Computer simulation for writing and evaluating multiple choice test items

Cheryl Olmstead Hausafus

Iowa State University
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in "sectioning" the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from "photographs" if essential to the understanding of the dissertation. Silver prints of "photographs" may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106
HAUSAFUS, CHERYL OLMSTEAD
COMPUTER SIMULATION FOR WRITING AND
EVALUATING MULTIPLE CHOICE TEST ITEMS.
IOWA STATE UNIVERSITY, PH.D., 1978

© 1978
CHERYL OLMSTEAD HAUSAFUS
All Rights Reserved
Computer simulation for writing and evaluating multiple choice test items

by

Cheryl Olmstead Hausafus

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major: Home Economics Education

Approved

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1978

Copyright © Cheryl Olmstead Hausafus, 1978. All rights reserved.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>Problem</td>
<td>4</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>5</td>
</tr>
<tr>
<td>Selection of Content</td>
<td>5</td>
</tr>
<tr>
<td>Objectives of the Study</td>
<td>7</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>8</td>
</tr>
<tr>
<td>Assumptions</td>
<td>8</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>9</td>
</tr>
<tr>
<td>Organization of the Report</td>
<td>11</td>
</tr>
<tr>
<td>REVIEW OF LITERATURE</td>
<td>12</td>
</tr>
<tr>
<td>Measurement and Item Writing Instruction</td>
<td>12</td>
</tr>
<tr>
<td>Item Writing Principles</td>
<td></td>
</tr>
<tr>
<td>Optimum number of options</td>
<td>18</td>
</tr>
<tr>
<td>Negatively oriented item stems</td>
<td>19</td>
</tr>
<tr>
<td>Repetition between the stem and option</td>
<td>20</td>
</tr>
<tr>
<td>Specific determiners</td>
<td>22</td>
</tr>
<tr>
<td>Grammatical inconsistencies</td>
<td></td>
</tr>
<tr>
<td>Length of option</td>
<td>23</td>
</tr>
<tr>
<td>Response set</td>
<td>25</td>
</tr>
<tr>
<td>Incomplete stems</td>
<td></td>
</tr>
<tr>
<td>Effect of Violations According to Student's Grade Level</td>
<td>25</td>
</tr>
<tr>
<td>Summary</td>
<td>28</td>
</tr>
<tr>
<td>DEVELOPMENT OF THE SIMULATION</td>
<td>32</td>
</tr>
<tr>
<td>Description of the MCTA Simulation</td>
<td>32</td>
</tr>
<tr>
<td>Initial Considerations</td>
<td></td>
</tr>
<tr>
<td>Design criteria</td>
<td>33</td>
</tr>
<tr>
<td>Selection of the computer system</td>
<td>34</td>
</tr>
<tr>
<td>Identification of item writing principles</td>
<td>36</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Mortality</td>
<td>97</td>
</tr>
<tr>
<td>Item analysis</td>
<td>102</td>
</tr>
<tr>
<td>Graduate Experiment</td>
<td>102</td>
</tr>
<tr>
<td>MCTA descriptive data</td>
<td>110</td>
</tr>
<tr>
<td>Item analysis</td>
<td>112</td>
</tr>
<tr>
<td>Summary</td>
<td>114</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>121</td>
</tr>
<tr>
<td>Introduction and Utilization of the Simulation</td>
<td>125</td>
</tr>
<tr>
<td>Refinement of the Instrument</td>
<td>126</td>
</tr>
<tr>
<td>Refinement of the Simulation</td>
<td>127</td>
</tr>
<tr>
<td>LITERATURE CITED</td>
<td>129</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>137a</td>
</tr>
<tr>
<td>APPENDIX A: PRO SPECIFIC DETERMINERS AND CON SPECIFIC DETERMINERS USED IN MCTA APPAYS</td>
<td>138</td>
</tr>
<tr>
<td>APPENDIX B: IGNORE FOR REPETITION ARRAY USED IN MCTA</td>
<td>139</td>
</tr>
<tr>
<td>APPENDIX C: NEGATIVE ARRAY USED IN MCTA</td>
<td>140</td>
</tr>
<tr>
<td>APPENDIX D: OPEN ENDED QUESTIONS ADMINISTERED TO PILOT GROUP</td>
<td>141</td>
</tr>
<tr>
<td>APPENDIX E: INSTRUCTIONS FOR USING THE PLATO TERMINAL AND MCTA SIMULATION</td>
<td>143</td>
</tr>
<tr>
<td>APPENDIX F: PRE-POST TEST INSTRUMENT</td>
<td>147</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Effects of item writing violations on item indices</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>Solomon Four-Group experimental design as used in the study</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>Item analysis data for instrument used in pilot testing</td>
<td>83</td>
</tr>
<tr>
<td>4</td>
<td>Analysis of variance for undergraduate educational psychology teachers</td>
<td>93</td>
</tr>
<tr>
<td>5</td>
<td>Mean scores for undergraduate educational psychology teachers</td>
<td>94</td>
</tr>
<tr>
<td>6</td>
<td>Paired t-tests of pretest and posttest for undergraduate class</td>
<td>96</td>
</tr>
<tr>
<td>7</td>
<td>Mean values for undergraduate class on simulation related variables</td>
<td>97</td>
</tr>
<tr>
<td>8</td>
<td>Partial correlation matrix for undergraduate teachers controlling for typing skill</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>Partial correlation matrix for undergraduate teachers controlling for total course points</td>
<td>101</td>
</tr>
<tr>
<td>10</td>
<td>Item analysis data for posttest instrument used in undergraduate educational psychology class</td>
<td>103</td>
</tr>
<tr>
<td>11</td>
<td>Analysis of variance for graduate evaluation class</td>
<td>106</td>
</tr>
<tr>
<td>12</td>
<td>Mean scores for graduate evaluation class</td>
<td>107</td>
</tr>
<tr>
<td>13</td>
<td>Paired t-tests of pretest and posttest for graduate teachers</td>
<td>109</td>
</tr>
<tr>
<td>14</td>
<td>Mean values for graduate teachers on simulation related variables</td>
<td>111</td>
</tr>
<tr>
<td>15</td>
<td>Intercorrelation matrix of simulation related variables for graduate teachers</td>
<td>113</td>
</tr>
<tr>
<td>16</td>
<td>Item analysis data for posttest instrument used in graduate evaluation class</td>
<td>115</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Figure 1.</td>
<td>Flow chart to check for repetitions</td>
<td>42</td>
</tr>
<tr>
<td>Figure 2.</td>
<td>Flow chart to check for repetitious phrase</td>
<td>43</td>
</tr>
<tr>
<td>Figure 3.</td>
<td>Illustration of complete item entered in the editor</td>
<td>65</td>
</tr>
<tr>
<td>Figure 4.</td>
<td>Illustration of initial analysis of item by MCTA</td>
<td>66</td>
</tr>
<tr>
<td>Figure 5.</td>
<td>Flow chart for teacher action in the MCTA simulation</td>
<td>69</td>
</tr>
<tr>
<td>Figure 6.</td>
<td>Illustration of final MCTA summary</td>
<td>74</td>
</tr>
</tbody>
</table>
Members of the educational community are continually seeking instructional strategies which improve the learning situation. As a result, recently they have developed numerous methods to enhance their teaching. These include programmed instruction, multi-media lessons, individualized instruction and simulation. With the advent of the digital computer a concerted effort has been made to develop the apparent potential of computers for instructional support. Computers have been used in both the management of instruction, and instruction itself. This study focused on the use of computers by students as a learning medium.

