8-9-2008

College Students' Behavior on Multiple Choice Self-Tailored Exams in Relation to Metacognitive Ability, Self-efficacy, and Test Anxiety

Jasna Vuk

COLLEGE STUDENTS' BEHAVIOR ON MULTIPLE CHOICE SELF-TAILORED EXAMS IN RELATION TO METACOGNITIVE ABILITY, SELF-EFFICACY, AND TEST ANXIETY

By

Jasna Vuk

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Educational Psychology
in the Department of Counseling and
Educational Psychology

Mississippi State, Mississippi

August 2008

COLLEGE STUDENTS BEHAVIOR ON MULTIPLE CHOICE SELF-TAILORED EXAMS IN RELATION TO METACOGNITIVE ABILITY, SELF-EFFICACY, AND TEST ANXIETY

By

Jasna Vuk

Approved:

David T. Morse Professor of Educational Psychology (Director of Dissertation)	Anastasia D. Elder Assistant Professor of Educational Psychology (Committee Member)
Linda W. Morse Professor of Educational Psychology (Committee Member)	Jianzhong Xu Professor of Curriculum & Instruction (Committee Member)
John S. Young Professor of Counselor Education (Committee Member)	Glen R. Hendren Graduate Program Coordinator for the Department of Counseling and Educational Psychology
Richard Blackbourn Dean of the College of Education	

Name: Jasna Vuk

Date of Degree: August 9, 2008

Institution: Mississippi State University

Major Field: Educational Psychology

Major Professor: Dr. David T. Morse

Title of Study: COLLEGE STUDENTS' BEHAVIOR ON MULTIPLE CHOICE SELF-TAILORED EXAMS IN RELATION TO METACOGNITIVE ABILITY, SELF-EFFICACY, AND TEST ANXIETY

Pages in Study: 142

Candidate for Degree of Doctor of Philosophy

The purpose of this study was to observe college students' behavior on five self-tailored, multiple choice exams throughout a semester in relation to: a) metacognitive ability, b) self-efficacy expectations, and c) test anxiety. Additionally, the effect of a self-tailoring procedure on exam scores and content validity of the tests was observed. Self-tailored testing was defined as an option in which students selected up to five questions they wanted to omit from being scored on an exam. Students' metacognitive ability was defined as the percentage of incorrectly answered questions out of the total number omitted.

Ninety-nine college students from two sections of an educational psychology undergraduate course participated in this study. Eighty students completed the study; seventy-one used an option to omit questions on all exams. Before taking exam 1, students answered measures of self-efficacy and test anxiety. After completing each of the five course exams, students marked on the back of their answer sheet up to five

questions they wanted to be omitted from scoring. After exam 5, students answered a questionnaire that addressed their perception of the self-tailoring procedure.

MANOVA, repeated measures ANOVA, Pearson correlations, *t*-test and one-way ANOVA were conducted. Students made a statistically significant increase in their scores on all exams by using the questions omitting procedure. There was a statistically significant linear increase of percentages of incorrectly answered questions out of the total number omitted across five exams. Frequency of items that students omitted from scoring were significantly negatively correlated with item difficulty values. The content validity of the test was affected on two out of five exams based on cognitive level of items and on three out of five exams based on chapter coverage. Students' self-efficacy expectations and test anxiety were not related to the likelihood to apply the self-tailoring procedure or to the degree of success students had in applying the procedure. The study provided a new perspective on self-tailored tests in college classroom with implications for teaching, assessment, and students' metacognitive abilities.

DEDICATION

I dedicate this research to my parents who taught me from an early age the value of education, and who have always believed in me. Furthermore, I also dedicate this research to my two wonderful sons Velimir and Vedran in hope of inspiring them to pursue their own goals and accomplishments in the future.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to people without whose assistance this dissertation would not have been possible. First of all, I would like to thank my major professor and dissertation director Dr. David T. Morse for his support and guidance throughout the doctoral program and dissertation process. I appreciate his knowledge, time, and effort, and I will always be thankful to him for encouraging me and selflessly assisting me to reach this accomplishment. Further appreciation is due to the other members of my dissertation committee, Dr. Anastasia D. Elder, Dr. Linda W. Morse, Dr. J. Scott Young, and Dr. Jianzhong Xu for providing me with their feedback and directions. I am thankful to Dr. Anastasia D. Elder and Dr. L. Morse for allowing me to collect data in their classes. Finally, I would like to thank my family, friends, and my husband Stanko for their understanding and encouragement during my doctoral program and dissertation process.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Self-tailored testing is not an ordinary method of assessment in college classrooms. Multiple choice exams are perhaps the most common assessment tool of college students' academic performance but are administered typically without any option to choose questions from the larger pool of questions or to make any adaptation of the test. Many times students complain that the test was either too hard, or it did not measure their real knowledge. An option to choose from a larger pool of questions is usually given to students only on an essay type exam. The current study allowed college undergraduate students to partially tailor their multiple choice exams by omitting a limited number of questions of their choice from the scoring on exams. After answering all questions on an exam, students omitted up to five questions that they believed they answered incorrectly or were uncertain as to the correctness of their response. In relation to this procedure, several types of outcomes were observed: a) students' metacognitive ability to distinguish between correct and incorrect answers when given an option to omit questions from scoring, b) the effect of a question omitting procedure on exam scores throughout the semester, c) impact on test content validity by examining difficulty and cognitive complexity of omitted questions, and d) self-efficacy and test anxiety as related to students' success in using an option to omit questions from scoring.

Theoretical frameworks necessary for understanding students' behavior in these areas such as self-tailored tests, metacognitive ability, item difficulty, self-efficacy, and test-anxiety will be explained in the next chapter. The remainder of this chapter briefly frames the background for the study and concludes with the purpose of the current study and the research questions of the study.

Self-tailored Tests

The terms "adaptive test" and "tailored test" have been used interchangeably in the literature by several authors (Crocker & Algina, 1986; Thorndike, 1982; Weiss, 1982; Weiss, 1983; Wright & Stone, 1979). Weiss (1983) described adaptive or tailored testing as a procedure which involves selection of test items during test administration that are appropriate in difficulty level for each examinee. The test results are then "'adapted' or 'tailored' to each individual's ability…items are selected out of larger pool using a set of rules, or 'strategies' that may operate differently" (p. 5). Similar ideas about tailored tests were explained by Thorndike (1982) and Crocker and Algina (1986).

Advancements in technology, the economics of testing and more positive attitude of examinees toward computers affected development of computerized adaptive testing (Gitomer, 2000). Despite all these, schools often do not have sufficient computers to implement a computerized adaptive type of testing to students in classrooms (Morse, 1988). Thissen and Mislevy (2000) reported that non-computerized adaptive tests could be designed on the example of "a self-scoring flexilevel test" described by Lord in 1971 (p. 102). Lord's general idea of flexilevel tests was that the examinee starts a test by answering a moderately difficult item, and then attempts an easier item if the answer is

wrong or a harder item if the answer is correct. This type of test included a modification of test directions, answer sheets, and scoring method. Because of its complex instruction, substantial effort on the part of an examinee and an additional effort to construct the test, this test design has been used rarely in practical settings.

Common for all adaptive tests is that an examinee moves to the next question, and the difficulty level of the next question is based on the outcome of previous question. Wood (1974) used the term response-contingent testing for the different names that were being given to this type of the testing. In ordinary classrooms, adaptive testing (computerized or paper-pencil) is difficult to apply. An alternative procedure that can be used in classroom is self-tailored test (Morse, 1988).

Morse (1988) investigated a form of self-tailored test (STT) in college classroom in which an examinee selected items from a larger set of questions with the only constraint of some minimum number of items that an examinee needed to attempt. After completing a test, 190 students in five undergraduate classes and one graduate college class were asked to review the test and mark on a separate sheet of paper those questions which would best show how well they learned material covered on the test. Students were asked to choose a minimum of five questions and no more than 50 % of the questions. When scores were calculated for the student-selected questions, only 21 of 190 students failed to improve their score by using the described method. The degree to which examinees increased their score by using the STT procedure appeared to be a function of the difficulty of item pool. On more difficult exams there was a larger difference between the average difficulty index of items selected by examinees and items not selected.

Examinees performed better on items they selected than on those they did not select. The results of the self-tailoring procedure on the test showed students' ability to distinguish items for which they knew the answers from those on which they were not certain about their answers.

According to Hunt and Hassmen (1997, p. 5) the question is, "where is the boundary between being certain enough and not certain enough." Certainty about knowledge might be important in practical application of knowledge. For practical performance, a person must be certain about his or her knowledge and to bear the consequences of his or her actions. In a testing situation, an examinee might choose the correct answer but still be unsure about its correctness. Test results do not differentiate between the examinee that is not sure but still correct about his or her answer and the examinee that is sure and correct. As a solution to this problem, Hunt and Hassmen offered Self Assessment Computer Analyzed Testing (SACAT) in which the level of certainty could be indicated for each answer by the examinee. An index of accuracy was calculated for each examinee. In addition to certainty about knowledge, a metacognitive ability to distinguish what one knows from what he or she does not know is equally important to understand.

Metacognition and Metacognitive Monitoring

"Knowing what one knows" (Sinkavich, 1995, p. 1) is the question that has been investigated within the field of metacognition since the early 1970s (McCormick, 2003). Flavell (1976) offered one of the first definitions of metacognition as the knowledge of

one's own cognition. Jacobs and Paris (1987) referred to metacognition as "thinking about thinking" (p. 255).

Metacognition in general can be separated into two main areas: knowledge of cognition and regulation of cognition (McCormick; Schraw 1998; Schraw & Dennison, 1994; Schraw & Graham, 1997; Schraw & Moshman, 1995). Knowledge about cognition involves three different types of metacognitive awareness: a) declarative-"knowing 'about' things", b) procedural-"knowing 'how' to do things", and c) conditional-"knowing the 'why' and 'when' aspects of cognition" (Schraw & Moshman, p. 352). Regulation of cognition involves activities that control and regulate a person's thinking and learning which includes planning, monitoring, and evaluation (Schraw & Moshman).

The areas of metacognition that were of greater interest in the current study were in regulation of cognition particularly in monitoring, evaluation of performance, and retroactive assessment. Several previous researchers used the term "monitoring accuracy" when referring to the process of matching perceived and actual performance. Monitoring accuracy was also labeled "calibration of performance" (e.g., Nietfeld & Schraw, 2002; Nietfeld, Cao, & Osborne, 2005).

Many researchers have examined college students' monitoring ability by comparing their perceived performance with actual performance on test items (Nietfeld & Schraw, 2002; Schraw, 1997; Schraw, Potenza, & Nebelsick-Gullet, 1993; Shaughnessy, 1979). Fewer studies have examined students' ability to discriminate between the answers they knew from those they did not know after taking the test. According to Lundeberg, Fox, Brown, and Elbedour (2000, p. 153) the ability of people "to

discriminate between what they know and what they do not know" might be more important in education than the ability to calibrate a confidence judgment. Still, there is much more research on confidence calibration than on discrimination (Lundeberg et al.).

Pressley and Ghatala (1988) examined college students' ability to discriminate between their correct and incorrect answers on multiple choice tests. Students answered three types of verbal section items of the Scholastic Aptitude Test: 15 items on reading comprehension, 10 items on opposites, and 10 items on analogies. They rated their certainty about the correctness of each answer on an Awareness Scale, supplied by the researchers that consisted of labels ranging from 20 % to 100% certainty. These labels included "just a guess" for 20% and "absolutely certain" for 100% (p. 458). According to Pressley and Ghatala, the most striking finding of their study was that students were often very certain their answers on comprehension items were right when in fact they were wrong. Another important finding was that students discriminated better between their correct and incorrect answers on easier questions as indicated by the item difficulty index values. However, students' awareness of performance was not related to their academic ability as indicated by their overall score on the verbal part of the SAT. Other studies have reported conflicting results concerning this relationship.

The Relationship between Test Performance and Metacognitive Monitoring

Sinkavich (1995) found that students who achieved higher scores on their exams were able to predict better what they knew from what they did not know. This finding led the researcher to conclude that students who scored higher on their exams had better

metamemory accuracy than poor students. A different finding was reported on final exam. On final examination, students were allowed to use 10 replacement items. Low-scoring students actually gained more percentage points by replacing the items on the final exam than did good students. This finding was explained by the fact that low-scoring students had more incorrect answers to replace.

Kruger and Dunning (1999) reported a similar ability of college undergraduate students to estimate their performance in relation to their ranking of performance on humor, logic, and grammar tests. They found that less competent individuals who scored in the bottom quartile overestimated their performance on these tests from different domains. In contrast to the students from the lowest quartile, students from the highest quartile underestimated their ability, but they were more accurate than the lowest performing students. According to Kruger and Dunning, less competent individuals lack metacognitive skills for an accurate estimation of their performance. Also, less competent students lack an ability to improve their self-assessment after observing the test results of their peers.

Another factor that might influence students' judgment of their performance is the effect of practice on successive exams (Hacker, Bol, Horgan, & Rakow, 2000; Nietfeld et al., 2005; Pierce & Smith, 2001); however, mixed results have been reported. Pierce and Smith indicated no improvement in metacomprehension accuracy on successive exams. Nietfeld et al. reported that monitoring accuracy remained stable over four exams in semester for all students; whereas, Hacker et al. found that high scoring students improved their prediction of performance over three exams in a semester.

Item difficulty and test difficulty have been reported by numerous studies as the important factors that affect students' judgment of performance (Hacker et al., 2000; Lichtenstein & Fischhoff, 1977; Morse, 1988: Nietfeld et al., 2005; Pressley & Ghatala, 1988; Schraw & Roedel, 1994; Sinkavich, 1995). However, judgment of performance might be affected by other factors. One of these factors, defined by social cognitive theory, is self-efficacy which plays an important role in self-regulation and monitoring of performance.

Self-efficacy, Self-regulation, and Monitoring of Performance

According to Bandura (1989), self-efficacy relates to "people's beliefs about their capabilities to exercise control over events that affect their lives" (p. 1175). These beliefs influence people's thoughts, feelings, motivation, and behavior. People construct scenarios of their actions based on their sense of self-efficacy. Those who have high self-efficacy visualize their success which guides their performance. Those who have low self-efficacy have difficulty in succeeding because they dwell on their failures from the past (Bandura, 1993).

A study by Nietfeld and Schraw (2002) found that self-efficacy scores were related to performance on the Raven Advanced Progressive Matrices, though not to the degree of accuracy in estimating success on items measuring knowledge of probability. Pajares (1996a) reported that self-efficacy beliefs influence people's choices, actions, thought patterns, emotional reactions, and motivation. Expectancy beliefs about "one's perceived competence" (p. 544) have been investigated in the area of education. Therefore, self-efficacy beliefs were identified as a relevant variable to be examined in

the current study. Self-efficacy beliefs were examined in relations to students' likelihood to apply and succeed in using a self-tailoring procedure.

Test anxiety is a factor that is related to self-efficacy, self-regulation, and to metacognitive monitoring but has not been researched in the context of self-tailored exams. Test anxiety in the current study was examined in relations to students' likelihood to apply an option to omit questions from scoring and to the degree of success students have in applying the optional omission procedure.

Test Anxiety in Relation to Self-efficacy and Metacognition

Test anxiety has been reported as a factor that negatively affects academic achievement of college students (Chapell et al., 2005; Culler & Holahan, 1980; Benjamin, McKeachie, & Lin, 1987; Benjamin, McKeachie, Lin, & Holinger, 1981; Pintrich, Smith, Garcia, & McKeachie, 1991). According to Pintrich et al., test anxiety has a cognitive and an emotionality component. A cognitive component refers to worrying, negative thinking, and preoccupation with performance. An emotionality component includes physiological and affective responses to anxiety. Test anxiety has been investigated in relation to students': a) self-efficacy (Bandura, 1989; Bandura, 1993), b) metacognitive skills (Everson, Smodlaka, & Tobias, 1994; Veenman, Kerseboom, & Imthorn, 2000), c) an information processing model (Benjamin, McKeachie, & Lin, 1987), and d) self-regulation (Pintrich & De Groot, 1990). Therefore, test anxiety was identified as a relevant variable to be examined in the current study.

Summary

There are several gaps in research in the areas previously mentioned.

First, self-tailored exams have not been well investigated in research. There is no study that has investigated examinees' behavior on self-tailored exams in the same manner as the current study.

Second, previous studies used different methods, materials and designs and reported inconsistent results which are difficult to compare.

Third, studies of whether practice on subsequent exams would improve accuracy of metacognitive monitoring on a test have yielded mixed results; therefore, this question needs to be investigated further. There are no other studies that have investigated the effect of practice over five examinations in a semester.

Fourth, none of the previous studies have examined the relation of metacognition in the context of a self-tailoring procedure on examinations. The topics of monitoring and regulation of test taking have received mixed presentations in literature.

Fifth, test anxiety, self-efficacy, and metacognition are not fully understood as to how they relate to one another, and they have not been investigated in the context of self-tailored exams.

Definitions of the Main Constructs

Self-tailored Tests

Self-tailored tests are tests in which an examinee exerts some control over the item he or she attempts. For the current study, self-tailored testing was operationally

defined as an option in which students may select up to five items they want to omit from being scored on an exam.

Metacognitive Ability

Metacognition or "knowing what one knows" (Sinkavich, 1995, p.1) is an underlying ability of an examinee on a self-tailored testing procedure. The early definition of Flavell (1976) that metacognition is a knowledge about one's own cognition has been commonly accepted by many researchers (McCormick, 2003). Metacognitive ability was operationally defined in the current study as the percentage of incorrectly answered questions out of the total number omitted.

Item Difficulty

Crocker and Algina (1986) defined item difficulty as "the proportion of examinees who answer the item correctly" (p. 311). In the the current study, item difficulty was operationally defined as the difficulty value reported by the Mississippi State University (MSU) Testing Services which is the same as the Crocker and Algina definition.

Self-efficacy

According to Bandura (1989) self-efficacy relates to "people's beliefs about their capabilities to exercise control over events that affect their lives" (p. 1175). In the current study, self-efficacy expectations were operationally defined as the total score on the Expectancy Component: Self-Efficacy for Learning and Performance subscale of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, et al., 1991).

Test Anxiety

Test anxiety is an anxiety that a test-taker experiences during a test that interferes negatively with performance on the test. According to Pintrich et al. (1991) test anxiety has a cognitive or worry and emotionality component. In the current study, test anxiety was operationally defined as the total score on the Affective Component: Test Anxiety subscale of the MSLQ (Pintrich et al.).

Purpose of the Study

The purpose of the current study was: a) to observe students' behavior on five self-tailored multiple choice exams in a college classroom, b) to observe the effect of a question omitting procedure on exam scores throughout the semester, and c) to observe the effect of exam self-tailoring on the content validity of the test. Students' behavior was observed in terms of: a) students' metacognitive ability to distinguish between correct and incorrect answers across five exams, b) students' strategies on self-tailored exams, c) students' self-efficacy expectations, and d) students' test anxiety.

Research Questions

The study addressed the following research questions:

1. How will an option to omit questions from scoring affect students' scores on five exams?
2. Do students improve their ability to distinguish between their correct and incorrect answers over five consecutive exams by using an option to omit

questions from scoring?

3. Will the frequency of items that students omit from scoring be correlated with item difficulty values?
4. How is an option to omit questions from scoring going to affect content validity of the test?
5. How do students describe their strategies of omitting questions from scoring?
6. Will students' self-efficacy expectations be related to their likelihood to apply an option to omit questions from scoring and to the degree of success students have in applying the omission procedure?
7. Will test anxiety be related to the students' likelihood to omit questions from scoring and their success in applying the omission procedure?

CHAPTER II

REVIEW OF THE LITERATURE

The chapter reviews the literature on self-tailored tests, metacognition and metacognitive monitoring as important aspects of a question omitting procedure. To provide a better understanding of monitoring accuracy measurement in current research the two terms, calibration and discrimination, are defined. The chapter continues with a description of factors that affect monitoring on a test such as item and test difficulty, format and timing of a test, practice on multiple tests, achievement on a test, examinee characteristics, and environmental factors. It concludes with a basic background of self-efficacy as related to self-regulation, use of strategies, judgment of performance, and test anxiety.

