0 being incorrect and 1 being correct – whereas the test score was aggregated on a continuous level. The point biserial of an item thus compared the number of students getting an item right/wrong with the total test score when taking that item into account. Cronbach's alpha was used to estimate internal consistency reliability.

Because ANOVA was used, statistics computed included F ratios, degrees of freedom for the particular factor and error, levels of significance, effect sizes, and observed power (Mertler & Vannatta, 2005). The F ratio is split to report between and within-variance. Between-group variance specifies variance not explained by main effects, or in other words treatment effects plus error variability. Within-group variance specifies error variability.

Null hypotheses of the study were:

Null Hypothesis 1: Test scores will not differ among students' scores from different test formats for combined ability levels.

Null Hypothesis 2: The interaction between student ability and test format will not have a significant effect on test scores.

These null hypotheses reflected the nature of the research questions and the logical argument underlying the undertaking of this dissertation. Specifically, students' total test scores were predicted to be better in the two non-randomized test-key formats. To recap, the rationale is based on the theory that students can more accurately guess the location of correct answers when the answer-key is perfectly-balanced (using the Underdog Strategy, for example), and also when the answer-key is arbitrarily-balanced (since both most correct options and most guesses are edge averse). No such successful strategy can be utilized for a randomized answer-key.

Note on Student Ability

Students were asked to estimate their current Grade Point Average (GPA) so as to compare them on different ability levels. This permitted an analysis of the ability-treatment interaction. Spreading out the distance between ability groups on the range of GPA scores decreased the standard deviation within groups and therefore increased the power of the test statistic (Trevisan & Sax, 1990). GPA was used as a proxy measure for ability. This feature of the design has been well-established in previous research literature. Measures of achievement such as GPA correlate highly with intelligence test scores and academic performances (Downing & Haladyna, 2006; Trevisan & Sax, 1990; Trevisan et al, 1991).

The GPA category was representative of a student's cognitive ability to understand the subject material from the general biology course lectures and readings. This was in contrast to

GPA being representative of a student's test-wiseness (refer to the 'Student Ability' section in Chapter 2). Although further research could examine the role of student ability to use test-wise strategies, this study did not make that connection.

Power Analysis

Experimental evidence from item-writing validation studies has supported the use of alpha levels set between 0.01 and 0.05 (Haladyna, 2004). This is also consistent with social science research methodology (Lipsey, 1990). Power analysis was conducted using an alpha level set at 0.05. This power analysis determined the adequate sample size required for each cell of the ANOVA tables (Cohen, 1988; Lipsey, 1990). Given that this is a social science study, power was recommended at 0.80 and alpha set at 0.05 (Cohen, 1988; Slavin, 2006). Effect size for this study was examined at 0.50 – a medium effect size – and 0.80 – a large effect size – as recommended in the measurement literature (Cohen, 1988; Lipsey, 1990). Shadish, Cook, and Campbell (2002) discussed the need to report effects sizes in research studies because of the difference between practical and statistical significance. In other words, effect size determines the strength or size of effects or differences regardless of significance (Mertler & Vannatta, 2005). A power analysis was undertaken to ascertain the required sample size needed to achieve power at 0.80 when alpha was set at 0.05 at various effect sizes. The results of this analysis are found in Table 4.

Table 4

Power Analysis of 3x3 Interaction and Main Effects with Required Sample Sizes for Various Effect Sizes

Type of Effect	Effect Size	Total Sample Size	Cell Sample Size
Interaction	0.50	196	22
	0.80	133	15
Main	0.50	158	17
	0.80	107	12

Note. These table figures are for a desired power of 0.80 when alpha is set at 0.05.

These required sample sizes from the power analysis were in accord with the use of the biology course in this study, which had a total sample size of 369 students.

Methodological Limitations of the Study

Two major design limitations were noted:

1. *Self-reported GPAs*. Self-reports have been found to be artificially over-inflated in many studies, especially those concerning student academic ability (Dobbins et al, 1993). The results of this study may have been over-inflated as a result of collecting self-reported GPAs. Students may have been placed incorrectly into one ability group based on their self-report when in fact they did not belong to that particular ability group.
2. *Use of noncontiguous ability groupings*. This procedure effectively eliminated some students from analysis whose reported GPAs fell outside the ranges used to mark the ability groups (see section on Data Analysis above for actual group cut-off points).

Chapter Four

Findings

Means, standard deviations, sample sizes for each form of the test and each ability group, as well as mean p-values are presented in Table 5. Table 6 presents reliability results of test scores from the different test formats using Cronbach's alpha. Note that validity evidence is discussed in Chapter 5. Tables 9 and 10 present the results of the univariate ANOVA test. Raw scores are located in Appendix D. Item statistics, including item p-values and variances, as well as point biserial correlation coefficients are located in Appendices E-G. A breakdown of item-option frequencies is found in Appendix H as part of the distractor analysis.

Table 5

Means, Standard Deviations, Sample Sizes, and Mean P-Values for Each Test Form and Ability Group

Test form	P	Items	M	S	Sample Size
Randomized					
L	0.58	50	29.28	5.58	32
A	0.60	50	30.26	5.73	57
H	0.75	50	37.74	3.12	19
C	0.62	50	31.29	6.08	108
Arbitrary					
L	0.56	50	28.38	5.16	26
A	0.61	50	30.52	4.57	58
H	0.71	50	35.83	3.68	18
C	0.61	50	30.91	5.17	102
Balanced					
L	0.57	50	28.57	5.65	28
A	0.60	50	30.42	4.28	48
H	0.73	50	36.87	4.21	15
C	0.61	50	30.91	5.44	91

Note. L = Low Ability, A = Average Ability, H = High Ability, C = Combined Ability Groups

Table 6

Cronbach alphas for each test format for N of 50 items

Test Form	Cronbach alpha
Randomized	.80
Arbitrary	.69
Balanced	.72

Data Screening

Descriptive data on students' ethnicity, age, and gender was examined to ensure that this study's sample of test takers was an accurate description of the student population at the institution. Data were screened for missing values and to ensure that the assumptions of factorial ANOVA would be fulfilled. The cases were transformed to fit a normal distribution with a low score of 23 and a high score of 42. Data were also examined for outliers through examination of box plots and histograms. Two outlier cases were deleted from the analysis because they did not fit the distribution of scores. Both outlier cases revealed very low raw scores (9 and 14) for two students who self-reported high ability levels in the GPA category. Investigation of histograms and tests for skewness and kurtosis revealed a normal distribution. To test for the assumption of homogeneity of variance, Levene's test was utilized when conducting the ANOVA.

Because of a design limitation to request students' GPA ranges instead of actual GPA scores, a large average ability group was obtained. This group was more than twice the sample size of the high and low ability groups. Frequency data from table 5 confirms this observation when comparing the average ability group to the high and low ability groups in each test format. Since the use of factorial ANOVA assumes equal sample sizes, random selection of cases was employed in order to produce a more equal cell sample size distribution. The procedure followed

was as follows. First, a random variable on an interval between 0 and 1 was assigned to each case. So for example, case 1 had a corresponding random variable equal to 0.61 whereas case 76 in the data set had a corresponding random variable equal to 0.36. The GPA range was then recoded into a different variable. This recoding was based on establishing a cutoff point for each case in the average and low ability groups that corresponded to their assigned random variable. The resulting cell and group sample sizes can be found in Table 7. These cell sample sizes satisfied the condition of power discussed in Chapter 3. The sample sizes in these categories were then used in the Two-way ANOVA interaction and main effect analyses. The means and standard deviations of the cases used in the ANOVA analyses are located in Table 8.

Table 7

Cross-tabulation of Cell and Group Sample Sizes for Student Ability x Test Format

		Test Format			
		Randomized	Arbitrary	Balanced	Total
Student Ability	High	19	18	15	52
	Average	18	20	19	57
	Low	19	17	18	54
Total		56	55	52	163

Table 8

Means, Standard Deviations, and Sample Sizes for Each Test Format and Ability Group in the ANOVA Analyses

Test Form	M	S	Sample Size
Randomized			
L	29.11	6.09	19
A	31.44	4.23	18
H	37.74	3.12	19
C	32.79	5.88	56
Arbitrary			
L	29.41	4.98	17
A	31.45	4.50	20
H	35.83	3.68	18
C	32.25	5.08	55
Balanced			
L	28.61	5.77	18
A	30.79	4.21	19
H	36.87	4.20	15
C	31.79	5.80	52

Note. L = Low Ability; A = Average Ability; H = High Ability; C = Combined Ability

Null Hypothesis Findings

Null Hypothesis 2: Test scores will not differ among students' scores across different test formats for combined ability groups.

A univariate ANOVA was conducted and a summary of results presented in Table 9. Levene's test of equality of variances found no significant difference, indicating homogeneity of variances within groups, $F(8, 154) = 1.42$, $p = .19$. Main effect results from Table 9 revealed no significant differences were found among student scores from different test formats at $p > .05$, $F(2, 154) = .32$, $p = .72$, partial $\eta^2 = .00$. Estimates of effect size revealed no strength in associations.

Null Hypothesis 1: The interaction between student ability and test format will not have a significant effect on test scores.

A univariate ANOVA was conducted and a summary of results presented in Table 9. Levene's test of equality of variances found no significant difference, indicating homogeneity of variances within groups, $F(8, 154) = 1.42$, $p = .19$. No interactions between student ability and test format were found to be statistically significant at $p > .05$, $F(4, 154) = .35$, $p = .84$, partial $\eta^2 = .00$. A line plot was also conducted and revealed no interaction. Estimates of effect size revealed no strength in association.

.