Instructional computer applications are implemented in either the batch or interactive mode. Batch applications require that all of the data and instructions be assembled and fed into the computer at one time. The computer then processes the data and produces the result without user intervention. If the batch user wishes to alter the data or instructions another job must be submitted. On the other hand interactive applications involve a conversation between the user and the computer, frequently requiring the user to intervene in the execution of the program to make critical procedural decisions. The interactive approach is especially suited to most learning situations where a student's
responses dictate the next procedures. The seemingly "immediate" feedback the student receives is advantageous for drill and reinforcement and more closely resembles traditional teaching strategies.

Tutorial and drill-and-practice type programs have held considerable appeal to most educators. The methodology for their design and implementation is well defined and straightforward. They conform to common instructional practices, substituting for the presence of the teacher in monotonous, repetitious activities while providing instruction at the student's convenience and pace. Such programs can record the student's progress toward prescribed competencies and document any deficiencies.

In spite of these appealing features, a trend toward student directed strategies such as problem solving, simulation and inquiry methods has developed. Occurrence of programs employing these techniques has steadily increased to account for 35% of the programs reported in 1976 (Kearsley, 1976). These programs give the student more control over the learning situation, thus appearing to encourage the development of problem solving skills, integration and testing (Rockart & Morton, 1975).

Many educators now view problem solving and simulation approaches as representing the path toward full realization of the computer's contribution to instruction. However,
skepticism has been expressed concerning present implementation of this movement. Papert (1972) expressed concern for the value of several simulation and inquiry approaches. In an effort to insure each student's learning success, Papert felt some strategies became so simplified that they provided foolproof operation, or discovery techniques were so manipulated that the student could not creatively approach the problem. When the learning experience is controlled to this degree of detail, the effects are counter to the hope for self-directed, proactive learning described by Knowles (1975) and Tough (1971).

Recent authors have called for programs using heuristic strategies which allow learners to take alternate approaches to the learning environment (Dwyer, 1974; Molnar, 1976a,b; Papert, 1972; Rockart & Morton, 1975). Such strategies employ open-ended procedures where neither the output nor algorithm is completely specified. Instead, the procedure addresses more global goals helping to "make both decisions and discoveries, but leaving open the question of the universe in which they are to operate" (Dwyer, 1974, p. 140). These programs often support an environment where the student is encouraged to manipulate and observe to achieve some goal. Within the confines of this environment the student provides stimuli and the computer reacts without editorialization. The student must develop criteria for
evaluating the reaction and adapt the stimuli to produce the global goal. Such programs have been classified by Rockart and Morton as having an "enrichment" function. The design, development and evaluation of this type of program in a sufficient number of areas to provide the necessary guidance to the movement is a challenging endeavor.

Problem

Currently most computer based learning materials are developed for college level instruction (Molnar, 1976a). This is to be expected when the expense of facilities to support an instructional computer is considered. However, when computer based learning materials are categorized according to the ten major disciplines which support this development, education ranks ninth with only 3.1% of the total materials developed for that field. It seems incongruous that computer assisted instruction continues to show promise for education, yet the education field makes little use of the computer for learning strategies. Program development for the educational disciplines should be encouraged not only to facilitate creative learning approaches, but also to provide educators with experiences to assess more realistically the capabilities of the computer for educational purposes.
Purpose of the Study

The purpose of this study was to develop a computer program for educators which facilitated heuristic learning strategies. Selection of content for the program met the following criteria:

a. content could be structured for development of a suitable algorithm
b. instruction in content was currently presented in a manner unsuited for the outcomes desired
c. instructional strategy could encourage self-directed, proactive learning.

Selection of Content

One of the basic instructional tasks of a teacher is evaluation. Recent educational trends in accountability, planning by objectives, and competency based instructional programs have brought about a greater demand for appropriate evaluation techniques. Yet, it is an unfortunate "fact that preservice programs in teacher preparation, by and large, do not begin to adequately provide for an acceptable set of (measurement) competencies" (Mayo, 1967, p. 1).
Despite the increasing use of multiple choice tests to evaluate pupils, teachers and programs, the development of such tests is often taught in very general terms. Preservice teachers get little opportunity to try their test writing skills, and even less opportunity to have their items evaluated in terms of item writing principles. With recent emphasis on teacher competencies rather than acquired knowledge, the opportunity to practice item writing skills and to receive pertinent evaluation of them becomes even more important. Typical instruction in test development at the undergraduate level has included identification of helpful rules for item writing, critiquing some previously developed test instruments, and constructing several items or an entire test while observing the identified rules. The instructor may have evaluated the test items developed, checking for accuracy, appropriateness to the content or test plan, and application of item writing principles. Such evaluation took a great deal of time and may have weakened the effect due to focusing on many different concerns at once. The student seldom was given additional opportunity to improve his or her test writing skills under the instructor's direction, or to administer the items to actual pupils. The computer can enhance the instruction on multiple choice item writing and on evaluating the items written. It can simulate the responses of a class answering the item. Such help would be
readily available for an unlimited number of items, thus allowing more specific instruction in an efficient and economical way.

Objectives of the Study

The design and development of a computer simulation for writing and evaluating multiple choice test items was the major thrust of this study. Thus the primary objectives were:

a. Develop an instructional program to assist preservice teachers in acquiring the ability to write multiple choice test items for use with middle and secondary level students.

b. Develop a computer simulation to evaluate multiple choice test items appropriate for middle and secondary level students according to selected item writing principles.

Secondary objectives were:

a. Determine the effectiveness of the computer simulation to increase the teacher's posttest score on writing multiple choice items.

b. Identify the effect of certain computer learning related variables on the teacher's posttest score.
It was not intended that the instructional program be self-contained but rather that it act as a resource for classroom instruction. Preliminary instruction before using the computer simulation was expected. The approach for evaluating items more closely resembled information available to a classroom teacher on items he or she developed such as class response, means, etc., rather than the common evaluative practice of instructors to identify and name the specific violations.

Hypothesis

The main hypothesis for investigation was:
There is no significant difference in posttest scores measuring the ability to write test items free from selected violations between preservice teachers using the computer simulation and those not using the simulation.

Assumptions

For this study the following assumptions were made:
a. The subject matter content of test items has no unique factors which affect the outcome of selected item writing violations.
b. A model can be developed which simulates the effect of selected factors on test scores of middle and secondary level students. Prediction equations used in the model, while not based on factual evidence, can model the effect on test scores.

c. The evaluation of test items constructed by preservice teachers encourages the teachers to observe these item writing principles when constructing multiple choice test items for classroom use.

d. Observing item writing principles when constructing test items increases the reliability of teacher made multiple choice tests.