Self-tailored Tests

Weiss (1982) reported that the first adaptive or tailored test was Binet's intelligence test. The length of the test administered to each individual varied. The difficulty level of questions on that test was adapted (tailored) to ability of each examinee. Wood (1974) also recognized that the principle of tailoring the test to an examinee's ability goes back to the Binet intelligence test. According to Weiss, the Binet test was administered and adapted to individual's ability level by a trained psychologist.

A characteristic of self-tailored tests, as described by Wright and Stone (1979) is that the decision where to start on the test and when to end the test is under control of the examinee.

Wright and Stone (1979) described self-tailoring as an individualized method by which an examinee is presented with increasing level of test item difficulty. Each examinee begins with answering the items which he or she finds "hard enough to interest them but easy enough to master" (p. 151). Answering continues until items become too hard for an examinee to answer. Consequently, performance on the test is measured on the content of a test segment determined by each individual.

Thorndike (1982) used the term "tailor-made testing" to mean the same as adaptive testing when describing an approach of "adjustment of task difficulty to the individual's estimated ability level" (p. 290). Thorndike also described a self-selecting procedure, based on the assumption that an examinee could estimate the difficulty of test items. Selection of items of appropriate level of difficulty would provide the maximum amount of information about examinee's ability. As cited by Thorndike, Prestwood and Weiss in 1977 provided evidence that examinees might choose too easy items which might not be psychometrically efficient indicators of their ability. The choice of test items might also depend on the examinee's level of test anxiety. Individuals with high test anxiety might choose to answer only easy questions.

According to Crocker and Algina (1986) the purpose of adaptive or tailored testing is to match items to an examinee's ability. For example, many times low ability examinees will guess on hard items of standardized tests which will only contribute to

measurement error without providing useful information about examinees' ability. According to Morse (1988, p. 2) "early attempts at developing paper-and-pencil adaptive tests, such as Lord's (1971) flexilevel tests, seem to have been abandoned in favor of computer-administered testing."

According to Thissen and Mislevy (2000, p. 105) "item selection algorithm" in flexilevel tests requires an additional effort from an examinee. Some of examinees might have difficulty with instructions on flexilevel tests which can affect an accuracy of scoring. In computerized tests "a machine does the work instead of the examinee" (Thissen & Mislevy, p. 105). However, computerized adaptive tests have little chance to be applied in ordinary classrooms. An alternative procedure that can be applied in college classroom is a self-tailored procedure (Morse, 1988).

Bedard (1974) and Morse (1988) used self-tailored procedures in college classrooms somewhat differently than described by Wright and Stone (1979). Bedard examined partly tailored exams in grade 12 Algebra and college I Introductory Calculus courses. As opposite to conventional classroom testing, in which all students were examined on the same items and with the test of the same length, partly tailored exams were split in two parts. One part consisted of items that the teacher wanted all of his students to answer; another part consisted of items that students were able to choose whether to answer them or not. The assumption was that students would perform better on exam if they had opportunity to choose those questions on which they believed know the answers. Thirty five girls participated in two experiments. In one study that tested students on algebra, the experimental group had 25 compulsory questions and 10

additional questions which they could choose to answer or not. The control group had the same exam, but all the questions were compulsory. Each exam had two scores: one for the first 25 questions and the other for all the questions answered. The experimental group that could freely answer additional questions had a higher overall mean score than the control group on both a midterm and final exam. In the second study, the significant difference was found between two groups only on the final exam of introductory calculus. Overall, the results supported the hypothesis that students could improve their scores if they had opportunity to choose additional questions.

Morse (1988) used a form of a self-tailoring testing in one graduate and four undergraduate college classrooms. After completing a test, students could choose those questions which would best show how well they learned material covered on the test. Students could choose a minimum of five questions and no more than 50% of the questions. Only 21 from the total number of 190 students who participated in the study failed to improve their score by using this form of a self-tailoring testing procedure.

Since these initial attempts of using partly tailored exams in classroom settings (Bedard, 1974; Morse 1988), the benefits of a self-tailoring testing procedure have not been investigated. Most likely because of the additional efforts to construct and score paper-pencil self-tailored tests, these tests have been abandoned in a favor of computerized adaptive testing (CAT). However, in computerized adaptive testing a computer algorithm selects items from an item bank and provides the individualized or tailored test for each examinee (Lunz & Bergstrom, 1995). Usually, CAT begins with an item of moderate difficulty. Based on the response to that item, an examinee's

proficiency is estimated and the next item is selected by computer (Pitkin & Vispoel, 2001). If an examinee correctly answers more difficult items, the next items presented will be more difficult. Therefore, the presented set of items will be matched to the examinee's ability by a computer algorithm (Lunz & Bergstrom; Wise, Plake, Johnson, & Roos, 1992).

A different form of computerized adaptive testing (CAT) is self-adapted testing (SAT). The main difference between the two forms of testing is that in SAT the examinee (not the computer algorithm) is allowed to choose the difficulty level of each item. Items on SAT are ranged by a difficulty level and grouped in categories ranging from easy to difficult. The examinee chooses the category, answers the item, receives feedback and chooses the next level of difficulty (Pitkin & Vispoel, 2001). Therefore, SAT gives a control over the test items to the examinee and decreases test anxiety (Wise, 1994).

The underlying examinee ability that is responsible for the process of exam self-tailoring is metacognitive ability. It was not until the middle 1970s, that metacognition received much attention from researchers.

Metacognition

Jacobs and Paris (1987) referred to metacognition as to "thinking about thinking" (p. 255). Flavell (1976) defined metacognition as "knowledge concerning one's own cognitive processes and products or anything related to them" (p. 232). Flavell's definition of metacognition in 1976 has been commonly accepted by many researchers in the field (McCormick, 2003). Schraw and Dennison (1994) defined metacognition as "the

ability to reflect upon, understand, and control one's learning" (p. 460). Thompson (1999) reported that metacognition is knowledge about cognitive processes and processes that control them.

Flavell (1979) perceived metacognition as the interaction of many different factors that fall under four categories: "a) metacognitive knowledge, b) metacognitive experiences, c) goals (or tasks), and d) actions (or strategies)" (p. 906). Flavell emphasized the importance of interactions of all these elements. According to Flavell, metacognitive knowledge includes knowledge about a person, task, and strategy, and it is stored in long term memory. It can be retrieved consciously or automatically depending on the task situation, and it can influence the outcome of cognitive processing without even entering consciousness.

According to Flavell (1979) metacognitive experiences, which might be cognitive or affective, accompany intellectual processes. They usually occur in situations that require conscious thinking such as school work and novel situations. Metacognitive experiences can activate both cognitive and metacognitive strategies. For example, a student has a sense that he does not know the chapter well to pass the exam and decides to read the chapter again or asks himself questions about it. By using different strategies a student tries to achieve a goal. Flavell distinguished between cognitive strategies that a student, for example, uses to study for an exam and metacognitive strategies that monitor progress toward a cognitive goal. In approaching a goal or task, a student is using interactively metacognitive knowledge, experiences, and strategies to achieve that goal.

Even though most of researchers in the field of metacognition agree on basic definitions, there are differences in explanations how metacognition works (Jacobs & Paris, 1987). For example, Flavell (1979) argued that metacognitive knowledge can influence cognitive process without entering consciousness. He also perceived affect as a part of metacognitive experiences that accompany regulation of intellectual processes. Ten years later, Jacobs and Paris reported that affect and automatic skill cannot be attributed to metacognition. They defined metacognition as "conscious awareness about cognitive aspects of thinking" (p. 258).

In the period of last ten years, two main areas of metacognition have been reported in the literature: a) knowledge of cognition, and b) regulation of cognition (McCormick, 2003; Schraw 1998; Schraw & Dennison, 1994; Schraw & Graham, 1997; Schraw & Moshman, 1995). Knowledge about cognition includes three types of knowledge: a) declarative, b) procedural, and c) conditional **(**Schraw; Schraw & Dennison). Declarative knowledge includes, for example, knowing about one's own memory and about other characteristics of oneself as a learner. Adults and good learners appear to know more about their cognitions than children and poor learners. Procedural knowledge relates to knowledge about strategies and their efficient use. For example, students with better procedural knowledge will know more strategies, used them more automatically and more effectively. Conditional knowledge refers to when and why to use different strategies. For example, a good learner will know which part of the material will need more rehearsal than the other and when (Schraw).

Regulation of cognition that helps students to control their learning includes three types of skills: a) planning, b) monitoring, and c) evaluation. Planning involves the selection of strategies and resources such as time and attention before beginning a task. Monitoring refers to self-testing while performing a task. Evaluation relates to appraisal of one's learning and might include re-evaluation of one's goals. All these regulatory skills are not isolated from knowledge about cognition, and they might influence each other. For example, better planning can improve other parts of regulation of cognition (Jacobs & Paris, 1987; Schraw, 1998; Schraw & Moshman, 1995).

Because of its role in performance on a task (in the current study on multiple choice exams), metacognitive monitoring will be described in more detail.

Metacognitive Monitoring

As defined by McCormick (2003) "monitoring includes identifying the task, checking the progress of task completion, and predicting the eventual outcome" (p. 80). Two related parts of monitoring are comprehension and performance (Pressley & Ghatala, 1990; Schraw, 1998). Nietfeld et al. (2005) referred to monitoring as an awareness of comprehending the task while in the process of performing the task.

Pressley and Ghatala (1990) described monitoring as an executive process that activates and de-activates other processes of thinking and is in the center of self-regulated thinking. It includes an evaluation of whether cognitive actions that are employed will result in reaching a learning goal or not. If a progress toward the goal has not been made, certain actions will be revised. Monitoring is considered to be critical for giving information to students as to both how well the material was learned and for making

decisions about new study strategies. Pintrich (2004) reported that monitoring provides a student with information about any discrepancy between a goal and progress made toward that goal. Students with good monitoring skills will reflect on their progress and change learning strategies if necessary. According to Pressley, Snyder, Levin, Murray, and Ghatala (1987) "monitoring is presumed to provide information about ongoing processing – information that is a form of metacognition – and in turn, this metacognition orchestrates subsequent reading" (p. 221). If a reader realizes that he or she does not comprehend or remember the material that needs to be learned, reading can be readjusted. A student can at least reread the text, read it more carefully or apply completely new strategies. Metacognitive monitoring might provide information to a student as to whether he or she is ready for a test or not (Pressley et al., 1987).

Metacognitive Monitoring of Performance on a Test

During performance on a test, students use regulation of metacognition called metacognitive monitoring as an executive process of cognitive actions that are taken (Pressley & Ghatala, 1990). Monitoring accuracy alerts students how to regulate their learning and performance. Several studies referred to monitoring accuracy as the process of matching between perceived and actual performance (Bol, Hacker, O'Shea, & Allen, 2005; Nietfeld & Schraw, 2002; Nietfeld et al., 2005). Monitoring accuracy was also called calibration of performance (Nietfeld & Schraw; Nietfeld et al.; Schraw, Potenza, & Nebelsick-Gullet, 1993). An important area for the current study is monitoring of performance on a multiple choice test, especially differentiation between correct and incorrect answers when taking the test. According to Kelemen, Frost, and Weaver (2000)

there is no one reliable and comparable index of monitoring accuracy established in current literature. To provide a background for better understanding of terminology in current research on measurement of metacognitive monitoring accuracy, two of the most frequently used terms, calibration and discrimination, will be discussed.

Calibration

Nietfeld et al. (2005) defined calibration as "the process of matching perception of performance with actual level of performance "(p. 10). Nietfeld et al. distinguished between calibration of comprehension and calibration of performance which are both estimates of monitoring accuracy. In calibration of comprehension, an individual provides a confidence judgment before answering a test question. In calibration of performance, an individual provides a confidence judgment after answering a question. Calibration of performance could be estimated locally on each item of the test and globally for the whole test. As the results of these estimates, an accuracy index of local monitoring and global monitoring could be calculated. Nietfeld et al. examined global and local monitoring accuracy across four multiple choice tests in educational psychology undergraduate course. They found that accuracy of confidence judgment was moderated by difficulty of test items.

According to Schraw et al. (1993) calibration of performance is measured by comparing perceived performance with real performance on test items. When perceived judgment of performance is close to the real performance, a learner is well calibrated. A big discrepancy between perceived judgment of performance and true performance means that a learner is poorly calibrated.

Schraw (1997) calculated bias score of confidence judgments of ninety-five undergraduate students on four tests: a) a lexical comparison test, b) the Nelson-Denny reading comprehension test, c) a syllogistic reasoning test, and d) a mathematics test that required computation of simple probabilities. The bias score was calculated as the difference between the mean confidence judgment for each test on a scale of 1-100 and the mean performance expressed as a percent of correct answers. The bias score ranged from -99 to 99. Positive scores expressed an overconfidence judgment and negative scores expressed an under confidence judgment for each test.

Discrimination

Another measure of metacognitive monitoring accuracy that has been reported in the literature is called discrimination or resolution. This measure refers to how well a person can discriminate correct from incorrect answers (Thompson, 1999). Yaniv, Yates, and Smith (1991) reported that the calibration and discrimination skill and assessment of these skills can be applied in different areas such as an investment, predicting the outcome of a game, forecasting, etc. However, good calibration ability does not necessarily correspond to good discrimination ability. For example, two probability judges can be equally well calibrated, but the better judge is the one who can identify the instances that would predict the exact outcome of the event.

Schraw, Dunkle, Bendixen, and Roedel (1995) used what they referred to as a discrimination index to measure the degree to which confidence for correct answers was higher than confidence for incorrect answers. Therefore, good discrimination ability means that a person is more confident about his or her knowledge on his correct than on

his incorrect answers. The index numbers could range from -1 to 1. A score close to zero indicates no ability to discriminate between correct and incorrect answers. According to Kelemen, et al. (2000) a discrimination score higher than zero means that "confidence is high for correct answers but low for incorrect answers" (p. 98). When a discrimination score is lower than zero, it indicates lower discrimination ability with low confidence for correct responses and high confidence for incorrect responses.

Conflicting Perspectives of Monitoring of Performance on a Test

According to Schraw (1997) there are two opposing views of how examinees monitor their performance on a test. One view is that test examinees monitor their performance on a test by using general metacognitive knowledge which is independent of the type and domain of the test. Examples of these general knowledge skills are rereading the questions and comparing the question with knowledge in memory. Another type of regulatory mechanism of test performance is domain specific and relates to the knowledge in a specific domain. According to the domain specific view, if a person's knowledge in a specific domain improves, the monitoring skill also improves.

Domain general and domain specific hypotheses were tested in several studies (Kelemen et al., 2000; Schraw, 1997; Schraw et al., 1995). Schraw conducted a study with 95 college students who answered multiple choice tests on lexical judgments, reading comprehension, mathematical reasoning, and syllogistic reasoning. If the domain specific hypothesis was correct, correlations between confidence judgment and test performance within one test domain would be high, and correlations on different tests would be lower. If a domain general hypothesis was correct, confidence judgments would

be intercorrelated across different tests. However, these intercorrelations would reveal little about the general test taking monitoring skills. In the study, students answered a 10-item general monitoring checklist before taking the four tests. The results of the study confirmed a domain general hypothesis because confidence judgments on different tests were significantly intercorrelated even though the performance scores on different tests were not. Confidence judgments were correlated with scores on the general monitoring skills checklist, but they were independent of performance scores on all tests. According to these results, domain knowledge, which was important for test performance, did not affect the confidence judgments. Most likely test takers use domain specific knowledge to perform on the test and domain general metacognitive knowledge that helps them to regulate their performance.

Schraw et al. (1995) also tested domain specific and domain general hypotheses of monitoring skill by assessing correlations among monitoring scores across different tests. Several measures were reported: performance, confidence, discrimination index, and the bias scores. The discrimination index and the bias scores measured monitoring proficiency. The discrimination index measured "the degree to which confidence for correct items exceeds confidence for incorrect answers" (p. 435). The study's results revealed that confidence judgments and bias scores were not related to domain specific performance, but discrimination indices were highly correlated to domain specific performance. Confidence and bias scores were highly correlated across domains; discrimination scores were uncorrelated across domains. It appeared that the ability to discriminate correct and incorrect items was a domain specific phenomenon. The

researchers explained this finding as “a compromise between a domain general and domain specific hypothesis” (p. 441).

When Schraw and Roedel (1994) investigated whether judgment of performance differed across different domains, they kept the test difficulty constant by selecting the items of similar difficulty from a larger pool of items administered in their previous study in 1992. The researchers predicted that the performance judgment errors would not be related to the domain of test that a student takes if the tests were of comparable difficulty. Thirty-five undergraduate students were tested in three different domains: reading comprehension, calculation of simple probabilities, and estimation of length of a line segment. Confidence judgments were statistically significantly lower for the probability test than for the reading comprehension and spatial judgment tests. Bias scores were not different across the tests. Schraw and Roedel reported that students might be aware of their limited knowledge in certain domains and more cautious when they judge their performance in less familiar domains. According to Schraw and Roedel, the study results suggested “the presence of a general monitoring skill that varies between individuals, but varies little within individuals” (p. 67).

The cited studies reported different measures of performance monitoring that resulted with inconsistent conclusions. Interpretation of the results also differed. For example, Pressley and Ghatala (1988) found a lower calibration on reading comprehension test items than on analogies and opposites test items. Schraw and Roedel (1994) reported a significantly lower confidence judgment for probability tasks than for reading comprehension and spatial judgment tasks. However, bias scores were not

different on these three types of tests. Based on these findings, Schraw and Roedel suggested that "monitoring one's performance may be a stable cognitive ability once difficulty is controlled" (p. 67). They explained the differences of their findings from Pressley and Ghatala's study by the way the results were reported. Pressley and Ghatala reported correlations between test performance and estimated test performance. Schraw and Roedel reported differences between true and estimated performance. According to Schraw and Roedel the differences between two studies were:

> Their study indicates that test performance is a poor predictor of estimated performance on the reading comprehension test as opposed to the opposites and analogy test. Our findings suggest that individuals were biased to an equal degree across the tests, regardless of correlations between test performance and estimated performance (p. 67).

Factors that Affect Metacognitive Monitoring on a Test

Based on reviewed literature, several factors that affect metacognitive monitoring emerged, and they will be presented in the following order: format and timing of a test, prediction of performance after taking a test, practice on multiple exams, and item difficulty and test difficulty. Additionally, examinee's characteristics and environmental factors might affect monitoring accuracy (Schraw et al., 1995). Gender and culture might affect calibration of performance (Lundeberg et al., 2000; Lundeberg, Fox, & Puncochar, 1994).

Format of a Test and Timing of a Test

According to Pressley and Ghatala (1990) the nature and timing of a test may influence monitoring accuracy and prediction of test performance. Ghatala, Levin, Foorman, and Pressley (1989) conducted a study with 70 fourth grade students to investigate the effect of test taking and the effect of test format on children's perceived readiness for examination performance (PREP). Children were assigned to one of four conditions: the Study, Test, Estimate, and Feedback condition. In the Study condition children read the passage as many times as they wanted before taking the test. In the Test condition children read the passage once and took the test, but no performance feedback was given to them after the test. Children could attempt as many repetitions of study and test trials as they wanted to reach 100 % mastery on the test. In the Estimate condition, after taking the test children estimated how many items on the test they answered correctly. In the Feedback condition children were informed about the number of items answered correctly. Only children in the Feedback condition regulated their studying to reach the mastery criterion while in other conditions children stopped studying prematurely. The results indicated that only taking the test without receiving feedback was not enough to improve children's PREP and studying regulation.