Table 9

Two-way ANOVA Summary Table of Interaction and Main Effects

Source	Df	Sum of Squares	Mean Squares	F ratio	F probability	Effect Size
Between						
treatments	8	1750.79	218.84			
Ability Level	2	1685.02	842.51	39.38	<.01	.33
Test Format	2	13.71	6.85	.32	.72	.00
Ability Level x						
Test Format	4	30.22	7.55	.35	.84	.00
Within treatments	154	3294.65	21.39			
Total	163	174979.00				

Table 10

Bonferroni Pairwise Multiple Comparisons of Significant GPA Factor Levels

					95% Confidence Interval	
GPA Range 5 (I)	GPA Range 5 (J)	Mean Difference (I-J)	Standard Error	Significance	Lower Bound	Upper Bound
1	2	5.60*	.89	<.05	3.45	7.75
	3	7.79*	.90	<.05	5.61	9.97
2	1	-5.60*	.89	<.05	-7.75	-3.45
	3	2.19*	.88	.04	.07	4.32
3	1	-7.79*	.90	<.05	-9.97	-5.61
	2	-2.19*	.88	.04	-4.32	-.07

*Mean difference is significant at the .05 level

Note. 1 = randomized test format, 2 = arbitrary test format, 3 = balanced test format

Chapter Five

Conclusions & Discussion

Based on this study's findings, arguments are presented that support or refute the rationale for using randomized answer-keys in classroom multiple-choice tests. Two comparisons of the test score group means among the test formats are presented with respect to the hypotheses findings. The first comparison looks at the combined ability group means across the different test formats; the second comparison focuses on the interaction of test format with student ability. Each comparison is followed with a discussion of the findings. These discussions are conceptual as opposed to empirical. They attempt to present a logical explanation for the findings with respect to the arguments presented in Chapter One for undertaking this study. The nature of the arguments for or against the use of a particular test-key format will depend on the resulting distribution of the test-key. Therefore, Table 11 summarizes the distributions of the test-keys used in this study. These distributions are referred to in the comparison discussions. Additional analyses follow which include a discussion of validity and student ability, and item analysis. Item analysis considers both item difficulty and item discrimination. A summary of findings and limitations, as well as recommendations for future study on the use of various multiple-choice answer-key test formats concludes the chapter.

Table 11

Distribution of Correct Options for Each Test Format

	Test Format		
Option	Randomized	Arbitrary	Balanced
A	.24	.18	.22
B	.10	.22	.20
C	.32	.18	.20
D	.14	.16	.18
E	.20	.26	.20

Note. One answer in the perfectly-balanced key was changed from D to A. This made the key slightly imbalanced, though not to any significant degree. The change was the result of conforming to a particular item-writing guideline to present the options in a logical order.

Comparison 1

Differences among test scores for the combined ability groups across test formats were not significant ($p > .05$). No significant differences in group means were found for the combined ability group means across test formats. The combined ability randomized group mean ($M = 32.79$) was slightly greater than the combined ability arbitrary group mean ($M = 32.25$), as well as the combined ability balanced group mean ($M = 31.79$).

Discussion of comparison 1.

Based on research expectations, it was expected that the combined ability randomized group mean would be lower than the combined ability arbitrary or combined ability balanced group means. This expectation was not supported by this study's findings. For guessing test-takers to be more successful in the randomized-key test format over the others, they must either: C1. Be able to use a randomized guessing device; C2. Be able to use a successful guessing

strategy such as edge-aversion to better effect than students would in an arbitrary test-key; or C3. Be able to use a successful guessing strategy such as the underdog to better effect than students would in a balanced test-key. Use of randomized guessing devices was perhaps not feasible for this particular test, making C1 unfeasible.

Edge aversion in the arbitrary test format exhibited by the test-maker and test-takers would give the latter an advantage when guessing on arbitrary test formats. No such occurrence would be found in the randomized group, unless the key happened to be edge-averse as well. To investigate this possibility, the distribution of correct options for each test format provided in Table 11 was examined. The results of the answer-key distribution do not confirm that the randomized key was edge-averse. Options A and E appear more times (0.24 and 0.20, respectively) than their neighboring options B and D (0.10 and 0.14, respectively). Middle-bias seemed to be present since option C appeared more than any other option (0.32). However, it is important to remember that a randomized key will change 'unpredictably' from key to key. Therefore, inferences from the distribution of this study's randomized key can only be made about this particular key's pattern. A different randomized key may likely produce a different distribution and hence different inferences. Although this expectation of unpredictability in test-key distribution is convenient for test-makers, it is difficult for educational researchers wanting to draw conclusive inferences from the data in studies that randomize test-keys. Nonetheless, investigation of the arbitrary test-key distribution also reveals no edge-aversion. Since neither the randomized key nor the arbitrary key were edge-averse, an edge-averse strategy was not feasible. Therefore, C2 is most likely unfeasible.

Examination of Table 11 shows an unbalanced randomized test key. Nonetheless, Attali and Bar-Hillel (2003) premise that use of the underdog strategy is unfeasible unless an

interaction between high ability and test format occurs. However, comparison 1 does not account for interactions, but instead looks at the main effects of combined ability groups across test formats. From the perspective of comparing combined ability groups between the randomized and balanced test formats, C3 is again perhaps unfeasible given Attali and Bar-Hillel's premise.

Although randomized keys are expected by chance to balance out over long runs, it is more likely that any true random device will produce non-perfectly balanced keys a great number of times. It follows that randomized keys can be equated more often than not with some form of a non-balanced key. The results of this study show that combined ability group mean test scores from the non-perfectly balanced keys (i.e. the arbitrary and randomized keys) do not significantly differ from their perfectly-balanced ones.

Comparison 2

Differences among test scores for the interactions between students' ability levels and the test format they were assessed under were not significant ($p > .05$). Interaction of student ability and test format showed no significant group mean differences across high ability levels for the different test formats. The high ability randomized group mean ($M = 37.74$) was slightly greater than the high ability arbitrary group mean ($M = 35.83$) as well as the high ability balanced group mean ($M = 36.87$). No significant group mean differences were found across the average ability levels for the different test formats. The average ability randomized group mean ($M = 31.44$) was equal to the average ability arbitrary group mean (31.45); both means were greater than the average ability balanced group mean (30.79). No significant group mean differences were found across low ability levels for the various test formats. The low ability randomized group mean (M

= 29.11) was slightly lower than the low ability arbitrary group mean (M = 29.41), though slightly greater than the low ability balanced group mean (28.61).

Discussion of comparison 2.

The findings of comparison 2 did not corroborate the arguments found in the educational measurement literature and summarized in Chapters One and Two of this study. The factors student ability and test format showed no significant interactions at any of the factor levels. Recall that edge-aversion exhibited at all levels of student ability seemed to have benefited students in past tests where the test-key format was arbitrarily formed by test-makers (Bar-Hillel & Attali, 2002). The expectation was thus that low ability student test scores for students in the arbitrary test format would be higher than those of low ability student scores in the balanced test format. Although this study's findings met these expectations, no statistical significance was found for any of the group mean differences. In contrast, the higher group mean for the high ability balanced group over the high ability arbitrary group may have been due to the former group's success at using the underdog strategy to better effect when guessing as opposed to the latter group's success at using edge-aversion. The underdog works by getting as many items correct prior to guessing (Attali & Bar-Hillel, 2003). Since the ability to get more items correct correlates with ability level, the higher group mean for the interaction of ability and test format would only be found at high ability levels. This study did not try to establish when guessing occurred, or what guessing strategies were used. Therefore, conclusions cannot be made that substantiate edge-aversion as a reason for any difference in group means.

Without knowing how often and what type of guesses occurred, use of the underdog strategy or edge-aversion can still be possible if the balanced key was indeed balanced and the

arbitrary key edge-averse. Examination of the balanced-key distribution shows a very balanced key as would be expected. However, the expectation to find edge-aversion in the arbitrary-key distribution was not met. Option E – an edge position – appears more frequently than any other option (0.26), especially when compared to option D (0.16). Option B appears more times than option A (0.22 compared to 0.18, respectively), indicating the possibility of edge-aversion. This would be satisfied if guesses in the arbitrary test format occurred for students exhibiting edge-aversion when guessing on items where option B was correct. Students would have then had to select option B as their guess as opposed to option D.

Although the interaction of ability level and the randomized test format was expected to occur at all ability levels, of particular interest was the interaction of this test format with high ability students. Since no true randomized device such as a deck of cards or unbiased coins could be used as a viable guessing strategy in the randomized test formats, ability would not be a factor when students had to guess on the randomized test forms. When comparing group means for high ability levels across test formats, randomized test-key formats were therefore expected to reduce CIV related to guessing for high ability students. In other words, high ability students answering the randomized test-key would have lower group means than high ability students answering the arbitrary or balanced test formats. The findings of this study did not meet these expectations. The high ability randomized test format group means were higher than the high ability arbitrary and balanced group means. This may have been due to the reliability of the tests and the relevant size of the test formats' standard deviations in comparison to each other. Examination of the standard deviations for the high ability randomized and balanced groups in table 8 reflect this observation.

Additional Analyses

Student ability and validity evidence.

Validity evidence can be found by investigating mean differences among the ability groups. It is expected that test scores will significantly differ in the order of high, average, and low ability. This positive correlation is evidence that the test items ranged in difficulty as well as discriminated between high and low ability students. Student ability was considered an independent factor in this study's analysis. The univariate ANOVA results in Table 9 provide effects of the main factors in this study. Main effects of student ability revealed that test scores were significantly different among students with differing ability levels, $F(2, 154) = 39.38$, $p = .00$, partial $\eta^2 = .32$.

Bonferonni's post hoc test was conducted to determine which ability groups were significantly different in test scores. Results of the Bonferroni test are found in Table 10. It was found that test scores of students with high ability are significantly different from all other ability groups. In addition, a significant difference, though small, was found between the low and average ability groups. Estimates of effect size indicated low strength in associations for all ability groups. These findings are consistent with the evidence from the educational measurement and testing literature that there are significant differences among ability levels for test scores, regardless of the test format. These differences provide support for the validity of the results since it is expected that scores will differ in decreasing order of magnitude for high, average, and low ability.

Item difficulty.

Item difficulty was assessed as part of the item statistics analysis. Item difficulty is usually defined as the proportion of students getting an item correct. The item difficulty index ranges from 0.00 to 1.00. The index range is indirectly proportional to the difficulty of items. Easier items have p-values approaching 1.00. Harder items have p-values approaching 0.00. University of Washington (2002) classifies items as easy for p-values less than 0.50, moderate for p-values between 0.50 and 0.84, and high for p-values greater than 0.85.

Examining the p-values of Appendix E across all ability groups and for the various test formats reveals warning trends in the item difficulty of this particular test. Items 2, 6, and 43 were consistently easy across all ability groups and formats. This is in contrast to items 1, 4, 5, 16, 17, 18, 28, 29, and 37 that showed a very difficult trend. These items should be reviewed further with the test-maker before assessment further use. Establishing the instructor's purpose to include these questions may shed further light on whether they should be retained or dropped.

Item discrimination.