Definition of Terms

Special terms and commonly used words are sometimes defined differently according to the author's background and purposes of writing. This study focused on preservice teachers in educational psychology and home economics education courses using a computer to simulate the response patterns of a hypothetical class to certain test items. To avoid confusion, it was necessary to define several terms as they were used in this study. These definitions follow:
Algorithm--a step-by-step procedure to accomplish a task or model a system

Cues--hints inherent within an item which a student may use to help choose an answer to a test item

Difficulty index--the percent of students responding correctly to an item

Discrimination--comparison quotient of students answering an item correctly to students having either high or low ability in the trait measured by the item

Instructor--the person who administered the computer simulation as part of his or her instruction in the educational classes

Item indices--statistical measures of the item's acceptability; measures include item difficulty and item discrimination

Preservice teacher--the teacher-in-training enrolled in education classes used in the study; although some teachers may have already taught in the classroom, all teachers in the study will be collectively called preservice teachers

Simulation--model of physical or social situations which portray reality in a reduced scale or simplified form

Student--the hypothetical learner being tested with the items written

Teacher--the preservice teacher (education student) who used the computer simulation
Testwise behavior—strategy employed by a wary student for answering test items for which the correct answer is not known.

Organization of the Report

This study may be of interest to both computer programmers and educators. Therefore, the chapters are structured in the following manner. Literature pertaining to trends in computer instructional programs appears in this chapter. Chapter two is limited to a survey of literature on measurement and test writing. Chapter three describes the Multiple Choice Test Analyzer (MCTA) simulation, literature citing previous developments with the PLATO computer system, and a detailed report of the algorithm and programming used for the MCTA. The final chapters describe the method, analysis, and findings and recommendations of the study. Specifically, the chapters are organized as follows:

Chapter 1 Introduction
Chapter 2 Review of literature on testing and item writing practices
Chapter 3 Development of the simulation
Chapter 4 Method of procedure
Chapter 5 Findings and discussion
Chapter 6 Summary
REVIEW OF LITERATURE

A survey of the literature was conducted to develop a research base for the assumptions made in development of the simulation. This review is divided into two sections. Section one reviews instruction in measurement and item writing. Section two focuses on specific item writing principles and the effect their violations would cause on test scores. Literature pertaining to computer based learning and PLATO system simulations is described in Chapters 1 and 3.

Measurement and Item Writing Instruction

Measurement and item writing skills are accepted as fundamental in the education of teachers. However, minimal research investigating the instruction of measurement and item writing theory has been reported. Mayo (1967) conducted a study of the preservice preparation of teachers in educational measurement. He asked 185 measurement experts to rank the importance of possessing 70 specified measurement competencies. He found agreement on the importance of certain core competencies, but a diversity in thinking about how and when the competencies should be learned. Among the competencies rated highly important were "knowledge of
advantages and disadvantages of teacher-made tests" and "knowledge of general principles of test construction."

Competencies judged to have medium importance included "knowledge of the advantages and disadvantages of various types of objective test items", "ability to construct different types of test items", and "knowledge of concepts of validity, reliability and item analysis" (Mayo, 1967, pp. 78-80).

A sample of 2877 graduating seniors from 86 teacher education institutions tested by Mayo did not possess the knowledges and skills in measurement which had been defined as important, and during a two year period after graduation, the teacher education graduates showed only a small improvement in their knowledges and skills. Mayo concluded that preservice tests and measurements courses could be improved by

(a) use of more and better audio-visual aids;
(b) more laboratory and field experiences;
(c) more meaningful presentation of material;
(d) improved evaluation of achievement;
(e) establishment of minimum or optimum standards for measurement courses (Mayo, 1967, pp. 56-57).

Mayo also suggested emphasizing the independent role of the student in an improved self-instructional environment.
In comparing two patterns of instruction for teaching item writing theory and skills, Masonis (1970) found no significant difference in ability to write test items between preservice teachers receiving one week of formal classroom instruction prior to student teaching and teachers receiving individualized instruction that consisted of discussion and analysis of the test items they had used while student teaching (Masonis, 1970).

There is an apparent need for better preparation of preservice teachers in evaluation and measurement. Unfortunately research to provide direction for such preparation is limited. The effects of both the form of instruction, whether it be individualized or in a class, and the time the instruction occurs in a teacher education program continue to pose questions for future investigation.

Item Writing Principles

Recently, many authors have shown interest in multiple choice test construction and the factors which influence a student's score on an objective test (Carter, 1971; Frisbie & Ebel, 1972; Jessell & Sullins, 1975; Miller & Williams, 1973; Powell & Isbister, 1974; Pyrczak, 1972a). Other authors have been concerned with specific models to account for guessing and other variance in a student's true test score (Ardiff,
Several additional studies have investigated the effects of violating item writing principles on students' test scores. The emphases in these studies have been in two areas, identifying change in item indices when a violation occurs, and examining test-wise behaviors of students. Test-wise behaviors are strategies a student uses to find cues for answering test items when he or she does not possess the knowledge measured in the items. The main purpose of this review was to determine the effect of violations on item indices. However, since some test-wiseness studies gave evidence of change in item indices, both types of studies were examined.

In a review of 57 educational psychology texts, Masonis (1970) made an unsuccessful search for a universally accepted list of principles for constructing test items. He compiled a list of 44 common principles which were mentioned by ten or more writers. Of those, 15 principles were listed for construction of multiple choice items. They were:

1. When an item is based on a controversial topic, the source of the content or idea should be mentioned in the stem.

2. Negative statements, especially double negative
statements, should be avoided.

3. The language of the item should be kept simple.
   Language should not be a barrier to overcome when
   the student is answering the question.

4. The stem should state a central problem or question
   that is complete enough to require homogeneous
   options.

5. The item should be stated as briefly as possible
   with no extraneous clauses or phrases, i.e., clauses
   or phrases that need not be used to answer the
   question.

6. Grammatical consistency should be maintained between
   the stem and every option.

7. Words in the stem that are the same or similar
   enough to words in the options to provide
   unnecessary clues should be avoided.

8. The options should be homogeneous enough in content
   so that all options are plausible to the uninformed
   student.

9. All options should be approximately the same length
   and written in the same form.

10. The options should not overlap in terms of the
    content they include.

11. The options should be placed at the end of the stem,
    not in the middle of it.
12. There should be only one correct answer for each item.

13. Each item should contain a minimum of four and a maximum of five options.

14. Each option should be placed on a separate line when the item is physically arranged.

15. Each item should be numbered and each option should be lettered.

(Masonis, 1970, pp. 31-32)

Determining the effects on test performance of violating multiple choice item writing principles was the major purpose of this review. Generally, item indices such as difficulty and discrimination levels, and total test reliability are used to evaluate the effects of the violations. Item difficulty is computed as the percent of students answering the item correctly. An optimum item difficulty ranges from 0.30 to 0.70. Item discrimination is the relationship of better students answering the item correctly to all students who answered the item correctly. An optimum discrimination index lies between 0.20 and 0.40 (Menne, 1976, p.7). Test reliability is an indication that the test will produce consistent results when administered at different times, or to different groups of similar students. A reliability of 0.65 or better for grouped results, or 0.85 or better for individual results is generally acceptable.
Item difficulty and discrimination indices are often computed to determine the appropriateness of an item. Both must be within acceptable ranges in order for an item to be judged a "good" item. Along with these indices, measures of choice attractiveness are often suggested as a final criterion. Distracters which are chosen by a high percentage of examinees may indicate an error in grading or ambiguity in the wording of the item. Furthermore, "distractors (sic) that are chosen by very few or none of the examinees should be regarded as implausible" (Pyrczak, 1972b, p. 13).