In the second experiment of the same study, Ghatala et al. (1989) investigated a phenomenon of overestimation of the test performance on multiple choice exams. Seventy-two fourth and third grade students were assigned to one of three conditions: Estimate-Decision, Estimate-No Decision, and No Estimate-No Decision. In the first condition students were asked to study a passage and take the test. Children could decide

when to stop studying, but the goal was to reach 100 % correct answers on the test. They wrote down the number of responses they believed they answered correctly after taking the test. In the Estimate-No Decision condition instructions were the same except there was no mentioning of necessity to make a study decision. In No Estimate-No Decision condition children were neither required to make study decision nor to judge their performance on the test. Half of the students in each condition were asked to give a certainty rating about the correctness of each answer by circling one of the numbers from 1-4 that were labeled as: "Not sure at all, just guessing. A little bit sure. Pretty sure. Really, really sure" (p. 58). Children overestimated the number of their correct answers in both estimate conditions. They gave significantly higher certainty ratings to their correct than to their incorrect answers.

In the third experiment of the same study, Ghatala et al. (1989) converted multiple choice tests to a short answer format. The procedure and instructions were the same as in the Experiment 1; only the format of the test was different. Forty-two fourth-grade students participated in the third experiment. In the Test and Estimate condition, students attempted more trials before reaching a mastery criterion on the test than in the Study condition. When they answered multiple choice tests, there was no difference in number of trials among the three conditions. According to Ghatala et al., when given the opportunity to monitor and assess their performance on a short answer test instead of a multiple choice test, students were more persistent in studying to reach the mastery criterion. The results of experiment indicated that multiple choice formats might be responsible for children's inflated sense of preparedness for the test.

Pressley and Ghatala (1988) found that college students often felt certain that their answers on multiple choice questions were right but in fact they were wrong. According to Pressley and Ghatala and Ghatala et al. (1989) the format of the test most likely affected monitoring of performance. Multiple choice questions included distractors that contained information based on the previous knowledge or familiar information from the text. Therefore, because of these distractors, an examinee might believe that his or her incorrect answers were correct when they were not.

As already mentioned in the Ghatala et al. (1989) study, the time when self-assessment was undertaken, before or after a test, could also be important for prediction of performance.

Prediction of Performance after Taking a Test

Several studies found a greater accuracy of a test performance prediction after test was taken than before test was taken (Pierce & Smith; 2001; Pressley et al., 1987). Pierce and Smith called an assessment of a test performance after test was taken "postdiction" and "the accuracy of postdiction relative to prediction judgments as the postdiction superiority effect" (p. 62). They tested two hypotheses in relation to this effect: a retrieval hypothesis and a test knowledge hypothesis. The test retrieval hypothesis referred to an examinee's memory of what was happening while answering the questions. The test knowledge hypothesis referred to an examinee learning about the nature of the test. Both hypotheses assumed that postdiction would be better than prediction of performance before taking a test. However, the test knowledge hypothesis predicted that postdiction superiority effect would disappear as an examinee moved from one set of questions to

another. Students read four narrative texts in the first experiment and three expository texts in another. Students predicted their performance on four sets of questions on each text before and after taking the test. The researchers found a significant postdiction superiority effect. Both prediction and postdiction did not change across the set of questions which confirmed the test retrieval hypothesis.

Pressley et al. (1987) reported that self-assessment after a test was taken was more accurate than before the test was taken. The participants of the study were 54 college students enrolled in an introductory psychology course. In the first experiment, students read a part of the chapter from the Human Development textbook book and answered 50 multiple choice questions with four choices per item. Questions were taken from the test bank that accompanied the textbook and tested factual knowledge. Students were randomly assigned to three experimental conditions: to make a prediction about their performance on the test before they read the chapter, after they read the chapter, and after they took the test. An observation of subjects through a one-way mirror confirmed that they indeed read the chapter. Students were also asked whether they needed to reread the chapter to get 20%, 40%, 60%, and 80% of the items correct. A prediction inaccuracy score was calculated by finding the difference between the predicted number of correct answers and the number of correct answers. A statistically significant difference was found only between the after test taking group and the before reading the chapter group on prediction of performance. The group of students who made their prediction after taking the test knew better whether they needed to reread the chapter or not than the group predicting performance before reading the chapter.

In the second experiment, students read a new chapter from the book they regularly used in the course. The exam consisted of 26 fill-in-the blank items taken from the study guide that accompanied the textbook. Two measures of students' ability were obtained: the performance on their midterm exam and the reading comprehension passages on the Scholastic Aptitude Test and the Graduate Record Exam. Predictions of performance were statistically significantly different between the after-testing condition and the before-reading the chapter condition. The study results did not reveal the ability differences between students who overestimated and underestimated their performance. Both experiments found that students' PREP was improved by a combination of reading the chapter and taking the test (Pressley et al., 1987).

In the third experiment, 162 undergraduate students who were enrolled the following year in the same introductory psychology course participated in the study. They were asked to provide the best estimate of their performance before reading the chapter, after reading the chapter but before testing, and after testing. In this experiment, adjunct questions were added to the text in a massed and an interspersed form. In the massed form, adjunct questions were added at the end of the chapter. In the interspersed form, questions were added throughout the chapter. Students were told not to look in the text to answer adjunct questions. The group of students who were estimating their performance after reading the chapter and having adjunct question in the text, differed significantly in estimation of their performance from students who did not read the chapter and did not have adjunct questions in the text. Within the after-reading the chapter condition, students were more accurate in prediction of their performance if they

answered interspersed adjunct questions than massed adjunct question. However, the results indicated that taking a test might be more important for self-evaluation than answering adjunct questions (Pressley et al., 1987).

Taking a test had a stronger effect on assessment of performance and expectations about future performance among students in grades 7-8 than among students in grades 1-5 grades (Pressley & Ghatala, 1989). Students were assigned to one of two conditions: before-test prediction and after-test prediction of performance. Students were administered one of two forms of a vocabulary test. Picture sets from the Peabody Picture Vocabulary Test-Revised served as alternatives for vocabulary words on multiple choice tests. On the easy test, 20 words were identified by two judges as the easiest and 10 as difficult. On the hard test, 20 words were difficult and 10 easy. Older students discriminated better between easy and hard items on the test. They were less confident about the correctness of their answers on hard items than younger students. After taking the hard test, students in grades 7-8 had lower expectations about their overall and item-level test performance in the future. The study results indicated that a developmental change occurs in monitoring of performance and self-regulation (Pressley & Ghatala).

The cited studies consistently reported that postdiction or retroactive self-evaluation is better in assessing performance. Taking a test appears to be "a metacognitive experience" (as cited by Ghatala, et al., 1989, p. 51; Flavell, 1979). Other researchers argue whether practicing on multiple tests could change metacognitive monitoring. The results in this area have not been consistent.

The Effect of Practice on Multiple Tests and Metacognitive Monitoring

Pierce and Smith (2001) examined metacomprehension accuracy of text on successive tests. Retrospective self-assessment or postdiction was found to be more accurate than prediction of performance before attempting a test. Across successive tests, neither prediction nor postdiction accuracy improved, which contradicted the hypothesis that accuracy should become better on successive exams because of accumulated knowledge about the nature of the test.

Hacker et al. (2000) examined the effect of practice on prediction and postdiction accuracy on three consecutive exams in a classroom context. Students who scored above 70% correct on their exams showed mostly accurate predictive and postdictive judgment of their performance with the highest scoring students showing consistent underconfidence on both judgments. The lowest scoring students, who scored below 50%, showed little accuracy but gross overconfidence on prediction and postdiction of test performance. There was a difference between lower and higher scoring students on their predictive and postdictive accuracy improvement. After having taken three exams, at the end of semester, high performing students had improved their accuracy on both judgments, especially on postdictive. Low performing students did not show improvement on their prediction or postdiction of test performance at the end of semester.

Nietfeld et al. (2005) examined differences between lower and higher performing students on monitoring accuracy in a classroom setting throughout the semester on three 25-items multiple choice exams and on a 50-item comprehensive final exam. Twenty-seven undergraduate students enrolled in an educational psychology course participated

in the study. The researchers investigated whether monitoring accuracy changed on consecutive exams throughout semester as measured by local and global monitoring and bias score. The hypothesis was that monitoring accuracy would not improve because no feedback or explicit training on monitoring was provided to students. After tests were returned, students could review their item by item performance, their confidence judgments, and they could ask questions about any item. Grade point average (GPA) score was obtained as a measure of students' general academic ability. All three indices of monitoring accuracy, local, global accuracy, and bias remained stable across four exams for all students. The researchers explained this finding with the lack of an explicit training on metacognitive monitoring and strategy use. The poorest monitoring accuracy occurred on the third test that had the lowest mean level of performance. Grade point average and test performance appeared to be the strong predictors of local monitoring accuracy. Overall, students were better calibrated on easy items with a tendency to be overconfident on difficult items and underconfident on easy items.

Important in both the studies of Hacker et al. (2000) and Nietfeld et al. (2005), was the observation of college students' metacognitive monitoring in classroom throughout semester. For example, Nietfeld et al. reported that their study was one of the first studies that observed this process "in a naturalistic setting over a substantial period of time" (p. 7). However, the researchers in these two studies reported different results on monitoring accuracy regarding the students' ranking in the class. Nietfeld et al. did not find any difference between higher and lower ranking students on monitoring accuracy over time. The poorest monitoring accuracy occurred on the third test which was

apparently the most difficult test as shown by the lowest average performance. Higher achieving students exhibited greater local accuracy of prediction on individual items of the test than lower achieving students, but overall the students were better calibrated on easy items.

Nietfeld et al. (2005) compared their finding with results of the Hacker et al. (2000) study regarding high and low achieving students. Hacker et al. examined 19 students at the top portion and 7 students at the bottom portion of a class of 99 students. The study reported a gain for higher achieving students in accuracy monitoring over time by "increasing amounts of variance accounted for in performance between the first and third exam" (Nietfeld et al., p. 23). Nietfeld et al. reported that a limitation of their study was a relatively small sample which included only 27 undergraduate students enrolled in an educational psychology survey course.

The effect of overt practice on calibration across five quizzes in a college undergraduate class over a 15-week period was assessed in the study conducted by Bol et al. (2005). Three hundred and sixty-five undergraduate students who were enrolled in an educational psychology course participated in the study. Students in calibration practice group entered their performance predictions and postdictions before and after completing five multiple choice quizzes on line. Students in a no-practice condition did not enter their predictions or postdictions of performance. Students were enrolled in either an on-line or traditional version of the course. Quizzes were posted online for both versions of the course. The hypothesis was that students who practiced their prediction and postdiction of performance on line would become more accurate on consecutive quizzes,

and they would score higher on the final exam than the students who did not practice. The results of the study revealed that the practice did not have a significant impact on prediction and postdiction accuracy or on achievement on the final exam. The study also investigated the difference in calibration between higher and lower achieving students. Students were assigned to a higher or a lower achieving group based on a median split of their performance on the final exam. There was a difference between the higher and lower achieving students in their calibration accuracy, especially on prediction accuracy. Higher achieving students were significantly more accurate in their prediction but slightly underconfident. Lower achieving students were less accurate but overconfident in their prediction of performance. Students exhibited greater postdiction than prediction accuracy but not as great a difference as was expected. For prediction accuracy, there was a significant quadratic trend on within subjects' contrasts over five quizzes. On postdiction accuracy the trend was linear.

Bol and Hacker (2001) assessed the impact of practice test on students' prediction and postdiction accuracy and on exam performance for multiple choice and essay type of test in an introductory research methods graduate level class. Fifty-nine graduate students were divided by a median split of their performance into a high and a low performing group. An interesting finding was that the students who took a practice test scored significantly lower on multiple choice items of the midterm exam and were less accurate in assessing their performance. The explanation for this finding was that students used the practice test as their study guide and probably expected identical questions on the midterm exam. Before they took the final exam, students already received feedback about

their performance and changed their study strategy. Another finding was that high achieving students were more accurate in calibration of their performance than low achieving students. There was no difference in calibration of their performance between an essay and a multiple choice type of test on both midterm and final exam. Low-achieving students showed better prediction and postdiction on an essay type than on a multiple choice test. They were less accurate but overconfident about their performance.

An important factor that affects metacognitive monitoring on a test that was mentioned in earlier discussion of several studies (e.g., Hacker et al., 2000; Nietfeld et al., 2005) is item difficulty and test difficulty.

Item Difficulty and Test Difficulty

Better calibration on easier items of a test and a tendency to be overconfident on difficult items and under confident on easy items was reported in several studies (Gigerenzer, Hoffrage, & Kleinbolting, 1991; Lichtenstein & Fischhoff, 1977; Nietfeld, et al., 2005; Pressley & Ghatala, 1988, Schraw & Roedel, 1994). Gigerenzer et al. referred to increase of overconfidence with the difficulty of the questions as the "hard-easy effect." The overconfidence effect happens "when the confidence judgments are larger than the relative frequencies of the correct answers" (p. 512). The difficulty of the questions was defined as "the percentage of correct answers." A decrease of overconfidence when questions become more difficult was called "a reversed hard-easy effect" (p. 512).

Pressley and Ghatala (1988) suggested that researchers should hold the test difficulty constant in studies of metacognitive awareness. The results of their study

indicated that college students had better ability to discriminate what they know from what they do not know on easier items of a text comprehension test. Lichtenstein and Fischhoff (1977) reported better calibration of examinees on easier items than on more difficult items of a general knowledge test. They found similar results on tests that differed in their difficulty. The participants overestimated accuracy on the hard test and underestimated accuracy of their answers on the easy test.

Schraw and Roedel (1994) indicated that overconfidence was due to the test difficulty since overconfidence did not differ on unrelated tests after test difficulty was controlled. They tested the hypothesis of whether overconfidence about performance was due to test (item) difficulty or person-driven errors. They also examined whether trend of judgment error across different levels of item difficulty was linear or not. In experiment one, 48 university undergraduate students calibrated their performance on the multiple choice Nelson-Denny Reading comprehension test. The items were divided in three groups: easy, moderate, and difficult. Participants marked their ratings for confidence judgment of test items on a 100 mm graphic rating scale. Bias scores and the difference between the average performance scores and estimated performance scores were calculated for each of the three groups of items. Students showed higher overconfidence or higher bias score on more difficult items. Average bias scores were low on easy items of the test. Bias score had a strong linear trend with test item difficulty. It also appeared that examinees adjusted their judgment to their perception of the test difficulty because moderately difficult items also led to overconfidence; therefore, to some degree judgment errors could be attributed to a person.

Morse and Morse (2002) investigated whether college students' choices of items as easy or difficult were correlated with item difficulty, defined as "the proportion of examinees who answer the item correctly" (Crocker & Algina, 1986, p. 311). Students in the Morse and Morse study were asked to mark during the test five items they believed were the easiest, and five items they believed were the most difficult. A statistically significant correlation was found between item difficulty and frequency of items chosen by students as difficult and easy. A majority of the questions chosen as difficult were judged as measuring comprehension or application level skills according to Bloom's taxonomy, whereas all the items chosen most frequently as easy were found to be at the knowledge level. The study results indicated that students had an awareness of the easiest and the most difficult questions on the test.

Studies inconsistently defined the item difficulty within a test. Two studies (Morse & Morse, 2002; Pressley & Ghatala, 1988) used the proportion of examinees who answer an item correctly as the numerical values of item difficulty. Several other studies presented item difficulty differently. Schraw and Roedel (1994) divided items of 28 multiple choice tests in three groups: easy, moderate and difficult. According to Schraw and Roedel, the criteria for these three difficulty levels were based on data gathered in their previous research in 1992. Nietfeld et al. (2005) provided examples of an easy question and difficult question on the multiple choice test used in their study. An easy question was defined as requiring simple identification and more difficult questions as requiring application of knowledge. Sinkavich (1995) indicated the percentage of the questions on all exams in his study that were "knowledge" and those which were

"application" items. The researcher did not specify this division of questions as equating to easiness or difficulty.

In addition to characteristics of a test and test items, examinee characteristics might affect metacognitive monitoring on a test.

Examinee Characteristics

Beside high and low achievement on a test, other factors related to examinee characteristics might affect metacognitive monitoring on a test. According to Schraw et al. (1995) these factors include individual characteristics such as mood, impulsivity, familiarity with the domain, and intellectual ability. According to Schraw et al., studies that examined the relationship between examinee's ability and metacognitive monitoring did not find significant relationships between the two variables. One of the examinee characteristics that might affect accuracy of metacognitive monitoring is their previous knowledge (Lichtenstein & Fischhoff, 1977; Nietfeld & Schraw, 2002). Findings of these two studies will be presented in more detail.

Lichtenstein and Fischhoff (1977) investigated relations between previous knowledge and calibration of performance. The participants in the study were students, volunteers who responded to an advertisement in the college newspaper. Students lacked an awareness of how little they knew and overestimated their knowledge on the questions for which their knowledge was very limited. After receiving some training on the task, the trained participants showed better calibration of their answers than untrained group. In an additional experiment, 120 participants were separated in three groups based on their answers on a general knowledge test: the best (40 with 51 or more correct answers

out of 75), the middle (39 with 46-50 answers correct), and the worst group (41 with fewer than 46 correct answers). Each item on the test had only two possible answers. The results of the study strongly suggested that calibration accuracy increased with knowledge. The interesting finding was that all groups had tendency to be overconfident, but the most knowledgeable participants were the least overconfident.

Nietfeld and Schraw (2002) examined the impact of previous knowledge on monitoring accuracy of college students on a 24 item multiple choice probability test. They divided 93 undergraduate college students in three groups based on the type and number of mathematics courses they had before: a high knowledge, mid-knowledge, and a low-knowledge group. The group of students with high prior knowledge performed significantly higher on probability test, and this group significantly better monitored their performance as indicated by higher accuracy score. However, prior knowledge did not affect bias and confidence judgment on probability test. Confidence judgment, or how confident each individual was that his or her response was correct, was measured on a scale of 1-100 for each question. Bias measured the severity of over and under confidence. Accuracy measured the degree to which confidence about the correctness of each response matched the actual performance.

Environmental Factors

Calibration of performance might be affected by environmental factors such as incentives given to examinees, feedback during testing, and adjunct questions included during studying (Ghatala et al., 1989; Pressley et al., 1987; Schraw et al., 1995; Schraw et al., 1993). Schraw et al. (1993) found that incentives given to college students in the form

of a double extra credit improved their calibration of performance. A verbal feedback which was given to students about their performance on the test did not have an effect on their performance, accuracy, or bias scores. Pressley et al. (1987) added adjunct questions to the text that were similar to test questions. Two types of adjunct questions were added: massed (at the end of the chapter) and interspersed (throughout the chapter). Students were told not to look in the text to answer the adjunct questions. The group of students who read the chapter and had adjunct questions added to the text, estimated their performance on multiple choice test better than the students who did not read the chapter and did not have adjunct questions. Within the after-reading condition, students were more accurate in their prediction of performance if they answered interspersed adjunct questions than massed adjunct questions. However, the results indicated that taking a test might be more important for self-evaluation than answering adjunct questions.

Additionally, culture and gender might affect calibration of performance (Lundeberg et al., 2000; Lundeberg, et al., 1994). The influence of culture and gender on confidence judgments was investigated by Lundeberg et al. (2000). The researchers examined the differences in confidence calibration and discrimination of college students across gender and culture in five different countries: Israel, Palestine, the Netherlands, Taiwan, and the United States. Significant differences were found in overall confidence, confidence when wrong, and confidence when correct across the countries, but gender differences within countries were small. Lundeberg, et al. (1994) reported that gender differences on confidence judgments of 70 male and 181 female college students were dependent on whether answers were correct or wrong and on the content of questions

being tested. Women showed more accurate perceptions of their incorrect answers; men, especially undergraduate men, showed overconfidence when they were wrong. Overall, the researchers indicated that women had a greater tendency to be more accurate in calibration of their performance. They did not necessarily lack confidence, but men showed too much confidence in wrong answers.

In relation to students' judgment of performance on a test and self-regulation, the construct of self-efficacy, which has been investigated in different fields, will be presented in more detail in the following section.

Self-efficacy, Self-regulation, and Monitoring of Performance

Human nature is not a set of unrelated entities that work independently from each other. Self-regulation and metacognitive monitoring, which is in the center of self-regulated thinking (Pressley & Ghatala, 1990), are examples of interaction among different factors. Self-regulation is described in the literature by overlapping theories. One of the theories that describes self-regulation and monitoring of performance is social cognitive theory. The main principles of social cognitive theory are the self-efficacy and triadic reciprocal determinism (Pajares, 1996a).