Item discrimination refers to how well an item discriminates across ability groups. It is based on a special type of correlation coefficient termed the point-biserial correlation coefficient. The correlation is between a dichotomous variable and a continuous variable. In this case, an item score of 0 (incorrect) or 1 (correct) is correlated with a total test score that takes the sum of all items scored 1. Because it is a type of correlation, the discrimination index ranges from -1.00 to +1.00. Negative point-biserials result when more students in the low performing group answer an item correctly as opposed to the high performing group. It could also mean a mistake in the answer-key has occurred. When students in the high performing group answer more items

correctly than those in the low performing group, the result is a positive index. Homogeneity across ability groups means that the students are more likely to perform relatively close to one another and therefore will have similar test performance. This results in low discrimination between high and low groups.

Based on the point-biserials provided in Appendix F and the classification data for the point-biserials in Appendix G, certain items are in need of stringent analysis and revision. In particular, items 1, 3, 16, and 20 have poor discrimination. Based on item 1's high difficulty and poor discrimination, it may be recommended that the test-maker drop this item from further test use. Item analysis can also be determined through the investigation of reliability results. Based on the reliability results presented in Chapter 4, Cronbach alphas were computed for items deleted. These are found in Table 12.

Table 12

Cronbach Alphas if Items Deleted Across Test Formats

	Test Format		
Item	Randomized	Arbitrary	Balanced
1	.80	.70	.73
2	.79	.69	.72
3	.80	.70	.73
4	.79	.69	.72
5	.80	.68	.73
6	.79	.69	.72
7	.80	.69	.72
8	.79	.69	.72
9	.79	.69	.72
10	.79	.69	.72
11	.79	.69	.72
12	.79	.68	.72
13	.79	.69	.73
14	.79	.69	.72
15	.79	.69	.72
16	.80	.69	.73
17	.79	.69	.72
18	.80	.69	.72
19	.79	.69	.72
20	.80	.68	.73
21	.79	.68	.72
22	.79	.69	.73
23	.79	.68	.72
24	.79	.69	.72
25	.79	.70	.72
26	.79	.70	.72
27	.79	.68	.72
28	.80	.70	.72
29	.79	.68	.72
30	.79	.68	.72
31	.79	.70	.72
32	.79	.68	.72
33	.79	.69	.72
34	.79	.68	.72
35	.79	.68	.71
36	.79	.68	.72
37	.79	.69	.73
38	.79	.69	.72

		Test Format	
Item	Randomized	Arbitrary	Balanced
39	.79	.68	.72
40	.80	.70	.73
41	.79	.69	.72
42	.79	.67	.71
43	.80	.69	.72
44	.79	.68	.72
45	.79	.68	.72
46	.79	.68	.71
47	.79	.68	.71
48	.79	.69	.71
49	.79	.68	.72
50	.79	.68	.72

An item with a Cronbach alpha that is lower than the Cronbach alpha if the item were deleted is cause for concern. Essentially, this means that the reliability of the total test scores improves if that item is taken out of the analysis. Therefore, that item is potentially problematic. The Cronbach alpha for the randomized test was .80. Cronbach alpha did not surpass this statistic for any item that was deleted in the randomized test format. The Cronbach alpha for the arbitrary test format was .69. Cronbach alpha surpassed this statistic when several items were deleted individually. These items need further attention. They are, in order, items 1, 3, 25, 26, 28, 31, 40. The Cronbach alpha for the balanced test format was .72. Several items again caused the Cronbach's alpha to increase if these items were deleted individually. These are items 1, 3, 5, 13, 16, 20, 22, 37, 40. The low Cronbach alphas may be the result of several things. It is possible that the high variance in the test scores for the arbitrary and balanced test formats were affecting the reliability of the test. This may have been due to the way the test was administered or from other internal validity threats that occurred.

Summary of Findings

The purpose of this study was to examine the impact of student ability and different answer-placement strategies in classroom multiple-choice tests. The following conclusions can be summarized from the data and analysis:

1. No significant differences were found for the main effect of answer-placement strategy on test scores. Combined ability group means for the different test formats did not differ to any significant degree.
2. No significant effect was found for the interaction of student ability and answer-placement strategy on students' test scores. Based on Bar-Hillel and Attali's (2002) study rationale, this study hypothesized that students in the randomized test format would do significantly worse than students in the arbitrary and balanced test formats due to the lack of viable guessing strategies available for randomized test-keys. Although no statistically significant differences occurred, the randomized group mean was actually slightly higher than the arbitrary and balanced group means.

The results of this study contribute validity evidence that strategies to vary the position of correct answers in classroom MC tests may have no significant impact on students' test scores. However, this inference may not extend to large-scale educational tests. It is probably the case that fewer guessing strategies are used in low-stakes classroom tests as opposed to high-stakes tests. This could be due to less test-wise training being available or offered for a typical classroom assessment as opposed to high-stakes educational assessments, such as the SAT prep courses. Most classroom assessment preparation is primarily focused on the content being tested. Classroom instructors may thus be advised to focus on assessing test content as opposed to combating use of test-wise guessing strategies.

Although these results are inconclusive, classroom instructors may feel at ease to vary the position of correct answers by any of the methods examined in this study. They can thus choose a method that is most convenient to them and be fairly confident that the type of test-key format will not affect their students' test scores significantly. In light of educational research that has been conducted in the past, the file-drawer problem is considered. The file-drawer problem refers to studies that have been looked at in a negative light because their results were insignificant. However, the insignificance of this study may ultimately shed light in a positive way for everyday classroom instructors undertaking the process of creating classroom MC tests.

Limitations.

1. Validity coefficients were not calculated because of a design limitation in collecting the data. Although GPA was used as a proxy variable for student ability, students were asked to mark which noncontiguous category of GPA ranges their GPA fell within. This is opposed to requesting their actual self-reported GPA scores before establishing the noncontiguous cutoffs for GPA categories, as both Cronbach and Snow (1977) and Trevisan (1990) had done. Because of this, raw scores were not correlated with GPAs to calculate exact validity coefficients.
2. Due to the collection of student's GPA within noncontiguous GPA categories as opposed to actual GPA scores, a large average ability group was formed. This average ability group had twice an n as the high and low ability groups. Analysis procedures were conducted to randomly select from the average ability group so that cell sample sizes could be equal in order to facilitate factorial ANOVA assumptions. This process discounted cases from the factorial analysis.

3. The subject-matter knowledge assessed in this study was undergraduate general biology material. Some of this material may only be representative of certain types of cognitive processing tasks.
4. Randomizing answer-keys is an 'unpredictable' process. Although randomizing theoretically leads to balanced keys in the long-run, a randomized test-key may turn out to be severely skewed or overly balanced on any particular trial. Attali and Bar-Hillel (2003) discuss the use of balanced-key policies to control for the outcomes of certain sequences, such as runs. Course instructors will be unable to control for such sequences if they truly randomize the answer-key.
5. There was no way to know the extent to which successful guesses were made. Furthermore, the extent to which established guessing strategies such as edge-aversion or the underdog were used is unclear. This lack of knowledge limits the degree to which comparisons could be made between interaction and main factor group means.
6. Based on the item analyses, several items warrant further attention. Several items warrant further attention before being used again. Most particularly noted are items 1, 3, and 40. Content concerns may have been the cause behind the low quality of many of these items.
7. Observed power statistics were recorded from the univariate ANOVA results. Observed power for the interaction effects between student ability and test format was 0.13. Observed power for the main effect of test format was 0.10. These results suggest that the actions taken to reduce cell sample sizes in order to create more homogenous groups significantly decreased the power of the study. The corresponding effect sizes for the main effect of test format and the interaction effect with student ability were extremely low. Increasing the cell sample size may therefore have no impact on the power of the

study. However, increasing the cell sample size originally is a different procedure from having a larger sample size to begin with and then randomly selecting cases out of this sample to include in inferential statistical analysis. Given that the latter occurred in this case, more research is needed on the impact of reducing sample sizes to facilitate ANOVA assumptions such as was done in this study. Although no statistically significant interaction or main effects were found, the underpowered study might be cause for concern that a type II decision error was committed. Type II errors occur when the null hypothesis is indeed false, but a conclusion is made that it is true.

8. Internal threats are always a concern in social science and educational measurement studies. A limitation of this study with reference to internal validity is based on the method of test delivery to the students. Although the course instructor was present for test administration, he was accompanied by several teaching assistants. These assistants varied depending on the section of the course. No protocols were developed or suggested prior to administering the test to ensure that the delivery procedures were similar for all the sections.
9. An external validity threat may have occurred when making inferences based on the inferential statistics from the univariate ANOVA test. The actions taken to reduce the cell sample sizes by random selection of cases may have had an impact on the results of this study. A persons-treatment interaction effect could have thus had an impact in this study.

Recommendations for further research.

The following recommendations are made based on the findings and limitations in this study:

1. Collect actual GPA scores to allow for the non-contiguous cutoffs to be made appropriately depending on the distribution of data so that each ability level will be roughly equal in sample size. Larger sample and cell sizes will thus be created from these cutoffs, and more cases therefore utilized in the ANOVA.
2. Validity evidence for or against the use of one particular answer-placement strategy must be ongoing (Haladyna, 2004). Further evidence should investigate the impact of answer-placement strategies – randomized, balanced, and otherwise – with other subject-matter knowledge material.
3. Further analyses that examine the impact of sequences (for example, runs greater than 3) on test scores should be conducted. Also, further studies should intentionally manipulate the correct answer-positions in ways other than those discussed in this study (for example, using palindromes). The impact of such test-key manipulations should then be investigated.
4. Students' actual guessing strategies for various test formats should be elicited by manipulating the condition where they are told the answer-key placement strategy.
5. Based on item analyses of difficulty, discrimination, and test score reliability, it is recommended that further discussions with the course instructor be held to determine the inclusion of some items – in particular items 1, 3, and 40 – in the test. It is recommended that a review of these items occur based on the table of specifications for creating MC tests, as well as using the item-writing guidelines discussed in Haladyna (2004).
6. Develop and validate a protocol for test administration that will be used for each section of the test. Respectfully request that the course instructor and his teaching assistants

follow this protocol for administering the test in order to minimize internal validity threats from occurring.

Attali and Bar-Hillel (2003) argue for randomizing test-keys as opposed to balancing or arbitrarily assigning the correct answer positions. Haladyna (2004) supports this notion in his review of the item-writing guideline that the location of correct options should be varied randomly. The research expectations in this study were based on this literature with a focus on classroom multiple-choice tests. However, the results of this study do not confirm most of the research expectations, and when they do, they do not do so to any significant degree. In the effort to establish a science of valid item-writing guidelines, more evidence is needed to corroborate the guideline to randomly vary the location of correct options in classroom multiple-choice tests.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME] (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40 (2)*, 109-128.