Several of the studies examined a combination of principles. Results of the studies are reviewed according to each principle investigated, rather than describing entire studies in order. The principles examined include optimum number of options, negatively oriented item stems, repetition between stem and option, specific determiners, grammatical inconsistencies between stem and option, relative lengths of options, response set, and incomplete stems.

**Optimum number of options**

Four is the minimum number of options usually suggested in rules for writing multiple choice test items. However, Tversky (1964) presented a theoretical calculation to support using three options to maximize discriminability, power and information of a test. For his calculations, he considered
the amount of time spent answering items with two, three or five options. In a study of 1566 college psychology students, Costin (1970, 1972) confirmed the three choice item as optimum when compared with four choice items. He concluded that a fourth option would not make an item more difficult since students use a systematic approach rather than guessing blindly when answering classroom test items they do not know.

**Negatively oriented item stems**

Dudycha and Carpenter (1973) investigated the effects of negatively stated item stems on item discrimination and difficulty. They found that tests containing negatively stated item stems administered to college students were significantly more difficult ($p<0.001$) than positively oriented items. They did not find any significant difference for item discrimination. No significant interaction between the positive/negative orientation of items and open/closed stem format was found. Dudycha and Carpenter (1973) concluded that "an increase or decrease in item difficulty produced by altering one format property is unlikely to be nullified due to its unique combination with another property" (p. 120).

In an exploratory study with graduate students, Slakter and Terranova (1966) as cited in Terranova (1969) found that
for items containing the same content and difficulty level, there were no significant differences in scores, or times required to complete the items containing negatives by the students. In that study however, scores were extremely skewed to the left indicating that the test was too easy for the students involved.

Terranova (1969) investigated the effects of negatively stated stems on fifth, seventh, ninth and eleventh grade students. The findings indicated that in comparing all positively stated stems with all negatively stated stems, the negative stems were more difficult than their positive counterparts, while their reliabilities remained similar within comparable grade levels.

Repetition between the stem and option

Kascnis (1970) noted in his review of literature that the most common superficial similarity between stem and keyed choice in a multiple choice item was repeating some portion of the stem in the correct choice but not repeating it in the distracters. Some types of repetition have also been identified as association (Diamond & Evans, 1972), and as clang (Gibb, 1964).

Diamond and Evans (1972) found that sixth grade students possessed and could verbalize test-wise behaviors which use cues from repetition of words, phrases and parts of words.
between the stem and the correct choice. McMorris and others (1972) found that when a test contained repetition of sounds, words or phrases in the stem and the correct choice, the items were easier for eleventh grade students.

A study by Slakter, Koehler and Hampton (1970b) found that students as young as fifth grade and through all grade levels to the eleventh grade exhibited test-wise behaviors using repetition as a cue ($p<0.05$). A replication of the study produced similar results.

Pyrczak (1973) used items from a previously published test which were unintentionally constructed using repetition of syllables or letters (pre-, -some, -ous) with undergraduate students. He found no significant differences in the mean scores for either the faulty or control test group. When the study was repeated using some correct and some faulty items in each test form, and the treatment group was given specific test-wise directions, a difference between the mean scores of the groups was significant ($p<0.01$).

Slakter, Koehler, and Hampton (1970a) taught high school seniors four test-wise behaviors to determine their effect on the students' test scores. One of the behaviors was to select the option which resembled an aspect of the stem. The study found that after receiving training in test-wise behaviors, students did not decrease the variance in their test scores.
Specific determiners

In an investigation of the test-wise behaviors of fifth through eleventh grade students, Slakter, Koehler and Hampton (1970b) found that behavior using specific determiners as cues in answering items did not occur frequently until the ninth grade. Dunn and Goldstein (1959) found using cues from specific determiners produced the highest mean scores on tests taken by Army trainees.

Grammatical inconsistencies

In testing Army trainees in basic training, Dunn and Goldstein (1959) found that inconsistencies in grammar between the stem and choices raised the difficulty level of items. However, no significant differential effect was found for reliability or validity.

Diamond and Evans (1972) found that students in the sixth grade could recognize and use grammatical inconsistencies for cues to answering fictitious content items. Students were not consistent in their use of grammar as a cue, however. Diamond and Evans also found a moderate correlation between IQ and ability to use grammatical cues.

Board and Whitney (1972) found in their investigation of college students that grammatical inconsistency had no major effect on test difficulty. In a replication study (Schmeiser
& Whitney, 1973), the authors found that grammatical inconsistencies significantly lowered the internal consistency of a test, that tests containing grammatical inconsistencies were significantly more difficult than a controlled test, and that there was no interaction between using grammatical inconsistencies as cues and the achievement level of the students.

**Length of option**

The relationship between the length of one option with other options in a multiple choice test item frequently has been investigated. Board and Whitney (1972) found a tendency for poorer students to gain more points than better students on items when the correct choices were noticeably longer or shorter than the item's distracters. In another study, the same authors (Schmeiser and Whitney, 1973) found that tests which contain items with the correct choice longer or shorter than distracters were significantly more difficult than tests containing no faulty items. They also found that the internal consistency of the length-varying option test was significantly lowered. In summarizing the findings of both studies the authors felt that results on the effect of the length of options were conflicting for college level students.
Dunn and Goldstein (1956) used alternate forms to test Army basic trainees. They found significant differences in mean scores when forms containing the over long correct choice items were compared to forms with equal choice length items.

Gibb (1964) trained college students to improve test scores by using seven test-wise behaviors. One of the behaviors was selecting the longer choice in a four choice item. Gibb found that after the short training session, students made significantly higher scores on the test than did the control group.

Diamond and Evans (1972) investigated the test-wise behaviors of sixth grade students on a test instrument containing fictitious material. They found that while the students did possess several test-wise traits, using cues from the length of an option was not a consistent test-taking behavior for sixth grade students in the study.

Chase (1964) investigated the relative length of options and their effect on 48 college students. He found students would select the option which was at least three times longer by word count more often \((p<0.01)\) than other options. When the word length ratio between options was from 2:3 to 1:2 this length option did not affect the student's choice.
Response_set

Chase (1964) also noted that a significant (p<0.01) response set was identified for either choosing the longer, or the shorter option when the correct option for less difficult items near that item were either longer or shorter than other options.

Incomplete stems

Board and Whitney (1972) found a significant effect (p<0.05) for incomplete stems to depress the test scores of all college level students studied regardless of their achievement level. In a replication study the effect of incomplete and grossly truncated stems did not produce significant differences (Schmeiser & Whitney, 1973).