According to Pajares (1996a) self-efficacy beliefs influence people's choices and actions, their effort, persistence, and resilience. Thought patterns, emotional reactions, and motivation are affected by self-efficacy perceptions. Expectancy beliefs or beliefs that are "specific to one's perceived competence" (p. 544) have been investigated in the area of education. Self-perceptions of academic competence can be general and domain

specific. Generalized self-efficacy beliefs are better predictors of general academic performance such as choosing major or obtaining grades. Domain specific beliefs have better predictive value for academic performance within specific domains. Pajares emphasized that efficacy-beliefs in educational research should be assessed "at the optimal level of specificity that corresponds to the critical task being assessed and the domain functioning being analyzed" (p. 547).

Another important principle of social cognitive theory described by Bandura (1986) is the model of "triadic reciprocality in which behavior, cognitive and other personal factors, and environmental events all operate as interacting determinants of each other" (p. 18). From the perspective of reciprocal determinism, social cognitive theory explains self-regulated learning. Self-regulated learning is influenced by personal processes, but these processes are influenced by environmental and behavior factors in a reciprocal fashion (Zimmerman, 1989; Zimmerman & Martinez-Pons, 1990).

According to Zimmerman (1989), personal characteristics that affect self-regulated learning are declarative knowledge, procedural knowledge (how to use strategies), conditional knowledge (when and why to use it), and decision making processes based on the learner's long term goals. Long term goals are influenced by one's self-efficacy perception and affective state.

Beside personal characteristics of a learner, self-regulated learning is influenced by behavioral and environmental factors. Behavioral factors include self-observation, self-judgment, and self-reactions to one's performance. Self-observation refers to monitoring of one's own performance. Self-judgment refers to students' comparison of

their performance with a standard or a goal. Self-reactions of one's performance involve, for example, goal setting, self-efficacy perception, and metacognitive planning that work together in a reciprocal fashion. All these factors are interrelated and affect each other. Examples of environmental influences include modeling and the structure of a learning context (Zimmerman, 1989).

The relationship between self-efficacy beliefs in mathematics, metacognitive monitoring accuracy on a probability test, and a training strategy for solving probability problems was examined in the study conducted by Nietfeld and Schraw (2002). One group of participants was trained in a two hours session on strategies for solving probability problems and another was not. The strategies taught did not explicitly address metacognitive monitoring of performance. Students in both the experimental group (n = 32) and the control group (n = 26), took the 24-item probability test before, after, and one week after strategy training. At the beginning of the study, students completed the Raven Advanced Progressive Matrices Test as an indicator of students' general cognitive ability. Before taking the pre-test on probability and the Raven Progressive Matrices test, students answered 10-item general mathematics self-efficacy questionnaire. The same self-efficacy questionnaire was answered by students during the immediate post-test period and again one week after the strategy training. After each probability test, students made their confidence ratings of having answered each test item correctly by using a 100-mm graphic line which was labeled as 0% confidence on the left end and 100% confidence on the right end. Monitoring proficiency was measured by two indices: the bias score and the accuracy score. Bias score was calculated by dividing by 100 the

difference between the average confidence and average performance scores on the probability test. “Scores greater than zero indicated overconfidence, and scores less than zero indicated under confidence” (p.134). The accuracy score measured “the degree to which confidence judgments for each item matched actual performance” (p.134). Accuracy was the average of absolute difference between confidence scores (0 -100) and item performance (0 or 1) across items. The values of accuracy score ranged from 0 (indicating perfect accuracy), to 1 indicating total inaccuracy.

Nietfeld and Schraw (2002) analyzed self-efficacy scores and monitoring proficiency scores. Self-efficacy scores were significantly positively correlated to the Raven Advanced Progressive Matrices, the pre-test and the immediate probability post-test score, all three confidence and monitoring accuracy scores, but not to the bias. Self-efficacy scores were highly related across the three testing periods. Strategy training had no impact on self-efficacy general mathematics score. The researchers explained this finding as attributable to the nature of self-efficacy questionnaire which contained questions about general mathematics ability and not about specific self-efficacy on probability problems. However, the results of the study indicated that the brief strategy training improved performance and monitoring accuracy on the immediate probability post test even though monitoring strategy was not explicitly included in the training. Monitoring accuracy was not improved on the one week delayed post-test. According to Nietfeld and Schraw, a lack of improvement of monitoring accuracy on the delayed post-test could be attributed to unreliability of the test or to the training itself. The researchers

suggested that strategy training over longer period of time (e.g., 10 weeks) with a component of explicit monitoring accuracy would produce a longer lasting effect.

Use of strategies (cognitive and metacognitive) in relation to self-efficacy and self-regulated learning will be presented in the following section.

Self-efficacy, Self-regulated Learning, and Use of Strategies

Self-regulated learning includes a person's use of learning strategies to achieve academic goals based on his or her self-efficacy perception (Zimmerman, 1989). Students will use strategies if they are motivated to achieve a particular academic goal. If students do not believe that strategies will produce a desired outcome, they will not use either cognitive or metacognitive strategies. One's perception of self-efficacy might have a critical influence on whether or not strategies will be used (Garner & Alexander, 1989; Schraw, 1998).

Garner and Alexander (1989) defined strategies (cognitive and metacognitive) as "means of reaching goals efficiently and effectively" (p. 149). The difference between cognitive strategies and metacognitive strategies is that cognitive strategies are "invoked to make cognitive progress" while metacognitive strategies monitor that progress (Flavell, 1979; Garner & Alexander). A student can use general strategies or domain specific. If a student does not have enough background knowledge in the subject, he might use general strategies to find necessary information and compensate for the lack of that knowledge. In other cases, the application of strategies will correspond to the specific domain. For example, certain strategies will be used for learning historical dates, and other strategies will be used to write an essay. A strong strategy in one domain might

be a weak strategy in another domain (Garner & Alexander). Another important issue is the relation between content knowledge and strategy knowledge. Strategy instruction may work only for students who have a certain amount of content knowledge. Garner and Alexander emphasized that "'Knowing' and 'knowing how to know' both matter for all sorts of academic tasks, and both can be enhanced with instruction" (p. 152).

In addition to having certain skills, students need to be motivated to use them in order to succeed in classroom (Pintrich & De Groot, 1990). As stated by Pintrich and De Groot (p. 38) "students need to have 'skill' and 'will' to be successful in classroom." Motivational components that are related to self regulated learning include: a) an expectancy component, b) value component, and c) affective reaction to the task. An expectancy component refers to students' beliefs about their ability to perform the task, and it is linked to students' self-efficacy, cognitive strategies, metacognitive strategies, and effort. A value component relates to students' beliefs about the importance of the task. Affective reactions to the task include the person's feelings about the task. In a classroom context, the most important affective reaction seems to be test anxiety. Pintrich and De Groot conducted a study with 173 seventh and eighth grade students from science and English classrooms. Students answered 56 items of the Motivated Strategies for Learning Questionnaire (MSLQ) that included items on motivation, cognitive and metacognitive strategies, and management of effort. Student performance was measured by the scores on their seatwork, quizzes, essays, and two exams. Pintrich and De Groot found that students who reported higher self-efficacy also reported the use of more cognitive and metacognitive strategies. Higher use of strategies was significantly

correlated with performances on all assignments except on the seatwork. According to Pintrich and De Groot self-efficacy *"*was not significantly related to performance on seatwork, exams, and essays when the cognitive engagement variables were included in regression analyses. These findings suggest that self-efficacy plays a facilitative role in relation to cognitive engagement" (p. 37).

Zimmerman and Martinez-Pons (1990) reported interesting differences between gifted and regular students about their verbal and mathematical self-efficacy and learning strategies. Forty-five boys and 45 girls of the 5th, 8th, and 11th grades from a school for academically gifted students and the same number of students from regular schools participated in the study. Two areas of academic efficacy were investigated in the study: verbal comprehension and mathematical problem solving. Students answered the Verbal Efficacy scale that involved 10 words, and the Mathematics Efficacy scale that involved 10 problems. Students were asked to rate their efficacy to define each word and solve each mathematical problem using a scale that ranged from 0% (completely unsure) to 100% (completely sure). Students also answered a structured interview that was developed to measure self-regulated learning strategies. Gifted students expressed higher verbal and mathematical self-efficacy and higher use of learning strategies than non-gifted students. Surprisingly, girls reported lower verbal self-efficacy than boys but reported using more learning strategies than boys. No data on students' performance was available in the current study to compare students' performance with their self-efficacy beliefs.

Pajares (1996b) reported that middle school gifted girls were biased toward underconfidence on solving mathematical problems when compared with gifted boys. In general, gifted students were less likely to overestimate their performance on algebra problems than non-gifted students. There were no difference on calibration measures between non-gifted girls and boys.

Self-efficacy and Judgment of Performance

According to Bandura (1993) people might have the same knowledge and skill, but they will perform differently because of their different self-efficacy beliefs. Bouffard-Bouchard (2001) investigated the relationship between self-efficacy beliefs on a verbal concept-formation task and performance of 64 Canadian college students. The task consisted of seven sets of six sentences. Within each set, the same target word was replaced by a nonsense word in each sentence. Students were asked to discover a meaningful word that would best replace a non-sense word. Students were tested on their cognitive skills and initial performance to guarantee that these skills were equivalent across groups. Students in the experimental group received positive feedback regardless of their performance on the task with intention to increase their perceived self-efficacy. Students in the comparison group received negative feedback regardless of their performance. Students' perceived self-efficacy was measured by asking them whether they believed they would succeed on a task or not. Additionally, they were asked to assess how confident they were in their success on a scale ranging from very unsure to completely sure. In the assessment of performance, students were asked to report the level of certainty about correctness of their responses. Students in the positive feedback

group solved more problems and were more confident about their eventual success than those in the control group. Also, students with a higher sense of self-efficacy evaluated more accurately correctness of their responses.

Klassen (2002) reviewed the studies that investigated calibration of self-efficacy beliefs of students with learning disabilities (LD). Calibration was defined as "the degree of congruence between efficacy beliefs and actual performance" (p. 89). The purpose of the review was to examine congruency of LD students' self-efficacy beliefs who had poor metacognitive skills with their actual performance. In most of the reviewed studies students overestimated their performance. Students with learning disabilities were more optimistic about their writing abilities than their mathematics abilities. In mathematics, students' calibration of their self-efficacy and performance was more accurate than in writing. None of the reviewed studies compared low achieving students with LD students. Klassen made an assumption that low achieving students might have a similar miscalibration issue as LD students that needs to be addressed in future research.

Pajares (1996a) asked the question: "But how much confidence is too much confidence, when can overconfidence be characterized as excessive and maladaptive in an academic enterprise, and what factors help create inaccurate self-perceptions?" (p. 565). According to Pajares, the emphasis should be on improvement of students' calibration, so students can distinguish what they know from what they do not know. When students are able to more accurately assess their performance, they will more likely use different cognitive and learning strategies to improve their performance. In the example of students with LD, Klassen (2002) reported that misjudgment in self-

evaluation can be potentially harmful. A knowledge deficiency might lead to faulty task understanding, not using appropriate strategies, and difficulties with self-regulation and self-monitoring. Otherwise, lower academic functioning leads to lower functioning of other skills and inaccurate judgment of one's own performance. According to Pajares (1996a) the main challenge would be to improve calibration without lowering students' confidence and optimism. This conclusion goes back to the already mentioned Kruger and Dunning (1999) study when students improved their estimation of performance after they were trained on the task and became more competent. A better competency provided students with better ability to estimate their knowledge.

A factor that is related to self-regulated learning, self-efficacy and metacognition but not well researched in this context is test anxiety. A basic background of these relations will be provided in the following section.

Test Anxiety in Relation to Self-efficacy and Metacognition

According to Bandura (1989) self-efficacy beliefs affect people's cognitive, motivational, and affective processes that influence their actions. People who believe they are not capable to control potential threats in certain situations will experience a higher level of stress and anxiety. A person's perception of coping deficiency with the situation might act as a cognitive mediator of anxiety arousal. A key factor in regulation of cognitively generated anxiety is a perceived self-efficacy to control one's intrusive thoughts (Bandura, 1989; 1993). In academic settings, students who have a low sense of self-efficacy to cope with academic demands will be at risk to develop scholastic anxiety.

Their perceived academic self-efficacy might depend on previous successes and failures (Bandura, 1993).

Test anxiety was examined in relation to students' metacognitive skills (e.g., Veenman et al., 2000) and an information processing model which explains a poor performance of test-anxious students due to problems in encoding and retrieval of information (Benjamin et al., 1987). Therefore, Benjamin et al. distinguished between two types of test anxious students. Type I students have inefficient study and concept organizational skills. They have difficulty with encoding information, and they perform poorly in both evaluative and non-evaluative contexts. Type II students have sufficient study skills and concept organization skills, but they fail to use them in a test situation. These students have difficulty retrieving information in an evaluation context. Veenman et al. extended this model to evaluate the relationship between metacognitive skills and test anxiety. They differentiated between the two types of students who suffer from test anxiety. The type I student suffers from a deficiency of metacognitive skills such as planning and monitoring. The type II student has the problem of production deficiency because of not knowing how and when to use metacognitive planning and monitoring. Veenman et al. investigated test anxiety in relation to metacognitive skills of secondary school students and their performance in mathematics. Metacognitive skills were examined by students' observation and thinking aloud protocol during the solving of word-math problems. The results indicated that the low anxious students had better metacognitive skills than the high-anxious students. Metacognitive skillfulness was

related to students' performance on the mathematics exam which consisted of six words problems.

According to Pintrich and De Groot (1990) metacognitive skillfulness is not enough; students need to be motivated to use those skills. There are three components of students' motivation: an expectancy component, a value component, and an affective component. The affective component includes students' different emotional or affective reactions to the task. Among different emotional reactions about the task, test anxiety seems to be the most important in educational settings.

Test anxiety was negatively related to self-efficacy beliefs and performance on exams and quizzes in the study that Pintrich and De Groot (1990) conducted with 173 seventh grade students. Test anxiety and self-efficacy beliefs were measured by the subscales of the motivational beliefs on the Motivated Strategies for Learning Questionnaire (MSLQ). The subscales of self-efficacy beliefs and test anxiety that are provided on MSLQ will be also used in the current study (Pintrich, et al., 1991).

Summary

Self-tailored exams in a college classroom have not been investigated in recent research. The attempts to develop this type of testing (Bedard, 1974; Morse, 1988), "have been abandoned in favor of computer-administered testing" (Morse, p. 2)**.** Within the literature on computerized adaptive testing, only self-adaptive computerized testing has allowed an examinee to exhibit some control over the test (Wise, 1994). The current study is specific by its form of a self-tailoring procedure that allows omitting questions from being scored on an exam after all questions have been attempted.

None of the previous studies examined the relation of metacognition in the context of a self-tailoring procedure on examinations. Flavell (1976) offered one of the first definitions of metacognition as "knowledge concerning one's own cognitive processes and products of anything related to them" (p. 232). In a question omitting procedure in the current study, students need to use their metacognitive ability to distinguish between their correct and incorrect answers. Students will be asked to omit questions from being scored that they believe were answered incorrectly or were not certain of the correct answers.

Metacognition has two interrelated parts: knowledge of cognition and regulation of cognition (e.g., McCormick, 2003; Schraw 1998). During performance on a test, students use regulation of metacognition called metacognitive monitoring as an executive process of cognitive actions that are taken (Pressley & Ghatala, 1990). Monitoring accuracy alerts students how to regulate their learning and performance. Several studies referred to monitoring accuracy as the process of matching between perceived and actual performance (e.g., Nietfeld & Schraw, 2002; Nietfeld et al., 2005). Monitoring accuracy has been a topic of investigation in numerous studies. However, a consistent conclusion of what affects monitoring accuracy, and how it needs to be measured has not been reached. According to Kelemen et al. (2000) there is no one reliable and comparable index of monitoring accuracy established in the research literature. Many studies examined students' confidence judgments between their perceived and actual performance on a test (e.g., Nietfeld & Schraw, 2002; Schraw, 1997), but not many

studies examined students' metacognitive ability to discriminate between their correct and incorrect answers (Pressley & Ghatala, 1988; Lundeberg, et al., 2000).

Another unanswered question in recent research is whether metacognitive monitoring is a general skill or domain specific. Different factors have been investigated in recent research that might affect monitoring accuracy: format and timing of a test, item difficulty and test difficulty, practice, previous knowledge, and a level of achievement on a test. However, these factors interact with each other when investigated in educational settings, which makes it difficult to assess the importance of one factor over the other.

Other factors that affect academic performance include self-efficacy and test anxiety. For example, Pintrich et al. (1991) reported a positive correlation between the score on the Expectancy Component: Self-Efficacy for Learning and Performance scale that will be used in the current study and the final grade in a college course. The data were gathered from a sample of 380 college students from 37 different classrooms. Pintrich et al. reported a negative correlation between the score on the Affective Component: Test Anxiety scale and the final grade on the same sample of students. The new angle in this research will be to investigate the relation of the scores on these scales with students' likelihood to apply an option to omit questions from scoring and to the degree of success students have in applying the omission procedure.

Based on the current literature, several outcomes in the current study were expected. First, students who exhibited better metacognitive skills and achieved higher scores on exams would be more likely to apply an option to omit questions from scoring and would have a higher degree of success in applying the omission procedure.

Second, the likelihood to apply an option to omit questions from scoring and the degree of success students have in applying the omission procedure would depend on the difficulty of the test and test items.

Another important aspect that the current study tried to answer is the effect of practice on multiple tests throughout semester on application of the question omitting procedure under investigation. The changes in metacognitive ability were indicated by the percentage of questions omitted that were answered incorrectly. Only a few studies observed the effect of practice on students' judgment of performance on successive exams and found conflicting results (Bol et al., 2005; Hacker et al., 2000; Nietfeld et al., 2005). As a result, it was unclear whether there would be changes in how often and how successfully students opt to omit questions from scoring over a series of examinations.

CHAPTER III

METHODOLOGY

This chapter presents the methodology of the study by describing participants, procedure, instruments, scoring method, and statistical analysis. Statistical analysis methods are reported in the same order as the research questions.

Participants

The participants in the current study were undergraduate college students in the Mississippi State University enrolled in two sections of an educational psychology Human Growth and Development course. Sixty-one students out of 99 from section one, and 38 students out of 39 from section two signed the informed consent. The total number of students who signed the informed consent was 99. Six students from section one and four students from section two dropped the course. Nine of the remaining 89 students did not take all exams. Therefore, 80 students completed the entire study. However, nine students did not omit questions on all exams. Therefore, 71 students used an option to omit questions on all exams. Out of the 80 students who completed the study, 69% were female, and 31% were male students. Sixty-five percent of students were Caucasian, 34% African/American, and 1% Native American. Thirty-one percent of students were juniors, 30% sophomores, 23% seniors, 15% freshmen, and 1% other. The

average age was 20.8 (M = 20.80, SD = 3.59, N = 80) with minimum reported age 18 and maximum of 41 years. The average self-reported ACT score was 22.45 (M = 22.45, SD = 3.99, n =71) with a range from 16 to 30. The average self-reported grade point average was 3.09 (M = 3.09, SD = .60, n = 72) with a range from 1.4 to 4.00.

Students from various majors took a Human Growth and Development course as a requirement for their degrees, or they took the course as a behavioral science elective. Twenty-six percent of students were education majors, 25% kinesiology majors, 10% engineering majors, 10% biological science majors, and 9% business majors. Approximately 11% of students were from majors of nursing, animal and diary science, and undeclared.

The average age, gender, and ethnic composition of both sections were very similar. The average age in section one was 21.1 (M = 21.10, n = 50) and in section two 20.3 (M = 20.30, n = 30). In section one, 68% of students were female and 32 % male students. In section two, 70% of students were female and 30% were male students. In section two, 64% percent of students were Caucasian and 36% African/American students. In section two, 67% percent of students were Caucasian, 30% African/American, and 3% Native/American.