Ayton, P., & Falk, R. (1995). *Subjective randomness in hide-and-seek games*. Paper presented at the 15th bi-annual conference on Subjective Probability, Utility, and Decision Making, Jerusalem, Israel.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician, 56 (4)*, 299-303.

Bar-Hillel, M., & Wagenaar, W.A (1991). The perception of randomness. *Advances in Applied Mathematics, 12 (4)*, 428-454.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple-choice tests: A case study in irrationality. *Mind & Society, 4 (1)*, 3-12.

Berg, I.A., & Rapaport, G.M. (1954). Response bias in an unstructured questionnaire. *The Journal of Psychology, 38*, 475-481.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31 (1)*, 37-50.

Brown, G.H, Carroll, C.G. (1984). The effect of anxiety and boredom on cognitive test performance. *Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA. April 23-27)*.

Burton, R.F. (2006). Sampling knowledge and understanding: How long should a test be? *Assessment & Evaluation in Higher Education, 31 (5)*, 569-582.

Christenfeld, N. (1995). Choices from identical options. *Psychological Science, 6*, 50-55.

Claman, C. (1997). *10 real SATs*. New York: College Entrance Examination Board.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd Ed.)*. Lawrence Earlbaum Publishing.

Cohn, E., Cohn, S., Balch, D.C., & Bradley, Jr., J. (2004). Determinants of undergraduate GPAs: SAT scores, high school GPA, and high school rank. *Economics of education review, 23 (6)*, 577-586.

Crehan, K.D., Koehler, R.A., & Slakter, M.J. (2005). Longitudinal studies of test-wiseness. *Journal of Educational Measurement, 11 (3)*, 209-212.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp.443-507). Washington, DC: American Council on Education.

Cronbach, L.J. (1988). Five perspectives of the validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp.3-18). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cronbach, L.J., & Snow, R.E. (1977). *Aptitude and instructional methods*. New York: Wiley & Sons, Inc.

Dobbins, G.H., Fahr, J.L., & Werbel, J.D. (1993). The influence of self-monitoring and inflation of grade-point averages for research and selection purposes. *Journal of Applied Social Psychology, 23*, 321-334.

Downing, S.M. (2002a). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference. *Academic Medicine, 77 (10)*, S103-S104.

Downing, S.M. (2002b). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education, 7*, 235-241.

Downing, S.M., & Haladyna, T.M. (2006). *Handbook of test development*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Drummond, R.J., & Jones, K.D. (2006). *Assessment procedures for counselors and helping professionals (6th edition)*. Upper Saddle River, NJ: Prentice Hall.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement (5th ed.)*. Englewood Cliffs, NJ: Prentice-Hall.

Educational Testing Service (2003). *ETS Standards for fairness and quality*. Princeton, NJ: Author.

Falk, R. (1975). *The perception of randomness*. Unpublished doctoral dissertation (in Hebrew, with English abstract), Hebrew University, Jerusalem, Israel.

Gardner, H. (1986). *The mind's new science: A history of the cognitive revolution*. New York: Basic Books.

Gibb, B.F. (1964). Test-wiseness as a secondary cue response. *Unpublished Doctoral Dissertation*. Ann Harbor, MI: University Microfilms (No.64-7643).

Graduate Management Admission Council (1992). *GMAT review, The official guide*. Princeton, NJ: author.

Green, K., Sax, G., & Michael, W.B. (1982). Validity and reliability of tests having different numbers of options for students of differing levels of ability. *Educational and Psychological Measurement, 42*, 239-245.

Guilford, J.P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Haladyna, T.M. (1992). The effectiveness of several multiple-choice formats. *Applied Measurement in Education, 5*, 73-88.

Haladyna, T.M. (2002). Supporting documentation: Assuring more valid test score interpretations and uses. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment for all students: Validity, technical adequacy, and implementation* (pp. 89-108). Mahwah, NJ: Lawrence Earlbaum Associates.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ: Lawrence Earlbaum Associates.

Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 37-50.

Haladyna, T.M., & Downing, S.M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51-78.

Haladyna, T.M., & Downing, S.M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement, 53*, 999-1010.

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23 (1)*, 17-27.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15 (3)*, 309-334.

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Item response theory: Principles and applications (2nd ed.)*. Boston, MA: Kluwer-Nijhoff.

Holtzman, M. (2008). Demystifying application-based multiple-choice questions. *College Teaching, 56 (2)*, 114-120.

Holtzman, K., Case, S.M., & Ripkey, D. (2002). Developing high quality items quickly, cheaply, consistently – pick two. *CLEAR Exam Review*, 16-19.

Lipsey, M.W. (1990). *Design sensitivity: Statistical power for experimental research*. California: Sage Publications.

Marks, A.M., & Cronje, J.C. (2008). Randomizing items in computer-based tests: Russian roulette in assessment. *Journal of Educational Technology & Society, 11 (4)*, 41-50.

Mertler, C., & Vannatta, R. (2005). *Advanced and multivariate statistical methods*. Pyrczak Publishing.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.13-104). New York: American Council on Education and MacMillan.

Miller, P.M., Fagley, N.S., & Lane Jr., D.S. (1988). Stability of the Gibb (1964) experimental test of testwiseness. *Educational and Psychological Measurement, 48 (4)*, 1123-1127.

Mohr, L.B. (1995). *Impact analysis for program evaluation.* California: Sage Publications.

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practices, 11*, 9-15.

Offir, R., & Dinari, N. (1998). *Chemistry for matriculation exams*. Tel Aviv, Israel: Lachman.

Open University, The (1998). *Psychological development: A study guide*. Tel Aviv, Israel: Author.

Phelps, R.P. (1998). The demand for standardized testing. *Educational Measurement: Issues and Practice, 17* (3), 5-19.

Powers, D.E., & Rock, D.A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93-118.

Roediger, H.L., & Marsh, E.J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31 (5)*, 1155-1159.

Rubinstein, A., Tversky, A., & Heller, D. (1996). Naïve strategies in competitive games. In W. Albers, W. Guth, P. Hammerstein, B. Moldovanu, & E. van Damme (Eds.), Understanding strategic interaction. New York: Springer-Verlag.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation (3rd ed.)*. Belmont, CA: Wadsworth Publishers.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Slavin, R. (2006). *Educational research in an age of accountability*. New Jersey: Allyn & Bacon Publishers.

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Stiggins, R. (1991). *Student-involved classroom assessment (3rd ed.)*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Supon, V. (2004). Implementing strategies to assist test-anxious students. *Journal of Instructional Psychology, 31 (4)*, 292-296.

Thurstone, L.L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press. (Reprinted in 1968 by the Psychometric Society).

Trevisan, M.S. (1990). Reliability and validity of multiple-choice examinations as a function of the number of options per item and student ability. *Dissertation Abstracts International*, DAI-A51/09. (UMI No. AAT 9104306). Retrieved March 02, 2010 from Dissertations and Theses Database.

Trevisan, M.S., & Sax, G. (1990). Reliability and validity of multiple-choice examinations as a function of the number of options per item and student ability. *Paper presented at the Annual Meeting of the American Educational Research Association (74th,Boston, MA. April 16-20).*

Trevisan, M.S., Sax, G., & Michael, W.B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement, 51*, 829-837.

Wood, R. (1977). Multiple-choice: A state of the art report. *Evaluation in Education: International Progress, 1*, 191-280.

Yoel, A. (1999). *A collection of questions from the matriculation tests in biology*. Tel Aviv, Israel: Lachma.

Appendix A

Student Consent Form

WASHINGTON STATE UNIVERSITY

CONSENT FORM

Researchers

Dane Christian Joseph, PhD Candidate, Educational Leadership & Counseling Psychology, 208-301-4136, danechristian@mail.wsu.edu
Michael S. Trevisan, Professor, Educational Leadership & Counseling Psychology, 509-335-8611, trevisan@wsu.edu

Research Statement

The purpose of this consent form is to provide you with the information needed to decide whether you wish for your test scores to be used in this study. Please read the form carefully. Please feel free to ask any questions about the purpose of this research, the risks and benefits of participation, your rights in participating, and anything else that is unclear about this research or consent form. When we have answered all your questions, you may decide if you want your test scores to be used in this study or not. This process is called 'informed consent'. You will be provided with a copy of this form for your records. The original copy will be kept by your class instructor – Professor Andrew Storfer.

The purpose of this study is to compare various methods for varying the location of correct answers in multiple-choice test answer keys. This study has the potential to help classroom teachers create better multiple-choice tests, specifically answer-keys. You will be required to take the classroom test your course professor has prepared for you. At the end of this exam, you will be asked to provide descriptive data on your GPA range, gender, race/ethnicity, and age. None of these variables can be used as personal identification information. However, you can allow your scores to be used but opt out of answering these descriptive variables if you choose.

Dane Christian Joseph	Dane Christian Joseph	02/08/2010
Printed Name of Researcher	Signature of Researcher	Date

Subject Statement

This study has been explained to me. I have had a chance to ask questions about my role in the study. I volunteer to let my test scores be used in this research study. If I have any further questions about the research, I can ask one of the researchers listed above. I will receive a copy of this consent form.

Printed Name of Subject	Signature of Subject	Date

Appendix B

Test Instrument

Name_____

Biology 102: Spring 2010 Exam I
Please be sure to fill out your name, ID# (under special codes) and lab section number on the scantron. This exam contains 50 multiple choice questions worth 2.5 points each. There is also one bonus question worth 2 points (total possible score 127 points). Choose the **best** answer for each question. Please also be sure to write down the **FORM (A, B, or C)** of your exam next to your name on the scantron form.