Dunn and Goldstein (1956, 1959) found that the item difficulty on tests taken by Army trainees was not affected when the stem format was either an incomplete sentence or a question. Interaction of stem format with other violations also produced no significant differences.
Effect of Violations According to Student's Grade Level

From the several studies previously cited, the development of certain test-wise behavior in students from fifth grade through college can be postulated. By the fifth grade, students seemed to be able to use cues from repetition in answering multiple choice test items (Slakter, Koehler & Hampton, 1970b). By the sixth grade students used and verbalized the use of cues from repetition and specific determiners (Diamond & Evans, 1972). Beginning at the fifth through all higher grade levels investigated, students found item stems which were negatively stated more difficult than their positively stated counterparts (Dudycha & Carpenter, 1973; Terranova, 1969). At the ninth grade, specific determiners were often used as test-taking cues (Slakter, Koehler & Hampton, 1970b). While cues from grammatical inconsistencies were not often observed for sixth grade students, by eleventh grade the presence of grammatical inconsistencies lowered the difficulty of items (Diamond & Evans, 1972; McMorris & others, 1972).

The length of an option did not consistently provide cues for students at the sixth grade level (Diamond & Evans, 1972), but by the eleventh grade, extra length in the correct option lowered the difficulty of test items (McMorris &
At the college and postsecondary level, differences in mean scores of groups varied from the findings for middle and secondary level students. Repetition in sound of words in the stem and the correct choice did not significantly increase mean scores of undergraduate and graduate students (Pyrczak, 1973). No differences were found in mean scores on tests containing grammatical inconsistencies (Board & Whitney, 1972).

The relative length of choices provided conflicting results for college students. Chase (1964) found that when an option had a word length of 3:1, students would choose that option more often than otherwise. However, a word length ratio of 2:3 to 1:2 would not cause a significant difference in option choice. In another study with college students, length of option produced a significant difference in the increase in scores of poorer students (Board & Whitney, 1972). A replication study by the same authors (Schmeiser & Whitney, 1973) found a significant difference in the difficulty level of items containing options of different lengths.

Open or closed stem format for items was investigated with postsecondary and college students. Open stem and grossly truncated stem formats were found to be significantly more difficult than a closed stem item regardless of the
achievement level of the college student (Board & Whitney, 1972; Dudycha & Carpenter, 1973; Schmeiser & Whitney, 1973). The format of the stem did not affect the difficulty level of items taken by Army trainees (Dunn & Goldstein, 1959).

Summary

Independent, self-instructional learning environments have been proposed to better prepare preservice teachers for the measurement and evaluation tasks of their profession. A simulation designed to predict the responses of a class to individual test items could provide realistic data for teachers to practice their skills. To develop an algorithm of testing behavior for such a simulation, a data base was needed.

For the simulation data base, a survey of literature was made examining the effects of item writing violations. Several researchers have investigated the testing response behaviors of students. Their findings provided trends and data for use with the algorithm. Items containing violations progressively increased their effect on students according to grade level. This was probably due to students' increasing awareness of vocabulary. By the college or adult level however, the violations had little effect. A greater experience base in test taking was a possible reason for the
switch.

Students received cues or became confused from poorly written items. For middle and secondary level students, the cues they received generally made the item easier, and more students answered it correctly. Confusion when reading an item caused more students to answer the item incorrectly. In either case, the discriminating power of the item was lessened.

The studies investigated the violations from differing viewpoints. The approaches included teaching test-wise behaviors for better test performance, examining the presence by grade level of certain test-taking strategies when answering fictitious items, and altering previously acceptable (i.e., reliable and valid) items to investigate the violation effect. Likewise, the location of the violation was inconsistent. Authors placed a violation in the correct choice, in the stem, or they did not indicate any specific location. Any attempt to synthesize the results must be done with extreme caution. However, for development of the simulation, a theory was hypothesized. Table 1 summarizes the results of each study.
Table 1. Effects of item writing violations on item indices

<table>
<thead>
<tr>
<th>Violation</th>
<th>Difficulty [level]</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative stems</td>
<td>no effect [grad student] (11)</td>
<td></td>
</tr>
<tr>
<td>Repetition of stem</td>
<td>no effect [college] (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>increases [Army] (5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>increases [11] (7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no effect [12] (10)</td>
<td></td>
</tr>
<tr>
<td>Specific determiners</td>
<td>increases [Army] (5)</td>
<td></td>
</tr>
<tr>
<td>Grammatical inconsistencies</td>
<td>no effect [college] (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>decreases [college] (9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>decreases [Army] (5)</td>
<td></td>
</tr>
<tr>
<td>Variable length</td>
<td>decreases [college] (9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>increases [college] (6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>increases [11] (7)</td>
<td></td>
</tr>
<tr>
<td>Stem format</td>
<td>no effect [Army] (5)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discrimination [level]</th>
<th>Miscellaneous [level]</th>
</tr>
</thead>
<tbody>
<tr>
<td>no effect [college] (4)</td>
<td>no effect on reliability <a href="12">5,7,9,11</a></td>
</tr>
<tr>
<td>no effect [12] (10)</td>
<td>students can recognize and use [6] (3)</td>
</tr>
<tr>
<td>decreases [college] (1)</td>
<td>students can recognize and use cues but not consistently [6] (3)</td>
</tr>
<tr>
<td>decreases [college] (1)</td>
<td>no effect on reliability or validity [Army] (5)</td>
</tr>
</tbody>
</table>

- difference only found when option three times longer [college] (2)
- students can recognize and use cues but not consistently [6] (3)
- significant difference when easy adjacent items have same option as correct choice [college] (2)
DEVELOPMENT OF THE SIMULATION

Recent legislation and educational trends have placed increasing emphasis on the role of evaluation in education. Classroom teachers commonly use objective tests as a major part of their evaluation plan. Yet teachers receive little opportunity in preservice education to practice item writing skills or to receive feedback on each item they construct. The Multiple Choice Test Analyzer (MCTA) was developed to simulate the type of feedback teachers would receive on test items if the items were administered to a class. The simulation encourages teachers to critically analyze and modify their items so that more reliable tests can be produced. This chapter describes the operation and computer implementation of the MCTA.

Description of the MCTA Simulation

The MCTA provides teachers with an opportunity to practice their item writing skills while observing the effects of item violations on students' scores. The MCTA was developed as an aid to be used in conjunction with classroom instruction on item construction. An understanding of item writing principles and the item indices of discrimination and difficulty are a necessary prerequisite for optimum use of
The basic operations of the MCTA are to accept a multiple choice test item supplied by the teacher, analyze it for the presence of selected item writing violations, and output the item indices and other information provided to determine if the item is free from violations. Finally, the teacher has the option of saving the item as constructed, modifying it, or withdrawing it.

Initial Considerations

Prior to development of the MCTA, design criteria were established, an effective computer system was selected, and suitable item writing principles were identified.