Small differences between two sections were noted in class classification and major composition. Out of the 50 students in section one who participated, 30% of students were sophomores, 30% seniors, 24% juniors, and 16% freshmen. Out of the 30 students in section two, 43% of students were juniors, 30% sophomores, 13% freshmen, 10% seniors, and three percent were undeclared. Apparently, in section one there was a

much higher percentage of seniors than in section two, but the percentage of freshmen was very similar in both sections.

There was a small difference in major composition between two sections. In section one, 30% of students were kinesiology majors, 24% education majors, 12% engineering majors, 10% biological majors, and 10% business majors. In section two, 30% of students were education major, 17% kinesiology majors, 7% engineering majors, 10% biological science majors, and 7% business majors. Perhaps because of the higher percentage of kinesiology and engineering majors in section one, the average self-reported ACT score was a little higher in section one (M = 23.02, SD = 3.88, n = 44) than in section two (M = 21.52, SD = 4.08, n = 27). The range of ACT self-reported scores from 16 to 30 was identical in both sections. The difference between self-reported grade point average of two sections was only 0.18. In section one, the average self-reported grade point average was 3.16 (M = 3.16, SD = .58, n = 45) with a range from 1.74 to 4.00. In section two, the average self-reported grade point average was 2.98 (M = 2.98, SD = .62, n = 27) with a range from 1.40 to 4.00. Therefore, all mentioned characteristics of participants in each section lead to conclusion that the variation of two sections was very small, and combining the two sections could not significantly influence the results of the study.

Procedure

Participants were administered five multiple choice exams throughout the semester, each exam consisting of 75 items. The test questions were taken from the test bank that accompanies the textbook, *The Developing Person through the Life Span*, 6th

edition, by Berger (2005). No technical information was provided about the test bank items; however, the instructors purposively chose the items. The lectures in both sections were not exactly the same, and the instructors in both sections did not necessarily use the same notes; however, students in both sections used the same study guides. Tests in both course sections were identical with exception of three items on exam 1 and one item on exam 5. These items were the result of differences between notes and lectures of two instructors. Tests in course section one had three forms. Tests in course section two had one form on exam 1 and two forms on exams 2-5. Forms of multiple choice tests had the same questions, but the order of alternatives was different.

Before exam 1, students were asked to participate in the study and to sign the consent form. Students were given oral instructions on how to omit questions from scoring with examples of scoring method and calculations of an adjusted score. Additionally, students were given written instructions on how to omit questions from being scored on the test (Appendix A). No incentives were given to students for their participation in the study because of the assumption that students would increase their score on exams by applying a question omitting procedure. In section one, only students who signed the informed consent were allowed to participate in the study. In section two, regardless of their choice to participate in the study, students were allowed to follow the procedure on all exams.

Before taking exam 1, students filled out a demographic form (Appendix B) and answered items on two subscales of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1991). The first subscale which students answered was

Expectancy Component: Self-efficacy for Learning and Performance (Appendix C), and the second subscale was Affective Component: Test Anxiety (Appendix D). Students answered one additional question constructed by the researcher (Appendix E).

After completing each of the five course exams, students marked on the back of their answer sheet up to five questions they wanted to be excluded from scoring. Calculation of this adjusted score was performed by the researcher. Tests were returned to students the next class period following the exam day. Correct answers were discussed by instructors in both sections. Each student was able to see the original score, the adjusted score, and how many of the omitted questions had been answered correctly or incorrectly.

After the fifth and last exam students answered a questionnaire that addressed their perception of the scoring method, its impact on their performance, and any changes in study or test-taking behaviors (Appendix F).

Instruments

The Motivated Strategies for Learning Questionnaire (MSLQ) has two large sets of scales: Motivational Scale (31 items) and Learning Strategies Scales (items 32-81) which were developed for college students. Both subscales, which were administered in the current study, the Expectancy Component: Self-Efficacy for Learning and Performance (8 items, $\alpha = .93$) and the Affective Component: Test Anxiety (5 items, $\alpha = .80$) belong to the Motivational Scale. These are all Likert type scales with scores ranging from 1 (not at all true for me) to 7 (very true of me) with item scores averaged for the scale score.

The Expectancy Component: Self-Efficacy for Learning and Performance

The Expectancy Component: Self-Efficacy for Learning and Performance subscale assesses two aspects of expectancy: expectancy for success which refers to performance expectations and self-efficacy which refers to "a self-appraisal of one's ability to master a task" (Pintrich et al., 1991, p. 13). An example of an expectancy for success item is: "I believe I will receive an excellent grade in this class." An example of a self-efficacy item is: "I am certain I can master the skills being taught in this class."

The Affective Component: Test Anxiety

The Affective Component: Test Anxiety consists of items which assess a cognitive component of test anxiety and an emotionality component. An example of an item that assesses a cognitive component is: "When I take a test I think about how poorly I am doing compared with other students." An example of an item that assesses an emotionality component is: "I feel my heart beating fast when I take an exam" (Pintrich et al., p. 13).

Additional Question

The students answered one additional question constructed by the researcher: "How good a test-taker do you think you are compared to others?" The response scale was a Likert type scale from 1-5 (not good at all, poor, good, very good, excellent). This question provided information about students' self-efficacy beliefs specific to their test taking skills. In the current study, it was examined in relation to students' likelihood to

apply an option to omit questions from scoring and to the degree of success students have in applying the omission procedure.

Questionnaire Given to Students after Exam 5

The students answered a questionnaire after exam 5 which assessed students' opinion about the question omitting procedure and its effect on their study and test taking strategies (Appendix D). The questionnaire had 12 items. Items 1, 2, 3, 4, 5, 7, and 8 used a Likert-type scale with the response options of "strongly disagree", "disagree", "agree", and "strongly agree". Item 6 assessed the change in amount of study time that the question omitting procedure might cause. Items 9, 10, and 11 included descriptive answers about test taking strategies, study strategies, and strategies that students used to decide which questions to omit from scoring. On item 12 students were asked to report how many hours they studied for the class. The estimated internal consistency reliability coefficient for the first five items was $\alpha = .82$.

Scoring Method

The decision to allow students to exclude up to five questions from scoring on an exam was mutually agreed to by both section instructors. Omitting up to five questions from a total number of 75 questions of the test would represent only a small percentage of the test (7%). By allowing students to omit up to five questions on each exam, students still had flexibility to choose how many questions they wanted to omit on each exam without the process having undue influence on the total score.

After tests were returned from the university testing services, the researcher compared the selected questions with the exam's answer key. If a student's choice of a question to be omitted from scoring happened to have been one that was answered correctly, then the number-right score was reduced by one from the raw score originally posted by testing services, as was the number of attempted items. If the question to be omitted had been answered incorrectly, then the number-right score was not adjusted, though the number of items attempted ("total") was reduced. A simple formula was applied in calculation. The new number of correct answers was divided by the new number of questions to be scored on exam for each student. For a better illustration of the scoring method, several examples of calculations are provided in the following paragraphs.

The general formula for adjusted score is:

100 x revised number correct/revised number attempted; where the revised number correct is the number of correct answers on all non-omitted items, and the revised number attempted is the count of all non-omitted items.

Example:

A student correctly answers 60 out of 75 items on a test. The unadjusted score, expressed as a percent, would be 80% [100 x 60/75]. If that student had identified five (5) items to be omitted from scoring, his or her adjusted score could be as high as 86% [100 x 60/70] if all of five omitted items had been incorrectly answered, or as low as 79% [100 x 55/70] if all five omitted items had been correctly answered.

If the student had omitted fewer than five questions from scoring, then the denominator (the missed number of items attempted) would be larger. In this example, it might be 71, 72, 73, 74, or 75, corresponding to the choice to omit 4, 3, 2, 1, or no items, respectively.

The researcher classified each question according to the level of knowledge (factual or application) that the question required from students. Another instructor, independently from the researcher, randomly chose 15 questions from each exam and judged whether the chosen questions required factual or application knowledge. Factual knowledge was defined as the type of knowledge that requires memorization such as names, terms, definitions, etc. Application knowledge was defined as an ability to apply the knowledge and solve problems in different situations. The agreement between the researcher and the instructor was 82% on 75 questions. The researcher and the instructor classified 82% of the 75 randomly selected questions identically. The agreement was highest on the 15 questions from exam 1 (100%). Therefore, inter-rater agreement was deemed satisfactory.

An example of factual knowledge question was: (exam 3)

The Piagetian term for centration in which a child thinks about the world exclusively from his or her personal perspective is called:

A) theory-theory

B) egocentrism

C) static perspective

D) world view

An example of application question was:

When Jennie sees her third-grade teacher in the grocery store, she does not recognize her. This is likely due to Jennie's:

A) static reasoning

B) abstract reasoning

C) concrete thinking

D) irreversibility

Questions were also identified (though not on the test as administered) by the chapter they covered. The researcher calculated the frequency of omission by students for each question. The results would show how the question omitting procedure affected content validity based on: a) chapter coverage, and b) cognitive level.

Statistical Analysis

Research Question 1: How will an option to omit questions from scoring affect students' scores on five exams?

The differences between scores before and after applying the score adjustment for omitted questions were calculated for each participant on each exam. The five difference scores were used as dependent variables that were analyzed via a single-sample multivariate analysis of variance (MANOVA).

Research Question 2: Do students improve their ability to distinguish between their correct and incorrect answers over five consecutive exams by using an option to omit questions from scoring?

A percentage of the number of incorrect answers relative to the total number of omitted questions was calculated for each person on each exam. The possible change of these percentages over five consecutive exams was assessed by repeated measures ANOVA. Tests of trend (linear, quadratic, cubic) were run to describe the nature of the change (if any). Students who did not use the omission option on all five exams were not included in this analysis.

Research Question 3: Will the frequency of items that students omit from scoring be correlated with item difficulty values?

Frequencies of omissions for each question were recorded. The item difficulty value for each question was used from the item analysis report. For any instances of mis--keyed questions, item difficulty values were calculated by the researcher. Pearson correlations were calculated between the frequencies and item difficulties.

Research Question 4: How does an option to omit questions from scoring affect content validity of the test?

Questions on each exam were classified in two groups based on the cognitive level (factual and application) and in groups based on the chapter they covered. The dependent variable was the frequency of omission from scoring for each question. An independent *t*-test was conducted to assess the difference in omission rate by cognitive level. One-way ANOVA was conducted to assess the differences among the groups of items based on their chapter coverage. These two analyses assessed the impact of the self-tailoring procedure on the potential content validity of the test

Research Question 5: How do students describe their strategies of omitting questions from scoring?

A questionnaire that students answered at the end of the study addressed this question (Appendix F). Analysis included descriptive statistics and categorizing of responses in which students described their strategies to omit questions from scoring.

Research Question 6: Will students' self-efficacy expectations be related to their likelihood to apply an option to omit questions from scoring and to the degree of success students have in applying the omission procedure?

Pearson correlations were calculated between scores on the self-efficacy scale (which measure expectancies for success and an appraisal of one's ability to master a task) and: a) the number of omissions made across the five exams, and b) the percentages of omissions that were incorrectly answered by the student.

Research Questions 7: Will test anxiety be related to the students' likelihood to omit questions from scoring and their success in applying the omission procedure?

Pearson correlations were calculated between scores on the test anxiety scale (which measures a cognitive and emotionality component of test anxiety) and: a) the number of omissions made across the five exams, and b) the percentages of omissions that were incorrectly answered by the student.

CHAPTER IV

RESULTS

This chapter presents the results of data analyses. The results are presented as statistically significant if $p < .05$. Data analysis for each research question follows with the answer to that question.

Research Question 1

Multivariate Analysis of Variance (MANOVA) was conducted to answer how the option to omit questions from scoring affected students' scores on five exams. Results of this analysis are presented in the following section.

Descriptive statistics of students' performance on all five exams before the question omitting procedure was applied are presented in Table 4.1. The means of exam performances reported as the percent of correct answers were fairly consistent across exams. Only about five percentage points separated the lowest mean performance, observed on exam 2 (76%) from the highest mean performance observed on exam 5 (81%). Distributions were slightly negatively skewed on all exams. Compared to previous classes, exam performances of this group of students were above the average. The analysis included students who took all five exams during the semester ($N = 80$).

Table 4.1

Descriptive Statistics for Exams before Applying the Question Omitting Procedure ($N = 80$)

	M (% correct)	*SD*	*Minimum* (% correct)	*Maximum* (% correct)
Exam 1	80.51	10.48	51	99
Exam 2	76.44	11.16	45	97
Exam 3	77.84	10.89	47	96
Exam 4	79.13	11.95	44	97
Exam 5	81.09	11.02	39	99

Multivariate Analysis of Variance (MANOVA) revealed that there was a statistically significant difference among exam scores before and after application of the question omitting procedure on the set of five exams $F(5, 75) = 74.93$, $\eta^2 = .83$, $p < .01$. Both unadjusted (before applying the question omitting procedure) and adjusted (after applying the procedure) exam scores were rounded to the nearest integer percent for analysis purposes. Descriptive statistics for score differences on all exams are presented in Table 4.2. The mean of the difference scores was lowest on exam 1 ($M = 2.39$) and highest on exam 5 ($M = 3.15$). In general, the means of the difference scores were close in value but tended to increase across the five exams. The analysis included students who took all exams in the semester ($N = 80$). Because not all students taking an exam elected to use the question omitting procedure, the differences reported in Table 4.2 are likely to be slight underestimates of the amount of score improvement realized by students who did apply the procedure.

Table 4.2

Descriptive Statistics for Score Differences between Adjusted and Unadjusted Scores on All Exams ($N = 80$)

Adjusted – Unadjusted Score	*M*	*SD*
Score Difference for Exam 1	2.39	1.82
Score Difference for Exam 2	2.93	1.71
Score Difference for Exam 3	2.51	2.01
Score Difference for Exam 4	3.03	1.95
Score Difference for Exam 5	3.15	1.88

Note. On each exam, the adjusted scores were significantly higher than unadjusted scores.

The univariate tests showed that score differences before and after question omitting procedure were statistically significantly different on all exams, $p < .01$. All the effect sizes were of the same magnitude. On average, students were able to increase their scores by approximately a quarter of a standard deviation (Table 4.3). The results of univariate tests for score differences between adjusted and unadjusted scores on all exams are presented in Table 4.3.

Table 4.3

Univariate Tests for Score Differences between Adjusted and Unadjusted Scores on All Exams ($N = 80$)

Adjusted – Unadjusted Score	$F(1,79)$	η^2	d
Difference Score for Exam 1	138.03 **	.64	.23
Difference Score for Exam 2	233.52 **	.75	.25
Difference Score for Exam 3	125.46 **	.61	.23
Difference Score for Exam 4	192.81 **	.71	.25
Difference Score for Exam 5	223.81 **	.74	.29

**$p < .01$

As indicated by the data analysis results, the option to omit questions from scoring significantly increased average scores on all five exams. On exam 1, 83.7% of students increased their score by using the option to omit questions from scoring. On exam 2, 92.5 % of students increased their score. On exam 3, 82.4% of students increased their score. On exam 4, 88.7 % increased their score. On exam 5, 93.7% increased their score. These figures are based on all students taking the exams ($N = 80$) and therefore, include some who did not apply the question omitting procedure. In no instance did more than 3.8% of the students see a decline in their score as a result of the question omitting procedure.

Research Question 2

Repeated measures ANOVA was conducted to answer whether students improved their ability to distinguish between their correct and incorrect answers over five consecutive exams when using the option to omit questions from scoring. The results of this analysis are presented in the following section. The analysis included students who omitted questions on all exams ($N = 71$). The DV was the percent of incorrect items out of the total number of items omitted, and it was recorded for each of the five exams.

The assumption of sphericity was not met ($p = .04$). The Huynh-Feldt adjustment for violation of sphericity was used to report the results of the analysis. Repeated measures ANOVA showed that the main effect of occasion for percentages of incorrect answers out of the total number of answers that students omitted on five exams was statistically significant $F(3.7, 261) = 3.89$, $MSE = 502.95$, $\eta^2 = .05$, $p = .005$. Descriptive statistics for percentages of incorrectly answered questions out of the total number omitted on five exams are presented in Table 4.4. The mean of percentages of incorrectly answered questions out of the total number omitted was lowest on exam 1 ($M = 59.60$) and highest on exam 5 ($M = 71.92$). From the first exam, in which about 60% of omitted questions were incorrect, on exams 2-5 about 70% of omitted questions were incorrect, on the average.

Table 4.4

Descriptive Statistics for Percentages of Incorrectly Answered Questions out of the Total Number Omitted ($N = 71$)

Incorrectly Answered Questions/ Total Number Omitted	*M*	*SD*
Percentage on Exam 1	59.60	26.59
Percentage on Exam 2	70.82	23.77
Percentage on Exam 3	65.61	29.80
Percentage on Exam 4	69.98	27.82
Percentage on Exam 5	71.92	26.30

The follow-up tests using the Least Significant Differences (LSD) criterion at $p < .05$ showed that percentages of incorrectly answered questions out of the total number omitted on exam 2 ($M = 70.82$, $SD = 23.77$, $n = 71$), exam 4 ($M = 69.98$, $SD = 27.82$, $n = 71$) and exam 5 ($M = 71.92$, $SD = 26.03$, $n = 71$) were statistically significantly higher than on exam 1 ($M = 59.60$, $SD = 26.59$, $n = 71$). There was a statistically significant difference between exam 3 ($M = 69.98$, $SD = 27.82$, $n = 71$), and exam 5. The test of trend showed a statistically significant linear increase of percentages over five exams $F(1,70) = 6.74$, $MSE = 596.98$, $\eta^2 = .09$, $p = .01$). Higher order trends (quadratic, cubic and order 4) were not statistically significant. Compared to exam 1, students exhibited the highest increase of percentages of incorrectly answered questions out of the total number omitted on exam 5. However, students made a statistically significant increase on exam 2 compared to exam 1. Increases on exams 4 and 5 were statistically significant in comparison to exam 1 but not to exam 2.

The results showed that students omitted an increasingly higher number of incorrectly answered questions out of the total number omitted over five successive exams, and this increase was linear in form. Therefore, students' metacognitive ability, which was operationally defined in the current study as the percentage of incorrectly answered questions out of the total number omitted, improved on consecutive exams for those students who omitted questions on all exams.

As indicated by the data analysis results, students improved their ability to distinguish between their correct and incorrect answers over five consecutive exams by using the option to omit questions from scoring. Therefore, research question 2 was supported in the affirmative.

Research Question 3

Pearson correlations were conducted to determine whether the frequency of omissions from scoring were correlated with item difficulty values. Item difficulty values were reported by the university testing services as proportions of students who answered each question correctly in a particular group of students. Lower values indicated more difficult questions; higher values indicated easier questions. Frequencies of item omissions were calculated by counting the number of students who omitted each item.

Descriptive statistics for item difficulty values and frequencies of item omissions are presented in Table 4.5. The mean item difficulty values were high on all exams, with about 76% of students answering items correctly, on average. The lowest average item difficulty value was found on exam 2 (.73) which was also the most difficult exam in the semester. The highest average item difficulty value was found on exam 5 (.79) which was

also the easiest exam. The consistency in mean item difficulty across the five exams mirrors that of the average scores, discussed earlier (Table 4.1). The average omissions per item were similar on all exams. Each item was approximately omitted 4.5 times on all exams except on exam 3 (4.31). However, some questions were not omitted at all; others were omitted substantially more often, as much as 26 times.