1. Which is the most lethal form of anthrax?
   a. cutaneous
   b. inhalation
   c. gastrointestinal
   d. liver
   e. neuronal

2. Your lawnmower won't run and you speculate that the reason is that it is out of gas. In scientific terminology, such reasoning would best be described as:
   a. forming conclusions from the results of experiments
   b. developing an observation based on a hypothesis
   c. developing a hypothesis based on an observation
   d. controlling variables in a repeated manipulation of nature
   e. testing a prediction generated from a hypothesis

3. Why might organisms, from fish to scorpion flies prefer symmetrical members of the opposite sex?
   a. symmetrical individuals are always larger
   b. symmetry implies that an individual might be better at fighting off parasites and thus more "fit"
   c. symmetry ensures no diseases
   d. symmetrical individuals always produce more offspring
   e. symmetrical individuals are older

4. Suppose bioterrorists introduce a new disease into the US. Which approach would best give us complete information about this disease and how to stop its spread?
   a. DNA analysis
   b. analysis of symptoms
   c. organismal analysis to determine how it is spread
   d. environmental analysis to determine where the disease lives
   e. an integrated approach that includes hierarchical study from genes to ecosystems

5. Dolphins, sharks and turtles all exhibit streamlined shapes that help them move smoothly through the water. Such similar forms in distantly related organisms that live in similar environments are examples of _____ evolution.
   a. divergent
   b. convergent
   c. co-
   d. punctuated
   e. adaptive

6. Where did the earliest human ancestors originate?

a. Africa
b. North America
c. South America
d. Asia
e. Europe

7. People who do not finish the entirety of their antibiotic prescriptions likely:
a. save money by saving some pills for the next time they get sick
b. make themselves immune to the antibiotic
c. artificially select for bacterial resistance to the drug, thereby reducing the drug's effectiveness in the future
d. promote the health of their immune system by reducing dependence on drugs
e. decrease exposure of bacteria to that antibiotic and make sure the antibiotic will be effective in the future

8. Whose theory that overpopulation in humans caused war, famine and disease and thus a "struggle for existence" influenced Darwin's theory of evolution by natural selection?
a. Charles Lyell
b. Alfred Wallace
c. Erasmus Darwin
d. Thomas Malthus
e. Lamarck

9. Which of the following does not provide evidence for evolution by natural selection?
a. Transitional fossils
b. Direct observation
c. All of the above provide evidence of natural selection
d. That all life forms are related in some way, and those relationships can be recreated
e. The fact that people who run a lot tend to have faster kids

10. Why is HIV/ AIDS so hard to treat?
a. it is susceptible to drugs, but most patients do not finish their treatment
b. it reproduces so quickly that drugs can't kill it
c. it reproduces very slowly
d. HIV's normal host is not humans
e. it mutates quickly and naturally resistant strains increase in frequency with drug treatment

11. The word "evolution" in biology literally means:
a. natural selection
b. mutation
c. descent with modification
d. divergence
e. randomness

12. A slight change in beak shape in a population of Darwin's finches after a couple of harsh winters with low food resources, whereby individuals with thicker beaks survived more and had more offspring of those finches with thinner beaks is an example of:
a. artificial selection
b. macroevolution

c. microevolution
d. inheritance of acquired characteristics
e. all of the above

13. According to the hypotheses of some sociobiologists and based on the material in class, what is the *second most important* trait males value in a female mate?
a. fidelity
b. youthfulness
c. maturity
d. attractiveness
e. resources

14. An evolutionary approach to treating patients with TB (tuberculosis) would not include:
a. observational therapy to ensure patients finish their prescription
b. testing patients for drug resistance
c. using the best drug straight off and stopping treatment when the patient feels better to avoid resistance
d. treating drug resistant strains with at least two additional drugs
e. treat patients with multidrug resistant strains for at least 18 months

15. Ancient whales are seen to have which feature from fossil evidence:
a. vestigal femur and pelvis
b. wings
c. front legs
d. a and c
e. all of the above

16. An example of a transitional fossil form would be:
a. horse
b. Archeopteryx
c. whales
d. all of the above
e. none of the above

17. Which of the following behaviors likely has the greatest genetic influence, as opposed to environmental influence?
a. imprinting
b. rooting reflex
c. habituation
d. learning the dialogue to a South Park episode
e. studying for biology 101/102

18. Which of the following is the hominid that lived with modern *Homo sapiens* most recently?
a. *Homo erectus*
b. *Homo sapiens neanderthalus*
c. *Homo florensis*
d. *Homo habilis*
e. *Homer simpsonus*

19. Which of the following was not one of Darwin's observations?
    a. there is intense competition among members of a population
    b. there is differential reproductive success among individuals in a population
    c. most individuals have an equal chance to survive and reproduce
    d. some characteristics are heritable and passed on to offspring
    e. there is considerable variation in members of a population

20. Young human babies respond to objects with a lot of contrast (e.g., black and white) so long as they are moving. It doesn't matter if the object is round or another shape – the baby just cues in on contrast and movement. This is an example of:
    a. sign stimulus
    b. imprinting
    c. adaptive learning
    d. habituation
    e. instinctive behavior

21. When researchers attempt to answer why various animals form social groups, they use:
    a. cost-benefit analysis
    b. homology
    c. phylogeny
    d. behavioral analysis
    e. bipedalism

22. All of the following are disadvantages to sociality, <u>except:</u>
    a. competition for mates
    b. competition for food
    c. disease spread
    d. access to scarce resources
    e. attracting predators

23. Why might species exhibit altruism, as in the example of ground squirrels who call to alert nearby squirrels that a predator is present, at the potential cost of being eaten and thus sacrificing themselves?
    a. altruism is a good thing to do
    b. inheritance of acquired characteristics
    c. some behaviors just don't make sense, although they are genetically engrained
    d. kin selection
    e. they are better equipped to escape predators than other individuals in the population.

24. Under a mating system of polyandry, you should see exaggeration in which gender?
    a. males
    b. females
    c. variable exaggeration
    d. neither males nor females
    e. both males and females

25. You conduct a study in a species that you think is monogamous, but you find that offspring in a cohort (e.g., nest) are not all fathered by the same dad. Which of the following might be reasons for the females to be "cheating?"
    a. "sexy son" hypothesis

b. looking for more symmetrical males
c. possibly looking for more resources
d. none of the above
e. all of the above

26. In the experiment from the "Evolution" series where females smelled t-shirts that guys slept in for a couple of nights, what feature did they preferentially choose?
a. males with complementary immune (MHC) complexes
b. males that smelled symmetrical
c. males that had testosterone markers in their smell
d. "sexy" males
e. all of the above

27. Louis Pasteur was responsible for which of the following?
a. germ theory of disease
b. establishing the field of microbiology
c. anthrax vaccine
d. Pasteurization
e. all of the above

28. You are working on your farm and start to notice that aphids are increasing in numbers in your potato crop. You have been using the best pesticide on the market for the last 5 years, but damage to your crop keeps increasing. Which of the following is **not** an application of evolutionary biology to deal with this problem?
a. start adding spiders to your crop
b. introduce a natural disease that kills aphids
c. purchasing potato stocks that have genes inserted in them that are toxic to insects
d. switching to the next best pesticide for awhile
e. bioengineering potatoes resistant to insect damage

29. Which of the following is **not** a prediction of descent with modification?
a. relatedness of life forms
b. homology
c. inheritance of acquired characteristics
d. change through time
e. life on Earth is old

30. Which of the following is the *best* example of microevolution?
a. relationship of humans and apes
b. a disease that wipes out every member of a population except those that are resistant
c. the fact that squids and dogs both have eyes
d. wolves and dogs share a common ancestor
e. all of the above are examples of microevolution

31. You are a geneticist who studies DNA variation brought in to study the threat of possible bioterrorism spread of small pox. Which type of biologists would you add to your team?
a. cell biologists
b. population biologists
c. landscape biologists

d. immunologists
e. all of the above

32. Sodium, a single atom and very small particle important for cellular function is in higher concentration outside the cell than inside the cell; this molecule will likely undergo:
a. diffusion
b. diffusion by osmosis
c. facilitated diffusion
d. active transport
e. exocytosis

33. Scientists are currently debating when which aspect of hominid evolution occurred that separates us from apes and chimpanzees?
a. use of tools
b. use of language
c. bipedalism
d. gossip
e. "memes"

34. After traveling into the nucleus of a T-cell, HIV needs to instruct the cell to make its proteins. Where are these proteins made?
a. nucleus
b. golgi apparatus
c. ribosome
d. mitochondria
e. cell membrane

35. Which of the following structures is a component of plant cells but **not** of animal cells?
a. cell wall
b. nucleus
c. ribosomes
d. cell membrane
e. mitochondria

36. Which of the following is a process by which cells get rid of large wastes?
a. endocytosis
b. nucleotosis
c. exocytosis
d. diffusion
e. passive transport

37. An animal cell that is hypertonic relative to its solution will:
a. gain water by osmosis and burst
b. gain solutes by diffusion
c. lose solutes by diffusion
d. lose water by osmosis and shrivel
e. experience neither a net gain or loss of water

38. Which of the following explains why we have been able to increase oil yields of corn crops steadily for over 100 years?
a. genetic engineering
b. biocontrol
c. natural selection
d. artificial selection
e. macroevolution

39. When might sexual selection oppose natural selection?
a. choosing symmetrical mates
b. the "sexy son" hypothesis
c. when male traits (e.g., peacock feathers) become so exaggerated as a result of female preferences that make males more susceptible to predators
d. choosing most fertile looking mates
e. none of the above

40. When natural selection is operating on a population, which of the following is most likely to be affected?
a. inheritable characteristics
b. adaptive traits
c. differences in survival
d. differences in reproductive success
e. all of the above

41. Which is the correct order of behaviors going from having mostly genetic influence to mostly environmental influence?
a. baby crying → habituation → imprinting → learning calculus
b. imprinting → baby crying → habituation → learning calculus
c. baby crying → learning calculus → imprinting → habituation
d. imprinting → baby crying → learning calculus → habituation
e. baby crying → imprinting → habituation → learning calculus

42. Peppered moths are generally light in color. In areas where soot has blackened the tree trunks, the frequency of moths shifted to a higher proportion of dark-colored. This example demonstrates:
a. microevolution
b. fossil influence
c. inheritance of acquired characters
d. macroevolution
e. random chance

43. Human and chimpanzee embryos look very similar at certain points in development and, upon close inspection, share a number of features. Why is this?
a. Humans and chimps share a common ancestor
b. Embryos of all organisms, plant and animal, look alike
c. Mutations in human and chimp embryos have caused them to look alike
d. They have come to look like one another due to chance
e. All of the above

44. An example of evolution that has been readily observed is:

a. a chameleon changing colors as it moves from a leaf of one color to a leaf of another color
b. chimpanzees learning sign language
c. the resistance of bacteria to an antibiotic that is used to kill them
d. humans teaching dogs to obey certain commands
e. all of the above

45. How do proteins and other larger molecules *most likely* pass through a cell membrane?
a. via gated channels or carrier protein molecules
b. by osmosis
c. by diffusion
d. exocytosis
e. they can't pass through a cell membrane

46. The close genetic relatedness of humans around the world supports the idea of a "Mitochondrial Eve" dating back about 200,000 years ago. These results support which of the following hypotheses related to human evolution?
a. bipedalism
b. increase in brain size
c. use of tools
d. "Multiregional" dispersal
e. recent dispersal "Out of Africa"