Design criteria

The design criteria established for the simulation were the following. First, the simulation should provide an opportunity for teachers to practice the principles of item writing they learned through reading and class lecture. The practice should allow for individual item construction rather than selecting from a group of items. Since teachers need to be able to write multiple choice items for several content areas, the simulation should avoid burdensome restrictions on item content. In this way the greatest concentration could
be placed on a teacher's individual weaknesses.

Secondly, the simulation should provide an outcome for evaluation or judgement. The outcome should not produce specific diagnosis of the item. Instead it should closely resemble the types of evaluative information the teacher would gain from administering test items to an actual class. Thus, the simulation would provide situations common to those the teacher would encounter in self-evaluation.

Thirdly, the simulation should be completed in a short period of time. The response time between the teacher's submission of the item and the evaluative outcome should be very small. When a simulation is interactive in nature, as is provided in a computer system, the response time should appear so negligible that it could be considered an immediate response.

Selection of the computer system

For the flexibility prescribed in the design criteria, a computer system was needed to execute the algorithm. Such a system could provide immediate feedback (at most a ten second delay for the MCTA simulation) of student responses. The evaluative data could be presented for an unlimited number of items while maintaining consistent and objective analyses.

The PLATO IV computer system was chosen to implement the MCTA simulation. The PLATO (Programmed Logic for Automatic
The Teaching Operation system consisted of a Control Data Corporation Cyber 73-4 computer connected by telephone lines to interactive terminals throughout the United States. Input to the system was provided by a keyboard similar to a typewriter and output was displayed on a plasma panel which resembled a television screen. Eastwood and Ballard (1975) gave a complete description of PLATO.

Studies using PLATO as an instructional tool have shown promising results. In a simulation of elementary grade level reading difficulties, Boysen (1976) concluded that the simulation using the PLATO system was a productive learning experience. She compared treatment and control groups and found the treatment group made a significant gain from pre- to posttest, while the gain in scores of the control group who received classroom instruction was not significant. Furthermore, a second posttest, administered four weeks after the first posttest indicated that there was no significant loss of retention by the treatment group during that time.

The PLATO system has been used in other teacher education simulations. Alexander (1976) investigated the effect of an instructional simulation dealing with behavioral objectives, tables of specifications, Bloom's cognitive taxonomy, and lesson planning procedures on application of principles in a microteaching situation. Blake (1973) found an increase in the questioning behaviors of preservice
mathematics teachers after using a program simulating the classroom response to questioning techniques.

**Identification of item writing principles**

Item writing principles commonly described in the literature were examined to identify those that met the content independent design criterion. Several principles which focused on the physical verbal characteristics of items and seemed feasible for implementing on the PLATO system were identified as possible violations to be checked by the simulation.

The studies examined supported the effect of selected principle violations on the test taking performance of fifth through twelfth grade students. From these, four violations were chosen for use in the simulation. The four original violations were:

a. grammatical inconsistency between the stem and one or more of the choices

b. repetition of sounds, words and phrases between the stem and one or more of the choices

c. use of negative words in either the stem or the choices

d. use of specific determiners in the stem or choices.

During development of the MCTA simulation, it appeared that two violations needed to be clarified. The violations
of concern involved grammatical inconsistency and specific determiners. The identification of grammatical inconsistencies in the simulation would require searching through several lists for agreement between the stem and choices. Lists would be required since no general rules could be established to handle singular or plural subject-verb-object agreement, and differentiation between the indefinite articles "a" or "an" for words beginning with "h". The time required to search through such lists would be prohibitive in an interactive situation. Therefore, grammatical inconsistencies were withdrawn as one of the violations considered in the simulation.

Specific determiners are words that describe a condition in the item stem or option which would give a cue to its acceptability. A list of specific determiners was identified. Further examination of the specific determiners (such as: all, always, generally, often, and usually) revealed polar conditions. While several words (all, always, each, every) described conditions which most likely would cause a student to avoid choosing an option, other words (generally, most, often, usually) made the condition more appealing so that students would choose the option more often. Therefore, the original list of specific determiners was divided into two lists reflecting attractiveness of the determiners. The lists, titled "con" specific determiners
for repelling words, and "pro" specific determiners for attracting words, were stored in memory as arrays (Appendix A).

Phases of Operation

After design criteria, computer system and appropriate principles were determined, the MCTA algorithm was developed. The function of the MCTA was to analyze multiple choice test items according to the selected item writing violations. The violations pertained to relationships between certain words used in the items. Since the individual test items were provided by the teacher without restriction to content, the simulation judged only violations which did not relate to the content of the item. Four major phases of operation for the MCTA were: entering and storing the item, analyzing presence of item violations, determining the effect of violations on item indices, and providing teacher action with regard to the item.

**Entering and storing the item**

Each multiple choice item was input in the MCTA by the teacher. The item could have six parts: the stem and each of five choices. The teacher entered each item part separately and stored it before entering succeeding item parts. Each
item part could be up to four lines in length with a maximum of 60 words per part.

After each item part was input, it needed to be stored in memory. The item part was first stored as a very long character string equal to the length of four lines of text. Then each written word was successively copied from the character string into the appropriate row of a 6x60 array (one row for the stem and one for each choice). The two-dimensional array of words was used for all further analyses of the item.

Analyzing presence of item violations

The violations checked in the MCTA simulation involved repetitions, negatives, and "pro" and "con" specific determiners. Values identifying the presence of each violation for all appropriate item parts were calculated.

Check for repetitions Determining repetition between the stem and the options included several steps. First, common words which are used often in communication were considered to be useless as cues for repetition. Secondly, special handling of root words and repetitious adjacent words was necessary. An array of common words (ignore for repetition array) to be ignored when checking repetition between the stem and options was developed (Appendix B). The repetition condition was then checked in the following
manner. The first word in the stem was selected, it was checked against the ignore for repetition array for a match. If a match was found, the word was ignored and the process began again with the next word in the stem. If no match was found, the word was considered a potential cue word and was checked against each option for a match.

Repetition of whole words was only one aspect of the repetition condition. Repetition of sounds and root words was also important. For operation of the simulation, it was decided that identifying the root part of a word from a prefix or suffix would require too much time to analyze. Likewise, the repetition of sounds not appearing in the initial part of a word would take too much processing time. Therefore, only the beginning of words was checked for repetition of sounds or root words. After examination of several three and four letter words and prefixes, it was decided to compare only the first four letters of words to determine if they matched.

Finally, repetition of an adjacent string of words was studied and referred to as a phrase. It was believed that repetition of an n-word phrase in an option should have greater attracting power than repetition of n nonadjacent words in the option. Therefore, a special procedure for identifying and weighting phrases was developed.
After a word in the stem was identified as a potential cue word, the first four letters of the word were checked against each option for a match. If a match was found, subsequent words in both the stem and the option were checked to determine if the match was for a single root word or for a phrase. If the match was a single root word, the repetition value was increased by one; if the match was for a phrase, the repetition value was increased by the number of words in the phrase plus a weighting factor of 0.2 for a two-word phrase and 0.5 for three or more words in a phrase. The program then returned to the next stem word. The new stem word was checked to insure that it had not been previously considered as a cue word, and if not, the process began again (Figures 1 and 2).