Table 4.5

Descriptive Statistics for Item Difficulty Values and Frequencies of Item Omissions on All Exams (k = 75)

	Exam	*M*	*SD*
Item Difficulty Value	1	.78	.14
Item Difficulty Value	2	.73	.17
Item Difficulty Value	3	.75	.18
Item Difficulty Value	4	.77	.13
Item Difficulty Value	5	.79	.16
Frequency of Item Omissions	1	4.55	4.19
Frequency of Item Omissions	2	4.45	4.63
Frequency of Item Omissions	3	4.31	3.76
Frequency of Item Omissions	4	4.48	3.75
Frequency of Item Omissions	5	4.55	4.36

There were negative correlations between item difficulty values and frequencies of omissions on all exams $r1$ = -.69, $r2$ = -.48, $r3$ = -.52, $r4$ = -.67, $r5$ = -.73. All correlations were statistically significant, $p < .01$. When given the option to omit

questions from scoring, students more frequently omitted more difficult questions which were indicated by the lower item difficulty values. The analysis included students who took all exams during the semester ($N = 80$).

As indicated by the data analysis results, frequency of items that students omitted from scoring were significantly negatively correlated with item difficulty values. Therefore, research question 3 was supported in the affirmative.

Research Question 4

Independent *t*-tests and one-way ANOVA were conducted to answer how the option to omit questions from scoring affected the content validity of the test. An independent *t*-test was conducted for each exam to determine whether the content validity of the test, based on cognitive level of items, was affected by the option to omit questions from scoring. One-way ANOVA was conducted to answer whether content validity of the test, which was based on textbook chapters, was affected by the option to omit questions from scoring.

Descriptive statistics for frequencies of item omission in the two groups of items based on their cognitive level (factual knowledge or application) on each exam are presented in Table 4.6. The mean frequency of item omissions was highest in the application group on exam 2 ($M = 6.87$, $SD = 4.90$) and lowest in the application group on exam 3 ($M = 2.27$, $SD = 2.98$). The total number of questions was the same on each exam ($k = 75$), but the number of questions in each group differed across exams. The frequencies of omissions were recorded from students who took all exams in the semester ($N = 80$).

Table 4.6

Descriptive Statistics for Frequencies of Item Omissions in the Two Groups (Factual Knowledge and Application) on All Exams

Group of Questions	Exam	*k*	*M*	*SD*
Factual Knowledge	1	62	4.48	4.07
Application	1	13	4.85	4.88
Factual Knowledge	2	60	3.85	4.40
Application	2	15	6.87	4.90
Factual Knowledge	3	53	4.94	3.89
Application	3	22	2.77	2.98
Factual Knowledge	4	56	4.25	3.74
Application	4	19	5.16	3.79
Factual Knowledge	5	68	4.53	4.33
Application	5	7	4.71	4.92

An independent *t*-test was conducted for each exam to assess the difference between frequencies of item omissions based on their two cognitive levels, factual knowledge or application. The results of independent *t*-tests for each exam are presented in Table 4.7.

Table 4.7

Independent *t*-tests for Frequencies of Item Omissions in the Two Groups (Factual Knowledge and Application) on Each Exam ($k = 75$)

Exam	$t(73)$	p
Exam 1	-.28	.78
Exam 2	-2.32	.02
Exam 3	2.35	.02
Exam 4	-.91	.37
Exam 5	-.11	.92

Assumptions of equality of variances were met on independent *t*-tests for all exams. On exam 2, the mean frequency of item omissions was statistically significantly higher in the application group ($M = 6.87$, $SD = 4.90$, $n = 15$) than in the factual knowledge group of items ($M = 3.85$, $SD = 4.40$, $n = 60$), $t(73) = -.2.32$, $p = .02$. On exam 3, the mean frequency of omissions was statistically significantly higher in the factual knowledge group ($M = 4.94$, $SD = 3.89$, $n = 53$) than in the application group of items ($M = 2.77$, $SD = 2.98$, $n = 22$), $t(73) = 2.45$, $p = .02$. On exam 2, students omitted application questions at a higher rate than factual knowledge questions. On exam 3, students omitted more frequently factual knowledge questions than application questions. On exams 1, 4, and 5 frequencies of item omissions were not statistically significantly different between factual knowledge and application groups of items ($p > .05$). Therefore, on exams 1, 4, and 5 students did not omit more frequently questions from any of the two groups.

As indicated by the data analysis results, the content validity of the test based on cognitive level of items was affected on exam 2 and exam 3 but not on exams 1, 4, and 5 by the option to omit questions from scoring. Therefore, this answers the first part of the research question. The content validity of the test based on cognitive level of items was affected on two out of five exams in the semester by the option to omit questions from scoring.

One-way ANOVA was conducted to answer the second part of the research question, whether content validity of the test based on textbook chapters was affected by the option to omit questions from scoring. Exam 1 included questions from four chapters. Exams 2-4 each included questions from three chapters, and exam 5 included questions from three chapters, and four additional questions that belonged to different chapters from the textbook.

One-way ANOVA for exam 1 revealed that there was a statistically significant difference among frequencies of item omissions based on the four chapters that exam 1 covered $F(3,71) = 2.87$, $MSE = 16.29$, $p = .04$. The follow-up Bonferroni test showed that frequencies of item omissions from chapter two were statistically significantly higher ($M = 6.14$, $SD = 5.26$, $n = 21$) than from chapter four ($M = 2.50$, $SD = 2.98$, $n = 18$), $p = .04$. Chapter two covered different development theories, such as, psychoanalytic, behaviorism, cognitive, sociocultural, and epigenetic. Chapter four covered the area of prenatal development and birth. Students omitted more frequently questions that covered chapter two than chapter four. Therefore, the content validity of the test based on chapters was affected by the option to omit questions from scoring on exam 1.

Descriptive statistics for frequencies of item omissions across four chapters on exam 1 are presented in Table 4.8. The mean frequency of item omissions was highest on chapter two (M = 6.14, SD = 5.26) and lowest on chapter four (M = 2.50, SD = 2.98).

Table 4.8

Descriptive Statistics for Frequencies of Item Omissions by Chapter on Exam 1

	k	*M*	*SD*
Chapter 1	17	4.06	3.07
Chapter 2	21	6.14	5.26
Chapter 3	19	5.16	4.01
Chapter 4	18	2.50	2.98
Total	75	4.55	4.19

One-way ANOVA for exam 2 revealed that there was no statistically significant difference among frequencies of item omissions based on the three chapters that exam 2 covered $F(2,72) = .70$, $MSE = 21.64$, $p = .50$. Students did not omit significantly more questions from any of the chapters on exam 2. Therefore, the content validity of the test based on chapters was not affected by the option to omit questions from scoring on exam 2.

Descriptive statistics for frequencies of item omissions on exam 2 are presented in Table 4.9. Even though the mean on chapter six was the highest (M = 5.24, SD = 4.47) and on chapter seven the lowest (M =3.68, SD = 3.84), there was not much difference among the means for frequency of item omissions on exam 2.

Table 4.9

Descriptive Statistics for Frequencies of Item Omissions by Chapter on Exam 2

	k	*M*	*SD*
Chapter 5	25	4.44	5.49
Chapter 6	25	5.24	4.47
Chapter 7	25	3.68	3.84
Total	75	4.45	4.63

One-way ANOVA for exam 3 revealed that there was a statistically significant difference among frequencies of item omissions based on the three chapters that exam 3 covered $F(2,72) = 3.68$, $MSE = 13.18$, $p = .03$. Follow-up Bonferroni test showed that frequencies of item omission from chapter nine were statistically significantly higher ($M = 5.84$, $SD = 4.67$, $n = 25$) than from chapter ten ($M = 3.12$, $SD = 2.83$, $n = 25$, $p = .03$). Chapter nine covered cognitive development of children from ages two to six, and chapter ten covered psychosocial development of children within the same age range. Students omitted more frequently questions from chapter nine which covered cognitive development of children than from chapter ten. Therefore, the content validity of the test based on chapters was affected by the option to omit questions from scoring on exam 3. Descriptive statistics for frequencies of item omissions on exam 3 are presented in Table 4.10. The mean frequency of item omissions was highest on chapter nine ($M = 5.84$, $SD = 4.67$) and lowest on chapter ten ($M = 3.12$, $SD = 2.83$), and this difference was for almost three points (2.72)

Table 4.10

Descriptive Statistics for Frequencies of Item Omissions by Chapter on Exam 3

	k	*M*	*SD*
Chapter 8	25	3.96	3.12
Chapter 9	25	5.84	4.67
Chapter 10	25	3.12	2.83
Total	75	4.31	3.76

One-way ANOVA for exam 4 revealed that there was no statistically significant difference among frequencies of item omissions from any of the three chapters that exam 4 covered $F(2,72) = .23$, $MSE = 14.34$, $p = .80$. Students did not omit significantly more items from any of the chapters on exam 4. Therefore, the content validity of the test based on chapters was not affected by the option to omit questions from scoring on exam 4. Descriptive statistics for frequencies of item omissions on exam 4 are presented in Table 4.11. There was not much difference among means frequencies of item omissions on any of the chapters covered by exam 4.

Table 4.11

Descriptive Statistics for Frequencies of Item Omissions by Chapter on Exam 4

	k	*M*	*SD*
Chapter 11	25	4.84	3.51
Chapter 12	25	4.48	3.98
Chapter 13	25	4.12	3.85
Total	75	4.48	3.75

One-way ANOVA for exam 5 revealed that there was a statistically significant difference among frequencies of item omissions based on the chapters that exam 5 covered $F(3,71) = 2.80$, $MSE = 17.69$, $p = .046$. However, the follow-up Bonferroni test did not show any statistically significant difference in frequencies of item omissions across the three chapters and four questions on adulthood. While the overall test was statistically significant, none of the pairwise chapter comparisons yielded statistically significant difference. Therefore, the content validity of the test based on chapters was affected by the option to omit questions from scoring on exam 5, though no specific pairwise differences can be identified. Descriptive statistics for frequencies of item omissions on exam 5 are presented in Table 4.12. The mean frequency of item omissions was the highest for the adulthood set ($M = 7.75$, $SD = 1.5$). However, there were only four questions on adulthood in the whole exam. The means frequency of item omissions were similar on Chapter 14 ($M = 3.76$, $SD = 3.00$) and Chapter 16 ($M = 3.23$, $SD = 4.17$).

Table 4.12

Descriptive Statistics for Frequencies of Item Omissions by Chapter on Exam 5

	k	*M*	*SD*
Chapter 14	25	3.76	3.00
Chapter 15	24	6.04	5.39
Chapter 16	22	3.23	4.17
Adulthood	4	7.75	1.50
Total	75	4.55	4.36

As indicated by the data analysis results, content validity of the test based on chapters covered on exams was affected on exams 1, 3, and 5 but not on exams 2 and 4 by the option to omit questions from scoring. Therefore, this answers the second part of the research question. The content validity of the test based on chapters covered on exams was affected on three out of five exams in the semester by the option to omit questions from scoring.

Research Question 5

Descriptive statistics of students' responses on questions 1-8 of the questionnaire given to students after exam 5 (Appendix F) was conducted to answer how students described their strategies of omitting questions from scoring. The analysis included students who took all five exams during the semester ($N = 80$). Additionally,

selected students' responses on questions 9-11 will be presented. Percentages of students' answers (strongly disagree, disagree, agree, and strongly agree) on questions 1-8 are presented in Table 4.13.

Table 4.13

Percent of Strongly Disagree (SD), Disagree (D), Agree (A) and Strongly Agree (SA) Answers on Questions 1-8

Question	SD	D	A	SA
1. The opportunity to exclude up to 5 questions helped me to improve score on my exams.		4%	56%	40%
2. I would recommend that this procedure be included on other multiple choice exams.		5%	38%	58%
3. I became more confident in distinguishing what I know from what I do not know.	1%	18%	55%	26%
4. I improved my selection of wrong answers to be omitted on successive exams.		14%	60%	26%
5. This optional procedure gave me a sense of control over my exams.		11%	60%	29%
6. This procedure caused me to A) study more for my exams 21% B) study less for my exams 4% C) not change my study time 75%				
7. The procedure made me aware of strategy changes that I needed to make when I was taking the test.		23%	56%	21%
8. The procedure made me aware of necessary changes that I need to make when I study for exams.	1%	40%	50%	9%

Note: Percentages may not sum to 100 due to rounding.

Students expressed a favorable opinion toward the option to omit questions from scoring. Ninety-six percent of students agreed that the option to omit questions from scoring helped them to improve their score on exams. This report was consistent with the percentage of students who actually increased their scores on each exam (83.7%, 92.5%, 82.4%, 88.7%, and 93.7%). However, the students' answers to question 1 included their opinion about score increase on all exams and not on any particular exam. Ninety-six percent of students would recommend this option to be included on other multiple choice exams. Eighty-nine percent of students reported having more control over their exams by using the option to omit questions from scoring. Eighty-one percent of students reported that the option to omit questions from scoring helped them to become more confident to distinguish between what they knew and what they did not know on exams. Additionally, 60% of students agreed, and 26% of students strongly agreed that they improved their selection of wrong answers on consecutive exams. These answers supported the two previous findings in the current study that were presented under research question 1 and research question 2: a) students improved their scores by using the option to omit questions from scoring, and b) students' metacognitive ability to distinguish between correct and incorrect answers improved on consecutive exams.

According to students' answers to question 12 (Appendix F), they studied on average 3.5 hours per week for the class, with a minimum of 1 hour and a maximum of 12 hours. Seventy-five percent of students self-reported that the option to omit questions from scoring did not change their study time for exams. However, fifty-nine percent of students reported that the option to omit questions from scoring made them aware of

changes in study strategies they needed to make, but only a small number of students (10) described study strategy changes. Seventy-seven percent of students reported that the option to omit questions from scoring made them aware of necessary changes of test taking strategies. This was also supported by students' descriptive answers on questions 9-11. Apparently, the option to omit questions from scoring affected more the students' test taking strategies than their study strategies.

Questions 9-11 required short descriptions. These questions and several selected answers will be presented in the following section.

Question 9: If you made any changes in test-taking strategies, please describe these changes.

Question 10: If you made any changes in your study strategies, please describe these changes.

Question 11: In a few sentences describe the strategy that you used to decide which question to omit from scoring.

Selected answers are presented within three categories based on similarities of students' answers on question 11. The first category of answers has in common a description of the question omitting procedure in which a student marks questions that he or she is not sure about and comes back to these questions later. If a student answered questions 9 and 10, these answers are also presented. A note is added indicating whether an answer was provided by a high, an average, or a low achieving student. A high achieving student was one having a mean unadjusted exam score higher than 80%. An

average student had a mean unadjusted exam score between 70% - 80%, whereas a low achieving student had an unadjusted exam average of less than 70%.

"I took the test putting question marks by questions I was not sure about. I then narrowed these questions down to five and omitted the final five that I was less sure of" (an answer to question 11 by a high achieving student).

"By using this procedure I changed my test-taking strategies to be sure to mark questions that I was unsure of while taking the test and comparing them to other questions to see which question was most appropriate to eliminate" (question 9). Analyzing the questions I was unsure of-and eliminating the five questions I was positive that I missed" (question 11, an average student).

"I marked each answer I wasn't sure about so I could omit those later" (question 9). "I omitted the ones I was least sure about" (question 11, a high achieving student).

"Skipped over ones I wasn't sure of and come back to them later" (question 9). "Studied more so I knew the material better" (question 10). "The ones I was least confident about" (question 11, an average student).

The second category of answers to question 11 has in common the specification of a confidence level during the question omission procedure. If a student answered questions 9 and 10, these answers are also presented.

"Knowing when to omit or not to, when to use all 5 or when to only use 1 or 2 omissions" (question 9). "The ones I had no clue on I would omit, if I was 95% sure I left it to scoring" (question 11, a high achieving student).

“I omitted only the questions I did not remember from studying if I recognized it, but was less than 50% sure-I omitted” (question 11, a high achieving student).

“Always aware that if not sure about an answer could omit” (question 9). “If I could not recall the answer from my studies with 80% of being sure would definitely omit” (question 11, a high achieving student).

The third category of answers to question 11 has in common specifically mentioning items that would require guessing when applying the question omitting procedure.

“Instead of guessing at questions I do not know I would come back later and possibly omit them” (question 9). “The questions that I knew I couldn’t make an educated guess on” (question 11, a high achieving student).

“I chose the ones that I did not even think I knew the answer to. The ones that I guessed on” (question 11, a high achieving student).

“Go through the entire test before deciding the questions to omit” (question 9). “Ones I omitted were guesses and that had many other possible answers“ (question 11, a high achieving student).

Several other responses, which did not belong to any of the previously mentioned categories, are presented in the following section.

“The questions that I chose to not to be counted helped me single out minor, but important information for later tests” (question 9).”If there was any doubt, I would wait and then I would go back and read the questions again at the end. If the doubt was still there, I would choose that questions” (question 11, a low achieving student).

"I tried not to change my answers" (question 9). "The questions I thought about changing my answers" (question 11, a high achieving student).

"Go over it twice" (question 9). "Study two days in advance", (question 10)."The question that I really don't know" (question 11, a low achieving student).
"I would overanalyze and make myself really confused and think way too hard. That's just me" (question 11, a high achieving student).

In summary, though the verbal descriptions may have differed, most of the comments imply that students relied on confidence level or certainty of their knowledge when making a decision to omit a question from scoring. Students exhibited their metacognitive ability to distinguish between what they knew from what they did not know on exams by using the option to omit questions from scoring. According to students' reports, the question omitting procedure did not affect their study strategies as much as their test taking strategies. Study strategies changes were mentioned by only a small number of students, and these changes were not systemic in nature.

Research Question 6

Pearson correlations were conducted to answer whether students' self-efficacy expectations were related to their likelihood to apply the option to omit questions from scoring and to the degree of success students had in applying the omission procedure.

Correlations between self-efficacy scores and the total number of questions omitted on each exam were not statistically significantly different from zero, $r1 = .05$, $r2 = .03$, $r3 = .19$, $r4 = .01$, $r5 = .14$, $p > .05$. Students who had higher self-efficacy expectations did not omit higher number of questions on any of the exams than those

with lower self-efficacy. The analysis included students who took all exams in the semester ($N = 80$). For the current study, the estimated internal consistency reliability coefficient of the Self-Efficacy subscale was $\alpha = .91$. The average score on the Self - Efficacy subscale was 5.76 ($M = 5.76$, $SD = .75$, $Min = 3.75$, $Max = 7$). Since the items used a response scale from 1-7, where 7 corresponds to a statement being"very true of me", and the statements were positively framed, the self-efficacy mean was high.

Correlations between self-efficacy scores and percentages of incorrectly answered questions out of the total number of questions omitted were not statistically significant on any of the exams $r1 = -.11$, $r2 = .00$, $r3 = .09$, $r4 = .09$, $r5 = .00$, $p > .05$. Students who had higher self-efficacy expectations did not omit higher percentages of incorrectly answered questions out of the total number omitted compared to those with lower self-efficacy. Students' self-efficacy expectations were not related to students' metacognitive ability which was operationally defined in the current study as the percentage of incorrectly answered questions out of the total number of questions omitted. The analysis included students who omitted questions on all exams ($N = 71$).

As indicated by the data analysis results, students' self-efficacy expectations were not related to their likelihood to apply the option to omit questions from scoring or to the degree of success students had in applying the omission procedure. Therefore, research question 6 was not supported.

Research Question 7

Pearson correlations were conducted to answer whether test anxiety was related to students' likelihood to omit questions from scoring and to the degree of success students had in applying the omission procedure.

Correlations between test anxiety scores and the total number of questions omitted were not statistically significant on any of the exams, *r1* = -.09, *r2* = .03, *r3* = -.16, *r4* = .00, *r5* = -.02, $p > .05$. The analysis included students who took all exams in the semester ($N = 80$). For the current study, the estimated internal consistency reliability coefficient of the Affective Component: Test Anxiety subscale was, $\alpha = .81$. The average score was 3.76 ($M = 3.76$, $SD = 1.36$, $Min = 1$, $Max = 6.8$). This mean on the scale 1-7 falls in the middle of the score range, suggesting students were reporting a moderate level of test anxiety, on the average.

Correlations between test anxiety scores and percentages of incorrectly answered questions out of the total number of questions omitted were not statistically significant on any of the exams, *r1*= .14, *r2* = -.15, *r3* = -.03, *r4* = -.01, *r5* =.10, $p > .05$. Therefore, test anxiety was not related to students' metacognitive ability, which was operationally defined as the percentage of incorrectly answered questions out of the total number of questions omitted. The analysis included students who omitted questions on all exams ($N = 71$).