47. Which of the following statements about the cell membrane is *false*?
a. It prevents some molecules from entering the cell
b. It allows some molecules to leave the cell via exocytosis
c. It is the storehouse for each cell's genetic information
d. It receives signals from the outside environment
e. It has both a hydrophobic and hydrophilic component

48. A population evolves when some of its individuals leave more offspring than other individuals from the same population. The result is:
a. extinction of the offspring
b. a change in the ancestors of the population
c. the population's becoming extinct
d. the population always increasing in size
e. an increase in the frequency of certain favorable traits

49. To which cellular structure would you assign the function of sorting materials that results in, for example, formation of a new cell wall in plants?
a. golgi apparatus
b. mitochondria
c. nucleus
d. smooth endoplasmic reticulum
e. rough endoplasmic reticulum

50. Microevolution refers to:
a. the evolution of microorganisms.
b. evolution that occurs on a large scale

c. the changes in trait frequencies that occur in a population over time
d. the splitting of a group of organisms into multiple species
e. all of the above

51. Bonus: Dr. Storfer worked on which animals during his sabbatical in Australia?
a. snakes
b. kangaroos
c. wombats
d. Tasmanian devils
e. frogs

52. You have Exam Form A, **please fill in bubble "A" on your scantron for Question #52.**

Questions 53-56 are OPTIONAL. By filling them out, you agree to participate in an ANONYMOUS study to evaluate different aspects of the multiple choice exam. Answers to these questions will be sent along with your exam answers and overall score, BUT WITHOUT YOUR NAME OR ID# to the department of Education at WSU to assist in a PhD student's research.

53. You are:
a. female
b. male

54. Your estimated GPA is:
a. 3.7-4.0
b. 3.0-3.4
c. 2.0-2.7
d. not shown

55. Your ethnicity is identified as:
a. Caucasian or White
b. African American or Black
c. Hispanic or Latino
d. Asian
e. Native American
f. Pacific Islander
g. Other or Unknown

56. Your age is between:
a. 18-19
b. 20-21
c. 22-23
d. 24 or older

Appendix C

Test Form Keys

Item	Form A Randomized Key	Form B Arbitrary Key	Form C Balanced Key
1	C	C	A
2	C	D	B
3	B	B	B
4	E	E	D
5	B	A	B
6	A	C	D
7	C	E	C
8	D	D	E
9	E	A	D
10	E	B	B
11	C	B	A
12	C	C	A
13	A	D	A
14	C	E	C
15	A	B	C
16	D	E	D
17	B	B	E
18	C	E	C
19	C	C	B
20	A	D	C
21	A	B	E
22	D	A	B
23	D	E	E
24	B	D	A
25	E	E	E
26	A	A	A
27	E	E	E
28	D	C	B
29	C	D	C
30	B	C	A
31	E	E	E
32	A	E	B
33	C	B	D
34	C	A	E
35	A	B	A
36	C	C	D
37	D	D	A
38	D	E	D
39	C	C	C
40	E	E	E

Item	Form A Randomized Key	Form B Arbitrary Key	Form C Balanced Key
41	E	B	B
42	A	A	D
43	A	A	A
44	C	A	C
45	A	C	C
46	E	E	D
47	C	D	C
48	E	B	E
49	A	A	B
50	C	B	A
51	D	D	D

Note. Item 51 is a bonus item. It was not included in the analysis of data since this item's options were not manipulated via one of the three answer-key strategies examined in this dissertation.

Appendix D

Raw Scores and Grade Point Averages

RANDOMIZED TEST FORMAT

ABILITY			
GPA NOT SHOWN	LOW (2.0-2.7 GPA)	AVERAGE (3.0-3.4 GPA)	HIGH (3.7-4.0 GPA)
SCORE	SCORE	SCORE	SCORE
32	30	32	39
35	27	21	38
32	24	22	31
25	31	39	39
32	28	27	40
35	38	28	35
41	26	27	35
35	21	31	39
33	26	30	37
	25	38	34
	30	21	43
	44	18	40
	24	34	42
	26	33	38
	40	31	44
	36	33	22
	27	21	40
	28	37	37
	31	32	36
	36	28	33
	38	28	9
	28	29	
	34	28	
	33	40	
	27	28	
	32	25	
	22	26	
	25	32	
	28	37	
	23	44	
	31	34	
	16	37	
		34	
		22	
		40	
		35	
		25	
		31	
		33	
		29	
		35	
		20	

	ABILITY		
GPA NOT SHOWN	LOW (2.0-2.7 GPA)	AVERAGE (3.0-3.4 GPA)	HIGH (3.7-4.0 GPA)
SCORE	SCORE	SCORE	SCORE
		31	
		37	
		33	
		26	
		23	
		33	
		26	
		32	
		35	
		32	
		29	
		33	
		37	
		23	
		14	

ARBITRARY TEST FORMAT

	ABILITY		
GPA NOT SHOWN	LOW (2.0-2.7 GPA)	AVERAGE (3.0-3.4 GPA)	HIGH (3.7-4.0 GPA)
SCORE	SCORE	SCORE	SCORE
31	26	26	41
29	37	26	30
41	26	30	34
38	28	37	33
33	28	28	36
20	26	34	36
33	30	38	38
24	31	28	29
37	17	31	34
28	31	28	37
33	23	35	36
24	20	29	36
31	21	28	31
22	31	30	37
24	31	23	36
31	29	28	43
	33	36	43
	22	27	37
	34	33	
	28	31	
	29	28	
	29	32	
	33	27	
	20	31	
	33	31	
	39	28	
		19	
		34	
		22	
		22	
		34	
		34	
		33	
		35	
		33	
		36	
		24	
		26	
		21	
		26	
		32	
		34	

	ABILITY		
GPA NOT SHOWN	LOW (2.0-2.7 GPA)	AVERAGE (3.0-3.4 GPA)	HIGH (3.7-4.0 GPA)
SCORE	SCORE	SCORE	SCORE
		34	
		30	
		30	
		37	
		29	
		33	
		37	
		35	
		37	
		27	
		32	
		38	
		28	
		31	
		37	
		26	

BALANCED TEST FORMAT

ABILITY			
GPA NOT SHOWN	LOW (2.0-2.7 GPA)	AVERAGE (3.0-3.4 GPA)	HIGH (3.7-4.0 GPA)
SCORE	SCORE	SCORE	SCORE
29	28	25	41
35	20	19	33
38	31	31	32
33	30	29	42
36	30	36	33
26	24	14	32
28	28	31	42
22	36	34	35
27	33	31	33
35	43	29	34
35	23	28	36
39	33	26	35
31	29	29	41
32	28	32	43
35	13	27	44
33	19	35	
31	20	29	
	30	30	
	36	33	
	32	33	
	20	28	
	25	33	
	36	33	
	31	30	
	30	26	
	28	31	
	31	31	
	26	28	
		37	
		36	
		31	
		28	
		28	
		26	
		37	
		39	
		34	
		37	
		28	
		34	
		25	
		26	

	ABILITY		
GPA NOT SHOWN	LOW (2.0-2.7 GPA)	AVERAGE (3.0-3.4 GPA)	HIGH (3.7-4.0 GPA)
SCORE	SCORE	SCORE	SCORE
		29	
		25	
		33	
		36	
		35	
		28	

Appendix E

Item P-Values and Variances for all Test Formats

RANDOMIZED TEST FORMAT

	ABILITY							
	NOT SHOWN		LOW		AVERAGE		HIGH	
ITEM	P	S^2	P	S^2	P	S^2	P	S^2
1	.33	.25	.25	.19	.37	.24	.33	.23
2	.89	.11	.94	.06	.91	.08	.90	.09
3	.78	.19	.75	.19	.84	.14	.95	.05
4	.56	.28	.28	.21	.46	.25	.67	.23
5	.33	.25	.31	.22	.26	.20	.38	.25
6	1.00	.00	.88	.11	.88	.11	.95	.05
7	1.00	.00	.84	.14	.81	.16	.90	.10
8	.44	.28	.31	.22	.44	.25	.76	.19
9	.89	.11	.63	.24	.60	.25	.71	.21
10	.89	.11	.81	.16	.88	.11	.86	.13
11	.78	.19	.88	.11	.75	.19	.81	.16
12	.56	.28	.66	.23	.61	.24	.86	.13
13	.33	.25	.41	.25	.46	.25	.62	.25
14	.89	.11	.78	.18	.84	.14	.86	.13
15	.67	.25	.41	.25	.58	.25	.67	.23
16	.44	.28	.41	.25	.32	.22	.29	.21
17	.67	.25	.38	.24	.42	.25	.48	.26
18	.22	.19	.25	.19	.07	.07	.29	.21
19	1.00	.00	.84	.14	.86	.12	.95	.05
20	.11	.11	.50	.26	.39	.24	.52	.26
21	.33	.25	.38	.24	.37	.24	.57	.26
22	1.00	.00	.87	.11	.82	.15	.95	.05
23	.56	.28	.53	.26	.60	.25	.90	.09
24	.67	.25	.59	.25	.56	.25	.62	.25
25	.11	.11	.59	.25	.63	.24	.71	.21
26	1.00	.00	.69	.22	.70	.21	.81	.16
27	.89	.11	.56	.25	.79	.17	.81	.16
28	.67	.25	.44	.25	.40	.25	.38	.25
29	.44	.28	.16	.14	.25	.19	.43	.26
30	.33	.25	.37	.24	.37	.24	.57	.26

ITEM	P	S^2	P	S^2	P	S^2	P	S^2
31	.89	.11	.59	.25	.72	.21	.81	.16
32	.67	.25	.47	.26	.35	.23	.48	.26
33	.56	.28	.44	.25	.47	.25	.71	.21
34	.56	.28	.47	.26	.42	.25	.57	.26
35	.89	.11	.81	.16	.86	.12	.95	.05
36	.89	.11	.81	.16	.75	.19	.71	.21
37	.22	.19	.19	.16	.14	.12	.14	.13
38	.33	.25	.34	.23	.39	.24	.62	.25
39	.78	.19	.81	.16	.88	.11	.90	.09
40	.78	.19	.87	.11	.82	.15	.86	.13
41	.33	.25	.31	.22	.53	.25	.57	.25
42	.56	.28	.44	.25	.49	.25	.71	.21
43	1.00	.00	1.00	.00	.93	.07	1.00	.00
44	.78	.19	.75	.19	.70	.21	.95	.05
45	.89	.11	.66	.23	.68	.22	.81	.16
46	.78	.19	.44	.25	.53	.25	.86	.13
47	1.00	.00	.81	.16	.82	.15	.86	.13
48	1.00	.00	.84	.14	.91	.08	.95	.05
49	.78	.19	.63	.24	.77	.18	.86	.13
50	.89	.11	.84	.14	.75	.19	.90	.09