A repetition value was calculated for each option in the item. Thus a five choice item would have five repetition values associated with it. Repetition between options had been included as a violation of repetition in the literature. When identical repetition existed in each option, that part of the option should be included in the stem. While the MCTA simulation did not check for this type of repetition, it did disregard the minimum repetition value between stem and each option when the simulation was determining the change in item indices.
Initialize a 6x60 word array, each row containing an item part and each cell containing separate words

Initialize a 1x6 repetition values array to zero; Initialize stem word pointer to one

Should stem word be considered for repetition

No

Increment stem word pointer by one

Yes

Is stem word a word previously used in the stem

No

Do for each option

Yes

Are the first four characters of the stem word repeated in an option word

No

Increment repetition value for that option by one

Yes

Is this the last stem word

Yes

Return to main program

No

Check for a phrase

No

Is this the last option

Figure 1. Flow chart to check for repetition
Figure 2. Flow chart to check for repetitions phrase
Check for negative words  The use of negative words in either the stem or the choices caused students to become confused since the common set of positively stated items was disrupted. Therefore, the presence of negatives was checked in all item parts.

Commonly used words with a negative orientation were identified and stored in the negative array (Appendix C). The list included never, no, none, not and the contraction n't so that words containing the contraction would also be searched. However, if the teacher used a word with the contraction n't but omitted the apostrophe, this word would not be identified. Several words (such as words ending in -ment, -ant, etc.) with the nt letters do not have a negative orientation. Likewise, words beginning with the prefixes a-, irr-, non-, and un- which sometimes connote negative words were not included in the list. In both these cases, no consistent rule could be determined to identify the negatively oriented words from other words containing the same letters and prefixes.

Each word from the negative array was checked against words in each item part. A negative value for each item part was computed as a count of the number of distinct words the item part contained which were also found in the negative array.
Check for specific determiners

Upon examination of specific determiners commonly identified in the literature, it was realized that the determiners could be categorized in two groups—those that attract and those that repel a student from choosing them. Thus two lists were identified and stored in arrays labeled "proSD" and "conSD" (Appendix A). To check for the presence of specific determiners each word from the proSD and conSD lists was compared with each item part for a match. A proSD value and a conSD value were determined by counting the different proSD's or conSD's appearing in each part.

Determining the Effect of Violations on Item Indices

After the teacher entered an entire item in PLATO, it was analyzed with violation values set for repetition in each choice, and negatives, proSD's and conSD's in the stem and in each choice. Then the violation values were entered in a series of equations to determine any change in the item analysis indices. Indices were computed for item difficulty, item discrimination, and test reliability.

Effect of violations on item difficulty

The four violations were categorized as having either attracting or repelling properties. Repetitions and proSD's
were attracting, while negatives and conSD's were repelling. When repetitions or proSD's were present in the correct or key choice, more students should choose that option, and thus the difficulty index would increase. If however, the repetition or proSD was present in a distracter, more students should be attracted to the wrong choice, and the difficulty index would decrease.

Negative words made an item confusing by reversing the frame of reference. In each case where a negative was present, the difficulty index was decreased.

ConSD's were repelling words. When conSD's were present in the key choice, the difficulty index was decreased, but if present in one or more distracters, the difficulty index was increased. More complex procedures were identified to handle the situation where a combination of proSD's and conSD's existed in the several parts of an item, so that the major effect of proSD's and conSD's was maintained as described above.

Item repetition value To determine a repetition value for the item, a weighting procedure was used with the option's repetition values. To avoid exaggerated changes for items whose options all contained a repeated word or phrase, a minimum repetition value for the item's individual options was identified. This minimum amount was then subtracted from each option's repetition value.
Repetition provided an attracting cue. Thus the repetition in the correct option was weighted more heavily than if it appeared in a distracter. This was accomplished by doubling the repetition value in the correct option. The item repetition value was computed by the following formula:

\[ \text{IRV} = \left( \sum_{c=1}^{5} \text{ORV}_c \right) + \text{ORV}_{kc} - (\text{NO} \times \text{MRV}) \]

where \( \text{IRV} \) is the item repetition value
\( c \) is the option
\( \text{ORV}_c \) is the option repetition value
\( kc \) is the key choice or correct option
\( \text{NO} \) is the number of options in the item
\( \text{MRV} \) is the minimum repetition value.

Item negative value A total negative value was computed as the sum of the negative values from each item part. Then the item negative value was converted from the total negative value in the following manner:

if total negatives = 0 1 2 3 or more
then item negative value (INV) = 0 .10 .15 .20

Item specific determiner value Simulating the total effect of specific determiners was the most complex part of

---

1 Asterisk is used to denote multiplication.
the algorithm. Because of the nature of the specific determiners either to attract or repel, the location of the specific determiner in the item (key choice, distracters or stem) was critical. Secondly, the cumulative effect of different possible combinations of proSD's and conSD's in the same option provided a considerable amount of conjecture.

The occurrence of proSD and conSD words in the correct option and distracters seemed to provide opposite cues, and the occurrence of both a proSD and conSD word in any item part yielded more intense cues. The following procedure was established to handle the specific determiner condition:

For distracters and the stem, if both a proSD and a conSD occurred in the same item part, then the original proSD value taken as the count of different proSD's occurring in the item part was weighted, and the conSD value was changed to zero.

This procedure was chosen since it was believed that in many cases where a proSD word and a conSD word exist in the same statement, the occurrence of the conSD identified any specific exception to the proSD general rule; therefore the general rule appeared stronger. The effect of proSD and/or conSD in the stem was believed to have a different effect than if the specific determiners occurred in the distracters. Therefore the item specific determiner value took into effect the special attribute of determiners in the stem. The
item specific determiner value was computed according to the following procedure:

If both a proSD and a conSD occurred in the same distracter, or in the stem, then PSDV was increased by $0.4 \times CSDV$, and CSDV was returned to zero.

where PSDV is the proSD value

$ip$ is the item part

CSDV is the conSD value.

Otherwise the proSD values and conSD values remained as originally computed. The item specific determiner value followed the general formula:

$$ISDV = 1.2 \times \left( \sum_{ip=0}^{5} PSDV - PSDV \right) - \sum_{ip=0}^{5} CSDV$$

$$- CSDV + 1.3 \times CSDV - 3 \times PSDV$$

where ISDV is the item specific determiner value

$ip$ is the item part

PSDV is the proSD value

$kc$ is the key choice or correct option

CSDV is the conSD value

$s$ is the stem.

An item specific determiner weighted factor was computed according to the presence of proSD's or conSD's in the correct choice. The weighted factor was computed as follows:
1) If neither proSD nor conSD occurred in the correct choice, then

\[ ISDWF = 0.04(-3*I_{SDV}) \]

where \( ISDWF \) is the item specific determiner weighted factor
\( I_{SDV} \) is the item specific determiner value.

2) If only a conSD occurred in the correct choice, then

\[ ISDWF = 0.04(-3 -4*I_{SDV}) \]

where \( ISDWF \) is the item specific determiner weighted factor
\( I_{SDV} \) is the item specific determiner value.