As indicated by the data analysis results, test anxiety was not related to students' likelihood to omit questions from scoring or to the degree of success students had in applying the omission procedure. Therefore, research question 7 was not supported.

Additional Analysis

Correlations between the scores on the scale 1-5 for the question, "How good a test-taker do you think you are compared to others?" and the total number of questions that students omitted were not statistically significant on any of the exams, *r1* = -.10, *r2* = -.04, *r3* = .11, *r4* = -.01, *r5* = .13, $p > .05$. This analysis included 76 students who took all exams during the semester because four students did not answer this question ($N = 76$). The average score was 3.29 ($M = 3.29$, $SD = .63$, $Min = 2$, $Max = 5$, $n = 76$). Based on the response scale (1 = not good at all, 5 = excellent; see Appendix E), this average would suggest that the typical student judged his/her test-taking skills as "good" in comparison with others.

Correlations between scores for the same question and percentages of incorrectly answered questions out of the total number of questions omitted were not statistically significant on any of the exams, *r1* = -.01, *r2* = .03, *r3* = .00, *r4* = .05, *r5* = -.10, $p > .05$. This final analysis included students who omitted questions on all exams during the semester ($N = 71$).

CHAPTER V

DISCUSSION

This chapter discusses the findings of the current study in relation to self-tailored tests, metacognitive ability, item difficulty, content validity of the test, test-taking strategies and study strategies, self-efficacy, and test anxiety. It concludes with limitations and implications of the study, and recommendations for future research.

Self-tailored Tests

On average, students increased their scores on all exams in the semester by using the option to omit up to five questions from scoring. Eighty examinees partially tailored at least one test by using this option. This finding was somewhat similar to the finding of Morse's study (1988). In Morse's study students selected items from a larger set of questions which would show best how well they learned the material covered on the test. In his study, only 19 out of 190 students did not improve their score by using this procedure. In the current study, students omitted up to five questions from each test. The remaining questions on the test were used to calculate the adjusted scores which were statistically significantly higher than unadjusted score on each exam. Eighty-four percent of students increased their score on exam 1. Ninety-three percent of students increased their score on exam 2. Eighty-two percent of students increased their score on exam 3.

Eighty-nine percent of students increased their score on exam 4. Ninety-four percent of students increased their score on exam 5.

No other studies have examined the score changes on multiple choice exams in similar context to the current study. Adaptive types of testing (paper-pencil or computerized) are difficult to apply in an ordinary classroom. Even alternative procedures such as self-tailored tests have not been well investigated. Self-tailoring in the current study was used as an option for students to omit questions from scoring. The benefits of this testing were reported by students on the questionnaire given to them after exam 5. Ninety-six percent of students reported that the procedure helped them to improve their score on exams. Eighty-nine percent of students reported that the procedure gave them a sense of control over exams. Ninety-six percent of students would recommend this procedure to be included on other multiple choice exams. However, instructors might be reluctant to apply the question omitting procedure in a large college classroom if they relied on a hand calculation of an adjusted score on exams. An alternative to this would be a calculation of an adjusted score by the school testing services, or a simple spreadsheet could be constructed to automate the process.

Metacognitive Ability

In addition to the increase of scores on exams, students made an improvement in the area of metacognition over the set of five examinations. The increase of adjusted scores compared to unadjusted scores was the result of omitting more incorrect than correct answers from scoring. The percentage of incorrectly answered questions out of the total number omitted, which was operationally defined as the measure of

metacognitive ability, increased linearly over five consecutive exams. By practicing the question omitting procedure on consecutive exams, students omitted a higher percentage of incorrect answers out of the total number omitted; therefore, students' metacognitive ability improved by practice. On exam 1, 18% of students omitted all incorrect answers out of the total number omitted. Twenty-five percent of students omitted all incorrect answers on exam 2, 30% on exam 3, 37% on exam 4, and 32% on exam 5. These findings were supported by 86% of students who reported that they improved their selection of wrong answers to be omitted on consecutive exams. Also, 81% of students reported that they became more confident in distinguishing what they knew from what they did not know because of this procedure.

The part of metacognition responsible for the question omitting procedure was metacognitive monitoring of performance on a test. Studies that investigated the effect of practice on metacognitive monitoring accuracy reported conflicting results (Bol & Hacker, 2001; Bol et al., 2005; Hacker et al., 2000; Nietfeld et al., 2005). Nietfeld et al. found no change of monitoring accuracy on four successive exams in a semester. Hacker et al. found that high achieving students improved their accuracy to predict their performance on a test after taking the test, but low performing students did not on three successive exams in a semester. Bol et al. examined the effect of overt practice on calibration across five quizzes on line and found no difference between performance on a final exam between students who practiced calibration, and those who did not. Calibration was defined as the ability of matching perceived performance with actual performance.

A factor that might affect monitoring accuracy is whether students received feedback or explicit training on monitoring of performance (Nietfeld et al., 2005). In the current study, students received oral instructions and written instructions on how to omit questions from scoring before the first exam (Appendix A). Students were instructed to omit questions that they believed they answered incorrectly, might have answered incorrectly, or were not certain of the correct answers. After tests were returned from the university testing services and adjusted scores were calculated by the researcher, students in both sections reviewed their answers on the test including the answers they omitted from scoring. Instructors in both sections discussed the correct answers on test questions. These practices might have had an influence on students learning how to apply the question omitting procedure and their increased success in applying the procedure on successive exams.

The highest increase in percent of incorrectly answered questions out of the total number omitted was observed between exam 1 and exam 2 (Table 4.4). The percent of incorrectly answered questions out of the total number omitted on exam 5 compared to exam 2 was only slightly higher but not significantly different. There was improvement on exams 3 and 4 compared to exam 1. Most likely, after the experience students had learned on the first exam and subsequent score feedback how to use the procedure more efficiently by exam 2, students continued to use the procedure successfully on other exams. However, nine students out of 80 students who took all exams did not omit questions on all exams. Most likely students who improved their scores and felt confident with the procedure continued to use it on all exams. With the exception of exam 3, the

mean improvement in scores due to self-tailoring option increased uniformly from exam 1 to exam 5 (Table 4.2). The positive effect of practice on predictive and postdictive accuracy of test performance was reported by Hacker et al., (2000). However, Hacker et al. found that high achieving students improved monitoring accuracy of a test performance, especially postdictive accuracy on three successive exams in semester, while low performing students did not improve their monitoring accuracy. High achieving students were the students who scored above 70% on their exams, and low performing students performed below 50% on their exams.

In addition to practice, completing all questions on an exam before omitting up to five questions from scoring might be a factor that contributed to students' accuracy in predicting whether the questions they omitted were incorrect. In the current study, students marked on a back of their answering sheet questions they believed they answered incorrectly after completing all questions on a multiple choice test. As cited by Ghatala et al., (1989, p. 51) and according to Flavell (1979) taking a test appears to be "a metacognitive experience." Several studies found a greater accuracy of a test performance prediction after test was taken than before test was taken (Pierce & Smith, 2001; Pressley et al., 1987). Pierce and Smith called this "the postdiction superiority effect" (p. 62).

Item Difficulty

Besides their metacognitive ability to omit a higher percentage of incorrect answers out of the total number omitted, students omitted more frequently more difficult questions. Pearson correlations between frequencies of item omissions and item difficulty

values were strong. These results were similar to the findings of Morse and Morse (2002). Morse and Morse found significant correlations between frequency of items that students chose as the five easiest and five most difficult items on the test and items' difficulty values. Results of both studies indicated that students had metacognitive awareness of easy and difficult questions on the test.

Students omitted with higher frequencies more difficult questions. However, Pressley and Ghatala (1988) found that students discriminated better between their correct and incorrect answers on easier questions. Similar findings were reported by several studies that investigated calibration or ability to match perceived performance with actual performance. These studies reported that students had better calibration on easier items than on more difficult items (Gigerenzer, et al., 1991; Lichtenstein & Fischhoff, 1977; Nietfeld et al., 2005; Schraw & Roedel, 1994). In the current study, students omitted questions they believed they had answered incorrectly, might have answered incorrectly or were not certain of the correct answer. Students were able to assess whether their answers were correct or incorrect on these more difficult items which resulted in an increase of mean scores on all exams.

Content Validity of the Test

In deciding whether to allow the option of omitting questions from scoring, the concern of the instructor might be whether the procedure affects the content validity of the test. The potential influence on content validity of the exams was observed in the current study using two bases: a) cognitive level of questions (factual or application), and b) chapters of the textbook that the exam covered.

Application questions are usually considered to be more difficult than factual knowledge questions. Therefore, it would be expected that students would omit with higher frequency application questions. In the current study, frequency of omissions of application items was statistically significantly higher than factual knowledge items only on exam 2. On exam 3, frequency of omissions of factual knowledge items was statistically significantly higher than for application items. Exam 2 was the most difficult exam in the semester as indicated by the students' performance average (M = 76.44, SD = 11.16). Exam 3 was the second most difficult exam in the semester as indicated by the students' performance average (M = 77.84, SD = 10.89). Exam 2 covered biosocial, cognitive, and psychosocial development of infants. Exam 3 covered the same areas of development of children between ages two and six. Most likely students found application items from development of infants more difficult than factual knowledge items. The opposite happened on exam 3 which covered development of children between ages two and six.

Several reasons might be responsible for these findings. Exam 2 was the first exam that assessed development of a certain age group. Application items of development of infants might be more difficult to students than application items from the other age groups. Exam 3 covered much vocabulary related to development of preschool children. Most likely students found these factual knowledge items on exam 3 more difficult than application items. On each exam in the semester, there were far fewer application items than factual knowledge items, which could have affected the results.

The content validity of the test based on chapters was affected on exams 1, 3, and 5. On exam 1, the frequency of item omissions from the chapter that covered developmental theories (chapter 2) was statistically significantly higher than that from the chapter on prenatal development and birth (chapter 4). Out of the total number of omissions, 20% of omissions included items from chapter 1, 38% from chapter 2, 29% from chapter 3, and 13% from chapter 4. On exam 3, the frequency of item omissions from the chapter that covered cognitive development of preschool children (chapter 9) was statistically significantly higher than that from the chapter on psychosocial development (chapter 10). Out of the total number of omissions, 31% of omissions included items from chapter 8 (biosocial development), 45% from chapter 9 (cognitive development), 24% from chapter 10 (psychosocial development). On exam 5, the differences of frequencies of item omissions were statistically significant across the exam but not on specific chapters. This information provides feedback to the instructor about students' learning in the course and about teaching the material those chapters covered. Most likely students found the chapters from which they omitted items with higher frequency to be more difficult, and these chapters should be taught more carefully in the future.

Study Strategies and Test Taking Strategies

Another concern for educators might be whether the question omitting procedure made students study less for exams. Only 4% of students reported that they studied less; 21% reported they studied more, and 75% reported they did not change their study time. Therefore, it would appear that students do not study more or less when the option to

omit questions is the part of the testing procedure. Based on these results of students self-reports, instructors need not fear that students would reduce their study efforts in preparation for exams. On the question of whether the procedure made them aware of necessary changes that they needed to make when they studied for exams, only 1% of the students strongly disagreed with the statement, 40% disagreed, 50% agreed, and 9% strongly agreed. Relatively few students actually described the study changes they made. Several students only reported that they studied more. Other descriptions, which are illustrated by the following examples, described positive but not significant changes of study strategies. “I read more and more through material we went over in class (book).” “I started taking better notes.” “I read chapters more carefully.” “I paid more attention to the information I read/study.” “I made my study guide more in depth by answering the questions and then going on line and taking quizzes and the tests.”

Students changed their test taking strategies while using the question omitting procedure. Seventy-seven percent of students reported that the procedure made them aware of strategy changes they needed to make when they were taking the test. A common test taking strategy for the question omitting procedure that students mentioned was the level of certainty. Hunt and Hassmen (1997) emphasized the importance of certainty about knowledge in practical applications and inability of multiple choice tests to measure certainty of knowledge. On multiple choice tests, college students might answer correctly questions they are unsure about. This might explain why students in the current study did not omit all incorrect answers by applying the question omitting

procedure. In the current study, students were asked to omit questions they believed were wrong, but often they omitted right answers.

When describing test taking strategy, students mentioned their confidence level. According to Kelemen et al. (2000) low confidence for incorrect answers and high confidence for correct answers indicated good discrimination ability. The current study did not assess confidence levels of students' answers. However, several students mentioned level of confidence in correctness of their answers they omitted and indicated threshold percentages of this confidence. If the current study had measured students' confidence levels for individual items of the test, most likely the questions students omitted would have had the lowest reported confidence levels. On incorrect answers that students omitted and expressed low confidence level, students would have shown high discrimination ability. On correct answers that students omitted and expressed low confidence level about their correctness, students would have shown low discrimination ability.

When describing test taking strategy, students mentioned guessing. When students guess on multiple choice exams, they might still circle the right answer. By using the question omitting procedure students eliminated questions for which they did not know the correct answer, and on which they were guessing. For example, a student wrote: "I chose the ones I did not even think I knew the answer to. The ones that I guessed on."

Students indicated that questions they guessed on had the lowest certainty rate. Ghatala et al. (1989) labeled guessing as the lowest certainty rating on the scale 1-4 for

the correctness of each answer on multiple choice tests. The lowest rating was labeled "Not sure at all, just guessing." Pressley and Ghatala (1988) used certainty ratings of students' responses on multiple choice tests on a scale from 20% to 100%. The lowest rating (20%) of certainty of correctness of students' answers was labeled as"Just guess. Gave 5 answers and I picked one. 1 in 5 = 20%" (p. 458). Therefore, the underlying mechanism of guessing appears to be similar to a certainty rating of someone's answer.

The format of multiple choice tests might contribute to accuracy of students' monitoring of performance on a test. Ghatala et al. (1989) pointed out that multiple choice questions can include distractors with familiar information from the text or from the previous knowledge. Therefore, these distractors might decrease students' accuracy in monitoring their performance, but no evidence about this was provided by the researchers. In the same study Ghatala et al. found that fourth grade students were more persistent in studying to reach the mastery criterion on short answer than on multiple choice format tests. The researchers suggested that a multiple choice format of a test might be responsible for students' inflated sense of preparedness for the test. The current study used only multiple choice items on each exam. However, the question omitting procedure could be investigated on the test that includes different types of items such as multiple choice items, short answer items, and true and false items to allow comparison of the frequencies of omissions of different types of items.

Nine students out of 80 students did not omit questions from scoring on all exams in semester. The mean of an unadjusted score on exam 1 for these nine students was slightly lower ($M = 76.66$, $n = 9$) than the mean of 71 students who omitted questions on

all exams (M = 81.51, n = 71). Means on exams 2-4 of nine students (76.89, 77.55, 79.67, n = 9), were almost identical to the means of 71 students (76.38, 77.87, 79.06, n = 71). The mean on exam 5 was higher for these nine students (M = 84.55) than for 71 students who omitted questions on all exams (M = 80.65). Only one student in this group of nine was a low achiever with an average on all exams of 56.2. Three students had an average on five exams 70 to 80, and four students had an average on five exams higher than 80. Apparently, students who did not omit questions on all exams were not low achieving students. The average self-efficacy score (M = 5.39, n = 9) was not different from the group of 71 (M = 5.80, n = 71). The average test anxiety score (M = 3.55, n = 9) was not different than the average test anxiety for the group of 71 (M = 3.76, n = 71). The group of students who did not omit questions on all exams was not different on self-efficacy and test anxiety scores from the group who omitted questions on all exams. There was not any observable pattern among these nine students in discontinuing use of the question omitting procedure. Though the reasons that prompted these nine students to stop using the question omitting procedure are not known, the students appear to be quite similar to those who omitted questions on all exams.

Self-efficacy

In an attempt to investigate the relation between self-efficacy expectations and students' likelihood to apply the question omitting procedure and the degree of success they had in applying the procedure, students completed the Expectancy Component: Self-efficacy for Learning and Performance subscale (Appendix C). Very low or no

correlations at all were found between the subscale scores, and the total number of questions omitted or percentages of incorrect answers out of the total number omitted. There are two possible reasons for these results. First, the reason for these low and insignificant correlations might be the nature of the scale itself, which was not specific enough for the variables under investigation. Pajares (1996a) emphasized that expectancy beliefs should be specific to the task that is assessed. The items of the self-efficacy scale used in the current study assessed students' expectations to perform well in class and did not directly assess expectations of their success in applying the question omitting procedure.

Second, students' self-efficacy expectations and their actual performance in class might not be necessarily congruent with each other. Several studies reported that low performing students overestimated their performance while high achieving students underestimated their performance. However, high achieving students were still more accurate than low achieving students (Bol et al., 2005; Hacker et al., 2000; Kruger & Dunning, 1999). Klassen (2002), and Pajares (1996a) emphasized the importance of accuracy of students' self efficacy beliefs in relation to their academic performance. If students have high self-efficacy expectations but do not put enough effort to study or lack adequate study strategies, their performance may not improve in the future. It is possible that unrealistic expectations of some students affected the results in the current study. Also, students from different majors might have different expectations about a Human Growth and Development class. Some students might have expected such a class to be easier than it actually was.

Test Anxiety

In an attempt to investigate the relation between test anxiety and students' likelihood to apply a question omitting procedure and the degree of success they had in applying the procedure, students completed the Affective Component: Test Anxiety subscale (Appendix D). Very low or no correlations at all were found between the subscale scores, and the total number of questions omitted or percentages of incorrect answers out of the total number omitted.

A relation between test anxiety and metacognitive ability has not been well investigated in current research. Veenman et al. (2000) investigated two types of test anxious students in relation to their metacognitive ability. Type I students lack metacognitive monitoring and planning, and type II students do not know how and when to use metacognitive monitoring. Metacognitive skills were examined by observation and thinking aloud protocol. Therefore, no specific test was constructed to measure and compare metacognitive monitoring and students' test anxiety. The test anxiety subscale in the current study was apparently not specific enough to assess the students' likelihood to apply the question omitting procedure and the degree of success they had in applying the procedure. However, 89% of students reported that the optional procedure to omit questions from scoring gave them a sense of control over their exams. This report could lead to the conclusion that the question omitting procedure might decrease test anxiety for students who suffer from it. A questionnaire which would assess how the question omitting procedure affected anxiety on the test given to students after the last exam

would probably provide better answers about the possible relationship between test anxiety and students' metacognitive ability.

A score on an additional question, which was similar to the first item on the test anxiety subscale, yield insignificant and low correlations with the total number of questions omitted and the percentages of incorrect answers omitted out of the total number omitted. Apparently, students' perception of themselves as test-takers was not correlated with their likelihood to apply the question omitting procedure and the degree of success they had in applying the procedure.

Limitations of the Study

The current study was conducted in two sections of a Human Growth and Development class. Twenty-five percent of students who participated in the current study were education majors, and 69% were female students. Students' behavior on self-tailored exams may be different in different majors and in classes with different gender composition. For example, Lundeberg et al. (1994) indicated that college women showed more accurate perceptions of their incorrect answers, whereas men showed too much confidence in wrong answers. Therefore, gender composition might affect the likelihood to apply the option to omit questions from scoring and the degree of success participants have in applying the omission procedure.

Multiple choice questions on all exams in the current study were used from the test bank that accompanies the textbook, *The Developing Person through the Life Span*, 6th edition, by Berger (2005). No technical information or information on the origin of these items was provided. Even though instructors purposively chose the items from test

bank for exams, questions from the test bank might differ from questions constructed by the instructor of the course. These differences could affect students' likelihood to apply the option to omit questions from scoring and the degree of students' success in applying the omission procedure.

The length of each exam in the current study was specified to 75 multiple choice questions. The self-tailoring procedure was defined as the option to omit up to five questions from being scored on an exam. An estimation of students' metacognitive ability might be different on tests of different length with different test fractions being omitted from scoring.

The mentioned limitations of the current study could be addressed in future research. Implications and recommendations for future research will be discussed in the two following sections.