ARBITRARY TEST FORMAT

	ABILITY							
	NOT SHOWN		LOW		AVERAGE		HIGH	
ITEM	P	S^2	P	S^2	P	S^2	P	S^2
1	.25	.20	.46	.26	.36	.24	.28	.21
2	1.00	.00	.96	.04	.97	.03	.94	.06
3	.87	.12	.81	.16	.81	.16	.78	.18
4	.50	.27	.35	.24	.29	.21	.44	.26
5	.19	.16	.15	.14	.33	.22	.44	.26
6	.94	.06	.85	.14	.98	.02	.94	.06
7	.75	.20	.85	.14	.84	.13	.94	.06
8	.63	.25	.50	.26	.50	.25	.72	.21
9	.81	.16	.73	.21	.76	.19	.67	.24
10	.87	.12	.77	.19	.88	.11	.89	.11
11	.81	.16	.69	.22	.81	.16	.72	.21
12	.69	.23	.62	.25	.48	.25	.83	.15
13	.37	.25	.35	.24	.50	.25	.67	.24
14	.63	.25	.85	.14	.86	.12	.94	.56
15	.81	.16	.77	.19	.79	.17	.94	.06
16	.31	.23	.12	.11	.31	.22	.33	.24
17	.24	.20	.42	.25	.34	.23	.50	.27
18	.25	.20	.15	.14	.19	.16	.17	.15
19	.87	.12	.88	.11	.84	.13	1.00	.00
20	.25	.20	.27	.20	.29	.21	.61	.25
21	.50	.27	.31	.22	.45	.26	.67	.24
22	.81	.16	.65	.24	.90	.09	1.00	.00
23	.81	.16	.73	.21	.67	.22	.94	.06
24	.44	.26	.54	.26	.59	.25	.72	.21
25	.00	.00	.04	.04	.07	.07	.06	.06
26	.44	.26	.77	.19	.67	.22	.78	.18
27	.81	.16	.42	.25	.71	.21	.94	.06
28	.50	.26	.19	.16	.52	.25	.39	.25
29	.44	.26	.35	.24	.26	.20	.44	.26
30	.31	.23	.38	.25	.31	.22	.61	.25

ITEM	P	S^2	P	S^2	P	S^2	P	S^2
31	.63	.25	.69	.22	.74	.20	.61	.25
32	.38	.25	.31	.22	.36	.24	.61	.25
33	.69	.23	.69	.22	.55	.25	.78	.18
34	.69	.23	.46	.26	.76	.19	.67	.24
35	.75	.20	.73	.21	.83	.15	1.00	.00
36	.75	.20	.69	.22	.90	.09	.94	.06
37	.21	.20	.23	.19	.12	.11	.28	.21
38	.19	.16	.42	.25	.21	.17	.61	.25
39	.75	.20	.50	.26	.79	.17	.89	.11
40	1.00	.00	.85	.14	.88	.11	.67	.24
41	.44	.26	.65	.24	.57	.25	.56	.26
42	.63	.25	.62	.25	.50	.25	.83	.15
43	1.00	.00	.96	.04	.95	.05	1.00	.00
44	.75	.20	.65	.24	.72	.20	.94	.05
45	.50	.27	.50	.26	.64	.24	.83	.15
46	.44	.26	.50	.26	.55	.25	.72	.21
47	.69	.23	.58	.25	.66	.23	.83	.15
48	.87	.12	.92	.07	.91	.08	1.00	.00
49	.69	.23	.65	.24	.78	.18	.89	.11
50	.75	.20	.73	.21	.79	.17	.94	.06

BALANCED TEST FORMAT

	ABILITY							
	NOT SHOWN		LOW		AVERAGE		HIGH	
ITEM	P	S^2	P	S^2	P	S^2	P	S^2
1	.61	.25	.39	.25	.27	.20	.33	.24
2	1.00	.00	.93	.07	.90	.10	1.00	.00
3	.56	.26	.71	.21	.73	.20	.87	.12
4	.17	.15	.29	.21	.19	.16	.47	.27
5	.22	.18	.21	.18	.25	.19	.27	.21
6	.89	.11	.89	.10	.92	.08	1.00	.00
7	.83	.15	.86	.13	.83	.14	1.00	.00
8	.67	.24	.50	.26	.79	.17	.93	.07
9	.94	.06	.79	.18	.83	.14	1.00	.00
10	1.00	.00	.79	.18	.88	.11	.87	.12
11	.67	.24	.46	.26	.62	.24	.73	.21
12	.50	.27	.54	.26	.50	.26	.40	.26
13	.78	.18	.79	.18	.85	.13	1.00	.00
14	.72	.21	.68	.23	.87	.11	.87	.12
15	.50	.27	.39	.25	.35	.23	.27	.21
16	.67	.24	.25	.19	.33	.23	.67	.24
17	.11	.11	.14	.13	.21	.17	.33	.24
18	.89	.11	.89	.10	.96	.04	1.00	.00
19	.39	.25	.39	.25	.25	.19	.40	.26
20	.44	.26	.36	.24	.33	.23	.73	.21
21	.67	.24	.79	.18	.92	.08	.93	.07
22	.67	.24	.79	.18	.65	.23	.80	.17
23	.67	.25	.57	.25	.62	.24	.67	.24
24	.61	.25	.43	.25	.69	.22	.73	.22
25	.61	.25	.57	.25	.73	.20	.87	.12
26	.78	.18	.50	.26	.67	.23	.87	.12
27	.33	.24	.39	.24	.27	.20	.40	.26
28	.22	.18	.14	.13	.27	.20	.47	.27
29	.39	.25	.25	.19	.44	.25	.40	.26
30								

ITEM	P	S^2	P	S^2	P	S^2	P	S^2
31	.78	.18	.71	.21	.60	.24	.73	.21
32	.44	.26	.29	.21	.37	.24	.73	.21
33	.56	.26	.57	.25	.56	.25	.73	.21
34	.67	.24	.57	.25	.52	.26	.73	.21
35	.83	.15	.68	.23	.87	.11	1.00	.00
36	.78	.18	.82	.15	.81	.16	.93	.07
37	.33	.24	.21	.18	.23	.18	.20	.17
38	.33	.24	.25	.19	.27	.20	.27	.21
39	.89	.11	.64	.24	.77	.18	.93	.07
40	.94	.06	.79	.18	.90	.10	.87	.12
41	.44	.26	.46	.26	.50	.25	.80	.17
42	.56	.26	.43	.25	.46	.26	.93	.07
43	.94	.06	.93	.07	.96	.04	1.00	.00
44	.72	.21	.61	.25	.60	.24	1.00	.00
45	.67	.24	.32	.23	.56	.25	.87	.12
46	.67	.24	.64	.24	.65	.23	.93	.07
47	.89	.11	.75	.19	.79	.17	.87	.12
48	.94	.05	.82	.15	.81	.16	1.00	.00
49	.89	.11	.79	.18	.81	.16	.93	.07
50	.83	.15	.79	.18	.77	.18	.93	.07

COMBINED ABILITY GROUPS

	RANDOMIZED		ARBITRARY		BALANCED	
ITEM	P	S^2	P	S^2	P	S^2
1	.33	.22	.36	.23	.37	.23
2	.92	.08	.97	.03	.94	.06
3	.83	.14	.81	.15	.72	.21
4	.45	.25	.36	.23	.25	.19
5	.30	.21	.29	.21	.24	.18
6	.90	.09	.94	.06	.92	.08
7	.85	.13	.85	.13	.86	.12
8	.46	.25	.55	.25	.40	.24
9	.65	.23	.75	.19	.72	.21
10	.86	.12	.86	.12	.86	.12
11	.80	.16	.77	.18	.87	.11
12	.66	.23	.59	.25	.61	.24
13	.46	.25	.47	.25	.50	.25
14	.83	.14	.84	.14	.84	.13
15	.55	.25	.81	.15	.80	.16
16	.34	.23	.27	.20	.38	.24
17	.44	.25	.37	.24	.41	.25
18	.17	.14	.19	.15	.19	.16
19	.88	.11	.88	.11	.94	.06
20	.42	.25	.33	.22	.33	.22
21	.40	.24	.46	.25	.41	.24
22	.87	.11	.85	.13	.84	.14
23	.63	.24	.75	.19	.71	.21
24	.59	.24	.58	.25	.62	.24
25	.60	.24	.05	.05	.61	.24
26	.74	.19	.68	.22	.69	.22
27	.74	.19	.69	.21	.67	.22
28	.43	.25	.42	.25	.33	.22
29	.27	.20	.33	.22	.26	.19
30	.40	.24	.37	.24	.38	.24
31	.71	.21	.69	.21	.68	.22

ITEM	P	S^2	P	S^2	P	S^2
32	.43	.25	.39	.24	.41	.25
33	.51	.25	.64	.23	.59	.25
34	.47	.25	.67	.22	.59	.25
35	.87	.12	.82	.15	.83	.14
36	.77	.18	.84	.14	.83	.15
37	.16	.14	.19	.15	.24	.18
38	.41	.24	.31	.22	.28	.20
39	.86	.12	.74	.20	.78	.17
40	.84	.14	.86	.12	.87	.11
41	.46	.25	.57	.25	.52	.25
42	.52	.25	.59	.24	.53	.25
43	.97	.03	.97	.03	.95	.04
44	.76	.18	.75	.19	.68	.22
45	.71	.21	.62	.24	.56	.25
46	.58	.25	.55	.25	.69	.22
47	.84	.14	.67	.22	.81	.16
48	.91	.09	.92	.07	.86	.12
49	.75	.19	.75	.19	.83	.14
50	.82	.15	.80	.16	.81	.16

Appendix F

Point Biserials for All Items, Options, and Test Formats

RANDOMIZED TEST FORMAT

ITEM	CORRELATION	ITEM	CORRELATION
1	-.02	26	.23
2	.31	27	.32
3	.04	28	.15
4	.35	29	.36
5	.17	30	.26
6	.35	31	.28
7	.12	32	.24
8	.38	33	.24
9	.24	34	.35
10	.28	35	.26
11	.25	36	.28
12	.28	37	.21
13	.27	38	.24
14	.23	39	.24
15	.26	40	.12
16	-.05	41	.25
17	.30	42	.25
18	.17	43	.19
19	.27	44	.21
20	-.00	45	.19
21	.32	46	.41
22	.42	47	.28
23	.32	48	.39
24	.33	49	.24
25	.24	50	.47