Implications of the Study

The current study has implications in the areas of assessment, metacognitive ability, and teaching in college classroom. The self-tailoring procedure can be incorporated on exams in college classrooms of different majors. Other types of exams, which include items such as short answer and true and false items, can incorporate the same procedure. Threats to content validity of the test can be minimized by allowing students to omit fewer than five questions, or the length of the test might be different. The reluctance of instructors to apply this procedure because of additional hand calculation could be eliminated by using a computer to calculate an adjusted score. The added benefits of using the self-tailoring procedure are improvement of students' metacognitive

skills and useful information to instructors about their teaching and students' learning in the class. However, there is no evidence as to whether the self-tailoring option affects how much or how well students learn the course content.

Students reported a favorable opinion about the optional procedure to omit up to five questions from scoring on multiple choice exams. They not only increased their score, but more importantly, they improved the percentages of incorrect answers omitted out of the total number omitted on consecutive exams. On each subsequent exam, a higher percent of students omitted all incorrect answers out of the total number omitted. Therefore, students improved by practice their metacognitive ability across five exams in a semester.

These findings were confirmed by students' reports about improving their selection of wrong answers on consecutive exams and greater confidence in distinguishing what they know from what they do not know. This is important because as educators we would like to improve students' confidence in their ability to distinguish what they know from what they do not know, which could lead to their better performance on exams, more accurate self-assessment, and changes in study behavior.

According to students' self-reports, the question omitting procedure did not affect significantly their study behavior. To induce changes in students' study behavior, most likely explicit training on study strategies would be necessary. Study changes in forms of an active questioning of material being learned would most likely improve students' ability to distinguish what they know from what they do not know when they study and

consequently improve their performance on exams. Furthermore, it might lead to more accurate self-efficacy expectations and lower test anxiety.

Better performance on exams can be influenced by more effective teaching. The results of the current study have implications for teaching a Human Growth and Development course but also for teaching other courses. The instructor could review the five most frequently omitted questions on exams which would provide guidance that further instruction was needed on topics or concepts covered by these questions. If students omitted more questions from a certain chapter, this chapter should be taught more carefully in the future. If students omitted more application items, more application examples in class should be provided. Based on observation of the most frequently omitted questions, the instructor of the course may decide whether the quality of these questions could be improved. Therefore, this information could be used to improve teaching of the course and assessment of students' knowledge in the future.

Recommendations for Future Research

Several issues that the current study highlights would benefit from further investigation. These issues are: a) the impact of improvement of metacognitive ability on future performance and preparedness for exams, b) an application of a self-tailored procedure on different types of exams, c) the impact of different test length and different numbers of questions to be omitted on content validity of the test and metacogntive ability estimation, d) the relationship between metacognitive ability, self-efficacy, and test anxiety in the context of self-tailored tests, e) differences in metacognitive ability of

high and low achieving students, and f) the possible impact of gender differences on the ability to apply a self-tailoring procedure.

The emphasis of this study was on gauging efficacy and improvement of self-tailoring skill by practice on consecutive exams and not on how well participants learned the subject matter. In the future, researchers may wish to see if those who improved their skill of self-tailoring are those who are better prepared and perform better on exams. Instructors might be interested to see whether the option to omit questions from scoring improves or worsens the mastery content. A survey that would assess instructors' willingness to apply the procedure in a classroom could be developed for future research.

Exams in the current study consisted from all multiple choice items. However, the question omitting procedure could be investigated on a test that includes items such as short answer items and true and false items to allow comparison of the frequencies of omissions of different types of items. Questions on the tests, instead of being chosen only from the test bank, can be also constructed by the instructor; therefore, the frequencies of omissions of the two types of items can be compared. Also, different types of tests can be compared in future research. For example, on a test which questions are constructed by the instructor of the course, students' behavior might be different than on a test which questions are chosen from the test bank.

Changing test lengths or the number of questions allowed to be omitted from scoring might have an impact on content validity of the test and metacogitive ability estimation. For example, in future research allowing students to omit fewer than five questions could be investigated. Allowing fewer questions to be omitted would minimize

further any threat to content validity of the test. It could also answer whether students' accuracy to discriminate between correct and incorrect answers would be higher if they had been restricted to a lower limit of items to be omitted from scoring than in the current study. Therefore, different combinations of test length and a test fraction that could be omitted from scoring might be used in future research of self-tailored exams.

More studies about students' metacognitive monitoring of test-performance on self-tailored tests are necessary to clarify the relations among variables investigated in the current study. For example, in future studies, self-efficacy and test anxiety scales should be more specific to test taking strategies, and whatever method is used for self-tailoring. New scales should be developed to assess the relationship between self-efficacy, test anxiety, and the results of any self-tailoring procedure.

Furthermore, studies should answer the question of how metacognitive ability, self-efficacy beliefs, and test anxiety are related to each other if they are related at all. For this purpose new or different indicators of metacognitive ability could be developed. Future studies could clarify whether an improvement of students' metacognitive ability to distinguish what they know from what they do not know would improve their performance on exams, and how it would affect students' self-efficacy beliefs, and test anxiety. Whereas in the current study, self-efficacy and test-anxiety were measured once, prior to the first exam, multiple instances of self-efficacy and test anxiety measurements might help to answer this question.

Researchers might want to investigate whether low or high achievement on a test makes a difference on how much students improve their metacognitive ability by

practice. Additionally, future research could answer whether high achieving students benefit more from a self-tailoring procedure than do low achieving students.

The current study did not examine the impact of possible gender differences on students' likelihood to apply the option to omit questions from scoring and the degree of success that students had in applying the question omitting procedure. Future research on self-tailored testing should include gender as a potential explanatory variable.

Though the current study extends our understanding of how metacognitive ability can and does change via practice on multiple exams, there is still much that we do not yet know. The recommendations listed here, if followed, would do much to enhance our understanding of how metacognitive ability, learning of content, and learner characteristics such as gender, achievement, self-efficacy, and test anxiety relate to the content of a self-tailored testing procedure.

REFERENCES

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist*, *44*(9), 1175-1184.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117-148.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bedard, R. (1974). Partly tailored examinations. *The Alberta Journal of Educational Research, 20*(1), 15-23.

Benjamin, M., McKeachie, W. J., & Lin, Y. G. (1987). Two types of test-anxious students: Support for information processing model. *Journal of Educational Psychology, 79*(2), 131-136.

Benjamin, M., McKeachie, W. J., Lin, Y. G., & Holinger, D. P. (1981). Test anxiety: Deficits in information processing. *Journal of Educational Psychology, 73*(6). 816-824.

Bol, L., & Hacker, D. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education, 69*(2), 133-151.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style of calibration accuracy and performance. *The Journal of Experimental Education, 73*(4), 269-290.

Bouffard-Bouchard, T. (2001). Influence of self-efficacy on performance in a cognitive task. *The Journal of Social Psychology, 130*(3), 353-363.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., et al. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology, 97*(2), 268-274.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont: CA: Wadsworth.

Culler, R. E., & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72*(1), 16-20.

Everson, H. T., Smodlaka, I., & Tobias, S. (1994). Exploring the relationship of test anxiety and metacognition on reading test performance: A cognitive analysis. *Anxiety, Stress, and Coping,7*, 85-96.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-development inquiry. *American Psychologist, 34*(10), 906-911.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231-235). Hillsdale, NJ: Erlbaum.

Garner, R., & Alexander, P. A. (1989). *Educational Psychologist, 24*(2), 143-158.

Ghatala, E., Levin, J. R., Foorman, B. R., & Pressley, M. (1989). Improving children's regulation of their reading PREP time. *Contemporary Educational Psychology, 14,* 49-66.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506-528.

Gitomer, D. H. (2000). Foreword to the second edition. In H. Wainer (Ed.). *Computerized adaptive testing: A primer* (2nd ed., pp. xiii-xv). Mahwah, NJ: Lawrence Erlbaum.

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160-170.

Hunt, D. P., & Hassmen, P. (1997). What it means to know something. *Reports from the Department of Psychology*, *Stockholm University, 835*, 1-16.

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22*(3), 255-278.

Kelemen, W. L., Frost, P. J., & Weaver III, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*(1), 92-107.

Klassen, R. (2002). A question of calibration: A review of the self-efficacy beliefs of students with learning disabilities. *Learning Disability Quarterly, 25,* 88-102.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessment. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159-183.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8*(3), 147-151.

Lundeberg, M. A., Fox, P. W., Brown, A. C., & Elbedour, S. (2000). Cultural influences on confidence. Country and gender. *Journal of Educational Psychology, 92*(1), 152-159.

Lundeberg, M. A., Fox, P. W., & Puncochar, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology, 86*(1), 114-121.

Lunz, M. E., & Bergstrom B. (1995). Computerized adaptive testing: Tracking candidate response patterns. *Journal of Educational Computing Research, 13*(2), 151-162.

McCormick, C. B. (2003). Metacognition and learning. In W. M. Reynolds, & G. E. Miller (Eds.). *Handbook of psychology: Vol. 7. Educational Psychology* (pp.79-102). Hoboken, NJ: John Wiley.

Morse, D. T. (1988). *An exploratory study of self-tailored tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Morse, D. T., & Morse, L. W. (2002). Are undergraduate examinees' perceptions of item difficulty related to item characteristics? *Perceptual and Motor Skills, 95*, 1281-1286.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *The Journal of Experimental Education, 74*(1), 7-28.

Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*(3), 131-142.

Pajares, F. (1996a). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578.

Pajares, F. (1996b). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology, 2*(4), 325-344.

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition, 29*(1), 62-67.

Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385-407.

Pintrich, P. R., & De Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 62(*1), 51-59.

Pintrich, P. R., Smith, D. F., Garcia, T., & McKeachie, W. J. (1991). A manual for the use of the motivated strategies for learning Questionnaire (MSLQ). (Tech. Rep. No. 91-B-004). Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning, The University of Michigan.

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement, 38*(3), 235-247.

Pressley, M., & Ghatala, E. S. (1988). Delusions about performance on multiple-choice comprehension tests. *Reading Research Quarterly, 23*(4), 454-464.

Pressley, M., & Ghatala, E. S. (1989) Metacogntive benefits of taking a test for children and young adolescents. *Journal of Experimental Child Psychology, 47*, 430-450.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25*(1), 19-33.

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly, 22*(2), 219-236.

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgment. *Journal of Experimental Education, 65*(2), 135-146.

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26,* 113-125.

Schraw, G., & Dennison, R. (1994). *Contemporary Educational Psychology, 19,* 460-475.

Schraw, G, Dunkle, M. E., Benedixen, L. D., & Roedel, T. (1995). Does a general monitoring skill exist. *Journal of Educational Psychology, 83*(3), 433-444.

Schraw, G., & Graham, T. (1997). Helping gifted students develop metacognitive awareness, *Roeper Review, 20*(1), 4-8.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351-369.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18,* 455-463.

Schraw, G., & Roedel, T. (1994). Test difficulty and judgment bias. *Memory & Cognition, 22*(1), 63-69.

Shaughnessy, J. J. (1979). Confidence-judgment accuracy as a predictor of test performance. *Journal of Research in Personality, 13*, 505-514.

Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *Journal of Instructional Psychology, 22*(1), 77-87.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed, pp.101-133). Mahwah, NJ: Lawrence Erlbaum.

Thompson, W. B. (1999). Individual differences in memory-monitoring accuracy. *Learning and Individual Differences, 11*(4), 365-376.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Veenman, M. V. J, Kerseboom, L. & Imthorn, C. (2000). Test anxiety and metacognitive skillfulness: Availability versus production deficiencies. *Anxiety, Stress, and Coping, 13*, 391-412.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press, Inc.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492.

Wise, S. L. (1994). Understanding self-adapted testing: The perceived control hypothesis. *Applied Measurement in Education, 7*(1), 15-24.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement, 29*(4), 329-339.

Wood, R. (1974). Response-contingent testing. *Review of Educational Research, 43*, 529-544.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*(3), 611-617.

Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*(3), 329-339.

Zimmerman, B. J., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*(1), 51-59.

APPENDIX A

INSTRUCTIONS GIVEN TO STUDENTS FOR THE QUESTION OMITTING PROCEDURE

On multiple choice exams in this class, all students will be allowed to omit questions from being scored on the test. The procedure is as follows. After you answer all the questions on the exam, you can mark up to five questions that you want to omit from scoring of the test. Be sure to mark only questions that you believe you answered incorrectly, or questions that you might answered incorrectly and are not certain of the correct answers. You will mark spaces on the back of the answer sheet. For example, if you want to exclude question 12 from being scored on the exam, on the back of your answer sheet you would mark question 112. If you want to exclude question 42, on the back of your answer sheet you would mark question 142 etc. You can mark up to five questions from the total number of questions.

The new score of your test will be calculated by a simple formula. The new number of correct answers (after excluding the items omitted) will be divided by your new total number of questions on the test (total number minus the number omitted).

APPENDIX B

DEMOGRAPHIC QUESTIONNAIRE

Demographics Form Code Number _____

Age:_____

Gender (please check appropriate space) Male____ Female _____

Classification in school (please check appropriate space)

Freshmen _____

Sophomore _____

Junior _____

Senior _____

Other _____

Major: _____

Please enter composite score for the following tests. If you did not take it, please write in N/A.

ACT score (composite)_____ SAT score (composite)____

GPA (current) _____

Ethnicity

Caucasian ____ African/American ____ Hispanic ____ Native American ____

Asian ____ Pacific/Islander ____ Other ____

APPENDIX C

EXPECTANCY COMPONENT: SELF-EFFICACY FOR LEARNING AND PERFORMANCE

The following questions ask about your motivation and attitudes about this class. **Remember there are no right or wrong answers, just answer accurately as possible.** Use the scale below to answer the questions. If you think this statement is very true of you, circle 7; if a statement is not at all true of you, circle 1. If the statement is more or less true of you, find the number between 1 and 7 that best describes you.

1 2 3 4 5 6 7
not at all very true of me
true of me

Expectancy Component: Self-efficacy for Learning and Performance

1. I believe I will receive an excellent grade in this class.

1 2 3 4 5 6 7

2. I'm certain I can understand the most difficult material presented in the readings for this course.

1 2 3 4 5 6 7

3. I'm confident I can understand the basic concepts taught in this course.

1 2 3 4 5 6 7

4. I'm confident I can understand the most complex material presented by the instructor in this course.

1 2 3 4 5 6 7

5. I'm confident I can do an excellent job on the assignments and tests in this course.

1 2 3 4 5 6 7

6. I expect to do well in this class.

1 2 3 4 5 6 7

7. I'm certain I can master the skills being taught in this class.

1 2 3 4 5 6 7

8. Considering the difficulty of this course, the teacher, and my skills, I think I will do well in this class.

1 2 3 4 5 6 7

APPENDIX D

AFFECTIVE COMPONENT: TEST ANXIETY

The following questions ask about your motivation and attitudes about this class. **Remember there are no right or wrong answers, just answer accurately as possible.** Use the scale below to answer the questions. If you think this statement is very true of you, circle 7; if a statement is not at all true of you, circle 1. If the statement is more or less true of you, find the number between 1 and 7 that best describes you.

1	2	3	4	5	6	7
not at all true of me						very true of me

Affective Component: Test Anxiety

1. When I take a test I think about how poorly I am doing compared with other students.

1 2 3 4 5 6 7

2. When I take a test I think about items on other parts of the test I can't answer.

1 2 3 4 5 6 7

3. When I take tests I think of the consequences of failing.

1 2 3 4 5 6 7

4. I have an uneasy, upset feeling when I take an exam.

1 2 3 4 5 6 7

5. I feel my heart beating fast when I take an exam.

1 2 3 4 5 6 7

APPENDIX E

THE ADDITIONAL QUESTION

The following question asks about your test-taking skills compared to others. Use the scale below to answer this question. Circle the answer that best reflects your feelings.

How good a test-taker do you think you are compared to others?

1	2	3	4	5
not good at all	poor	good	very good	excellent

APPENDIX F

QUESTIONNAIRE GIVEN TO STUDENTS AFTER EXAM 5

CODE NUMBER_____

For questions 1-8, please circle the response that best reflects your feelings.

1. The opportunity to exclude up to five questions from exams helped me to improve score on my exams.

Strongly Disagree Disagree Agree Strongly Agree

2. I would recommend that this procedure be included on other multiple choice exams.

Strongly Disagree Disagree Agree Strongly Agree

3. I became more confident in distinguishing what I know from what I do not know because of this procedure.

Strongly Disagree Disagree Agree Strongly Agree

4. I improved my selection of wrong answers to be omitted from scoring on successive exams.

Strongly Disagree Disagree Agree Strongly Agree

5. This optional procedure gave me a sense of control over my exams.

Strongly Disagree Disagree Agree Strongly Agree

6. The procedure caused me to

A) study more for my exams.

B) study less for my exams

C) not change my study time for my exams

7. The procedure made me aware of strategy changes that I needed to make when I was

taking a test.

Strongly Disagree Disagree Agree Strongly Agree

8. The procedure made me aware of necessary changes that I need to make when I study for exams.

Strongly Disagree Disagree Agree Strongly Agree

9. If you made any changes in test-taking strategies, please describe these changes.

_____

_____

_____

10. If you made any changes in your study strategies, please describe these changes.

_____

_____

_____

11. In a few sentences describe the strategy that you used to decide which questions to omit from scoring.

_____

_____

_____

12. On average, how many hours per week did you study and read for this class?

_____hour(s).

APPENDIX G

INFORMED CONSENT

Consent Form

College Students' Behavior on Self-tailored Multiple Choice Exams in Relation to Metacognitive Ability, Self-efficacy, and Test Anxiety

Project Purpose

The purpose of this project is to investigate college students' behavior on multiple choice tests. Students will be given an option to omit questions from being scored on the test they believe answered incorrectly and are not certain of the correct answers. Students' self-efficacy expectations and test anxiety will be examined in relation to the question omitting procedure. The study will investigate the type of questions that student omitted from being scored on the test and the effect of the question omitting procedure on students' grade.

Procedure and Confidentiality of Data

Before taking the first exam you will be asked to answer two short questionnaires: a) about your self-efficacy, and b) test anxiety. After answering all the questions on the multiple choice exam, you will be asked to mark on the back of an answer sheet the questions that you want to omit from being scored on the test. You can mark up to 5 questions. Your grade on the test will be calculated on a scale 1-100 based on questions that you did not omit. We expect the duration of your voluntary participation will be approximately 5-10 minutes in addition to the time required to complete the exam. After the fifth exam, you will be asked to complete a questionnaire that addresses your perception of the test scoring method used in the current study. This form requests your

consent to participate in a research study. Your consent to participate in the current study gives permission to Jasna Vuk to use for research purposes the questionnaires that you answered and the results of your multiple choice exams taken in this class. Your name will be only used for the grading purposes in this class, but it will not be released for any other purpose. Your name will be deleted from all records and data associated with this research. You may withdraw your consent (up until the point that your material are identified), and your records will not be used for the current study.

Risks of Participants

There are **NO** foreseeable risks or discomforts that might occur as a result of your participation in this research.

Benefits of Participation

You will be allowed to freely omit questions from being scored on the test that you believe answered incorrectly, or questions that you might have answered incorrectly and are not certain of the correct answers. By doing so you might improve your final score on the test. The broader benefits of your participation include the added value of additional scientific findings in this specific area of research.

Further Information

If you should have any questions about this research project, please feel free to contact Jasna Vuk by telephone at _____ or by e-mail at _____. For additional information regarding human participation in research, please feel free to contact MSU Regulatory Compliance Office at _____.

Voluntary Participation

Please understand that your participation is voluntary. Your refusal to participate will involve NO penalty or loss of benefits to which you are otherwise entitled. All students are allowed to omit 1-5 questions from the exam regardless of their decision to participate in the current study or not. You may discontinue your participation at any time without penalty of loss of benefits. You will be provided with a copy of this form for your personal records.

Participant: _____	_____	_____
Print your name	Signature	Date
Investigator: _____	_____	_____
	Signature	Date