ARBITRARY TEST FORMAT

ITEM	CORRELATION	ITEM	CORRELATION
1	-.02	26	.04
2	.03	27	.25
3	-.05	28	.04
4	.18	29	.23
5	.33	30	.26
6	.10	31	.00
7	.17	32	.27
8	.08	33	.16
9	.05	34	.23
10	.09	35	.35
11	.08	36	.30
12	.37	37	.19
13	.13	38	.17
14	.11	39	.23
15	.02	40	-.13
16	.12	41	.05
17	.19	42	.38
18	.11	43	.13
19	.09	44	.31
20	.28	45	.30
21	.32	46	.31
22	.18	47	.33
23	.34	48	.17
24	.20	49	.29
25	-.20	50	.30

BALANCED TEST FORMAT

ITEM	CORRELATION	ITEM	CORRELATION
1	-.08	26	.20
2	.32	27	.28
3	.03	28	.11
4	.20	29	.24
5	.05	30	.24
6	.37	31	.14
7	.30	32	.17
8	.14	33	.18
9	.22	34	.27
10	.18	35	.48
11	.16	36	.20
12	.30	37	.04
13	-.01	38	.13
14	.25	39	.28
15	.18	40	.06
16	.05	41	.24
17	.19	42	.36
18	.20	43	.16
19	.28	44	.29
20	.01	45	.28
21	.27	46	.41
22	.05	47	.40
23	.11	48	.41
24	.23	49	.18
25	.13	50	.19

Appendix G

Classification of Point Biserials for all Test Formats

CLASSIFICATION DATA FOR RANDOMIZED TEST FORMAT

POOR (Below .09)	FAIR (Between .10 and .29)	GOOD (Above .30)
1, 3, 16, 20	5, 7, 9, 10, 11, 12, 13, 14, 15, 18, 19, 25, 26, 28, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47, 49	2, 4, 6, 8, 17, 21, 22, 23, 24, 27, 29, 34, 46, 48, 50

CLASSIFICATION DATA FOR ARBITRARY TEST FORMAT

POOR (Below .09)	FAIR (Between .10 and .29)	GOOD (Above .30)
1, 2, 3, 8, 9, 10, 11, 15, 19, 25, 26, 28, 31, 40, 41	4, 6, 7, 13, 14, 16, 17, 18, 20, 22, 24, 27, 29, 30, 32, 33, 34, 37, 38, 39, 43, 48, 49	5, 12, 21, 23, 35, 36, 42, 44, 45, 46, 47, 50

CLASSIFICATION DATA FOR BALANCED TEST FORMAT

POOR (Below .09)	FAIR (Between .10 and .29)	GOOD (Above .30)
1, 3, 5, 13, 16, 20, 22, 37, 40	4, 8, 9, 10, 11, 14, 15, 17, 18, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 36, 38, 39, 41, 43, 44, 45, 49, 50	2, 6, 7, 12, 35, 42, 46, 47, 48

Appendix H

Distractor Analysis

RANDOMIZED TEST FORMAT

ITEM	OPTION A	OPTION B	OPTION C	OPTION D	OPTION E
1	.10	.56	.33*	.00	.02
2	.00	.06	.92*	.01	.02
3	.00	.83*	.06	.10	.01
4	.41	.01	.10	.03	.45*
5	.14	.30*	.05	.01	.50
6	.90*	.00	.01	.07	.03
7	.00	.04	.85*	.02	.09
8	.20	.07	.16	.46*	.11
9	.01	.04	.17	.13	.65*
10	.01	.13	.00	.00	.86*
11	.04	.12	.80*	.03	.01
12	.03	.04	.66*	.18	.08
13	.46*	.25	.06	.17	.06
14	.06	.00	.83*	.08	.03
15	.56*	.00	.01	.40	.03
16	.13	.33	.11	.35*	.09
17	.40	.44*	.17	.00	.00
18	.32	.40	.17*	.12	.00
19	.02	.03	.88*	.01	.06
20	.42*	.08	.02	.01	.48
21	.40*	.08	.01	.51	.00
22	.04	.01	.03	.88*	.04
23	.01	.10	.13	.63*	.13
24	.27	.59*	.06	.06	.03
25	.19	.05	.11	.06	.59*
26	.74*	.01	.13	.02	.11
27	.01	.03	.00	.22	.74*
28	.31	.05	.08	.43*	.13
29	.03	.21	.27*	.03	.46
30	.09	.41*	.04	.35	.10
31	.17	.02	.00	.10	.71*
32	.43*	.29	.19	.05	.03

ITEM	OPTION A	OPTION B	OPTION C	OPTION D	OPTION E
33	.09	.24	.51*	.07	.09
34	.18	.17	.47*	.13	.05
35	.87*	.03	.02	.04	.05
36	.08	.02	.77*	.05	.08
37	.27	.27	.26	.16*	.04
38	.41	.12	.01	.41*	.05
39	.02	.04	.86*	.01	.08
40	.08	.02	.05	.02	.84*
41	.22	.24	.03	.05	.46*
42	.52*	.03	.13	.33	.00
43	.97*	.01	.02	.00	.01
44	.03	.01	.77*	.01	.19
45	.72*	.11	.11	.06	.00
46	.07	.11	.03	.21	.58*
47	.00	.02	.84*	.06	.08
48	.02	.03	.01	.04	.91*
49	.75*	.10	.05	.05	.05
50	.07	.02	.82*	.04	.06

*Correct Answer

ARBITRARY TEST FORMAT

ITEM	OPTION A	OPTION B	OPTION C	OPTION D	OPTION E
1	.12	.53	.36*	.00	.00
2	.01	.00	.01	.97*	.02
3	.13	.81*	.00	.06	.00
4	.48	.00	.16	.01	.36*
5	.29*	.09	.09	.01	.53
6	.03	.00	.94*	.02	.01
7	.03	.03	.08	.02	.85*
8	.09	.14	.15	.55*	.07
9	.75*	.00	.01	.09	.16
10	.12	.86*	.00	.03	.00
11	.10	.77*	.04	.09	.00
12	.03	.13	.59*	.18	.08
13	.03	.27	.20	.48*	.03
14	.04	.02	.06	.04	.84*
15	.01	.81*	.09	.04	.05
16	.24	.35	.06	.09	.27*
17	.53	.38*	.09	.00	.01
18	.19*	.24	.48	.10	.00
19	.05	.02	.88*	.01	.04
20	.10	.06	.08	.33*	.42
21	.02	.46*	.09	.43	.00
22	.85*	.01	.03	.05	.06
23	.01	.14	.05	.06	.75*
24	.01	.29	.04	.58*	.09
25	.12	.01	.10	.72	.05*
26	.68*	.03	.20	.01	.09
27	.03	.02	.03	.23	.70*
28	.31	.07	.42*	.09	.10
29	.02	.03	.25	.33*	.38
30	.13	.04	.37*	.41	.05
31	.23	.00	.00	.07	.70*
32	.23	.07	.18	.14	.39*

ITEM	OPTION A	OPTION B	OPTION C	OPTION D	OPTION E
33	.03	.64*	.15	.05	.14
34	.67*	.09	.11	.09	.05
35	.00	.82*	.05	.03	.10
36	.02	.00	.84*	.10	.04
37	.20	.24	.34	.19*	.03
38	.50	.11	.03	.04	.31*
39	.03	.13	.74*	.03	.08
40	.05	.00	.02	.08	.86*
41	.12	.57*	.01	.28	.03
42	.59*	.02	.17	.21	.01
43	.97*	.02	.02	.00	.00
44	.75*	.02	.03	.00	.21
45	.14	.17	.62*	.06	.01
46	.06	.09	.02	.28	.55*
47	.01	.05	.18	.67*	.09
48	.00	.92*	.01	.00	.07
49	.75*	.10	.11	.03	.01
50	.06	.80*	.01	.07	.07

*Correct Answer

BALANCED TEST FORMAT

ITEM	OPTION A	OPTION B	OPTION C	OPTION D	OPTION E
1	.37*	.54	.06	.00	.03
2	.01	.94*	.04	.01	.01
3	.13	.72*	.03	.13	.00
4	.48	.00	.26	.25*	.02
5	.16	.24*	.09	.00	.51
6	.02	.01	.02	.92*	.04
7	.00	.05	.86*	.03	.06
8	.18	.09	.14	.18	.40*
9	.02	.02	.10	.72*	.15
10	.11	.86*	.01	.01	.01
11	.87*	.09	.00	.04	.00
12	.61*	.05	.09	.21	.05
13	.50*	.28	.16	.03	.04
14	.07	.01	.84*	.03	.05
15	.01	.04	.80*	.12	.04
16	.14	.38	.03	.38*	.08
17	.00	.41	.17	.00	.41*
18	.25	.45	.19*	.09	.02
19	.03	.94*	.02	.00	.02
20	.07	.09	.33*	.05	.46
21	.03	.08	.00	.48	.41*
22	.06	.84*	.02	.00	.07
23	.01	.11	.11	.06	.71*
24	.62*	.26	.05	.04	.04
25	.17	.04	.13	.06	.62*
26	.69*	.03	.16	.02	.11
27	.05	.04	.02	.23	.67*
28	.43	.33*	.05	.07	.12
29	.00	.36	.26*	.04	.35
30	.38*	.03	.17	.37	.06
31	.26	.01	.00	.06	.68*
32	.32	.41*	.17	.07	.03

ITEM	OPTION A	OPTION B	OPTION C	OPTION D	OPTION E
33	.06	.06	.20	.59*	.10
34	.04	.10	.16	.11	.59*
35	.84*	.01	.01	.05	.09
36	.05	.05	.02	.83*	.06
37	.24*	.33	.25	.17	.02
38	.53	.14	.01	.28*	.05
39	.03	.13	.78*	.04	.03
40	.03	.01	.05	.05	.87*
41	.20	.53*	.01	.23	.04
42	.02	.26	.17	.53*	.02
43	.95*	.03	.00	.00	.02
44	.06	.00	.68*	.00	.27
45	.18	.15	.56*	.06	.06
46	.13	.04	.02	.69*	.13
47	.02	.02	.81*	.08	.07
48	.10	.01	.03	.00	.86*
49	.04	.84*	.04	.05	.05
50	.81*	.01	.03	.06	.10

*Correct Answer