3) If a proSD or both a proSD and a conSD occurred in the correct choice, then

\[ ISDWF = 0.04(4* -2 * I_{SDV}) \]

where \( ISDWF \) is the item specific determiner weighted factor
\( I_{SDV} \) is the item specific determiner value.

These coefficients and formulas for determining the effect of specific determiners were developed as a result of inspecting numerous cases with varying numbers of proSD's and conSD's and their anticipated effects. Arbitrary weights were assigned to make the results appear more realistic.

**Summary effect of violations on difficulty** After item values were computed for repetitions, negatives and specific determiners, the effect on the difficulty index could be determined. The change in difficulty was computed as
\[ DC = RF - INV + ISDWF + \frac{---}{30} \]

where DC is the change in difficulty
RF is a random factor from -0.03 to +0.03
INV is the item negative value
ISDWF is the item specific determiner weighted factor
TRV is the total repetition value.

**Effect of violations on item discrimination**

For purposes of this simulation, an occurrence of an item writing violation decreased the discrimination index. The reasoning for this principle was that any violation which gave students unnecessary cues to answering the item would cause confusion to all students, including the students with high ability. Therefore, fewer of the high ability students would answer the item correctly. The change in discrimination was computed as follows:

\[ Disc = RF - INV - |ISDWF| - \frac{---}{30} \]

where Disc is the change in discrimination
RF is a random factor from -0.03 to +0.03
INV is the item negative value
\(|ISDWF|\) is the absolute value of the item specific determiner weighted factor
TRV is the total repetition value.
A procedural rule of the MCTA simulation held that items must have at least three choices. When an item had three or more choices, procedures for computing difficulty and discrimination indices remained the same regardless of the number of choices (Costin, 1970, 1972; Tversky, 1964). However, for items with less than three choices, the difficulty and discrimination indices were set to -1.99 which was below any possible computed index value. The simulation displayed "**" where the indices were normally printed, and an error message was given.

**Effect of violations on test reliability**

The reliability of a 50 item test which contained the items submitted by the teacher would change in proportion to the computed change in difficulty. For items submitted by the teacher, and subsequently saved for the final test, the reliability was computed as:

\[ R = BR + \frac{50}{IS} + \sum_{I=1}^{IS} \frac{DC}{50} I \]

where \( R \) is the reliability

\( BR \) is the base reliability (base level for the simulation)

\( IS \) is the number of items saved

\( I \) is an item

\( DC \) is the difficulty change.
Additional Features

The basic operations of the MCTA were a) to enter and store the item, b) to analyze the item for violations and c) to determine the effect of violations of item indices. Considerations for additional features were made in the following areas: grade level influence, teacher aids for interpreting item indices, percentage response for each option, and final teacher action on each item analyzed.

Grade level influence

In accordance with the literature, the effect on the difficulty of an item for the occurrence of specific determiners and grammatical inconsistencies was influenced by grade level. However, incorporating this gradation of effect would only lessen the impact of the simulation to identify a violation for items written for fifth through eighth grade students. Therefore, grade level effect was not included in the final MCTA simulation and teachers were instructed to write their items for secondary level students.

Supplementary aids provided to interpret analysis of the item

The MCTA provided the difficulty index and discrimination index for each item the teacher analyzed. The
teacher was then expected to interpret this information to determine if the item appeared to be a good one. Additional aids for interpreting the item indices were provided by a description of the index terms "reliability", "difficulty" and "discrimination" in both a concise sentence and a paragraph form, and the identification of acceptable levels for the indices and base levels for a good item. The acceptable levels for indices specified by the MCTA were:

<table>
<thead>
<tr>
<th>Index</th>
<th>Acceptable Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>test reliability</td>
<td>0.65 to 1.00</td>
</tr>
<tr>
<td>item difficulty</td>
<td>0.45 to 0.70</td>
</tr>
<tr>
<td>item discrimination</td>
<td>0.20 to 0.45</td>
</tr>
</tbody>
</table>

The base levels for good items with no violations were:

<table>
<thead>
<tr>
<th>Index</th>
<th>Base Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>test reliability</td>
<td>0.85</td>
</tr>
<tr>
<td>item difficulty</td>
<td>0.60</td>
</tr>
<tr>
<td>item discrimination</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Pyrczak (1972) stated that indices of choice attractiveness were apparently more helpful in identifying faulty items than the item difficulty level. As an indicator of choice attractiveness, the teacher also could request the students' percentage response to each option. Such information could indicate the more or less plausible option from the others.
Percentage response to each choice

Using techniques similar to those identified for determining difficulty level, the percentage of students responding to each item choice was computed. The percentage response to the correct option was equal to the difficulty index. When no violations were present in the distracters, the percentage response for each distracter was equally distributed from the remaining percentage. However, if distracters contained attracting words, they received a larger amount of the remaining percentage, while distracters with repelling words received a smaller amount of the remaining percentage. The teacher then could identify options which contained violations according to the variance in option response rate. The distribution of response for distracters containing violations was computed according to the following procedures.

Violation switches were provided for violations in each option. If a negative violation were present in an option, the negative violation switch was set to one; if no negative violation existed, the switch was set to zero. Similar rules were followed for proSD, conSD and repetition violation switches for each option. Then total negative, proSD, conSD and repetition values were computed by summing the switch values for all options.
For each distracter, if no violations existed the response factor for the distracter was one. If only one violation existed in the distracter the response factor was calculated as follows:

If the violation were negative, \( R_{esF} = \frac{0.4}{NV} \)

If the violation were proSD, \( R_{esF} = 1 + 0.75*PSDV \)

If the violation were conSD, \( R_{esF} = 1 + 0.25*CSDV \)

If the violation were repetition, \( R_{esF} = 1 + \frac{RV}{3} \)

where \( R_{esF} \) is the response factor
\( d \) is the distracter
\( NV \) is the negative value
\( PSDV \) is the proSD value
\( CSDV \) is the conSD value
\( RV \) is the repetition value.

When more than one violation type existed in the distracter, the response factor was calculated as follows:

1) If conSD and negative existed, \( R_{esF} = 0 \)

2) If repetition and conSD and/or negative existed together, \( R_{esF} = 1 + \frac{RV}{3} - 0.75*CSDV - 0.6*NV \)

3) If only a proSD and negative existed, \( R_{esF} = 0.75*PSDV \)

4) If proSD and repetition existed, with or without a
negative, \[ \text{ResF} = 2 + 0.75 \times \text{PSDV} + \text{RV} \times 3 - 0.6 \times \text{NV} \]

where \( \text{ResF} \) is the response factor
\( \text{EV} \) is the repetition value
\( \text{d} \) is the distracter
\( \text{CSDV} \) is the conSD value
\( \text{NV} \) is the negative value
\( \text{PSDV} \) is the proSD value.

Finally, if any distracter's response factor was a negative value, it was replaced by zero.

The response factors for each option were summed for a total response factor. Then this amount was divided into the percentage of response remaining for the distracters to give the common response amount. Finally the response factor for each distracter was multiplied by the common response amount for the percentage response to each distracter.

**Final teacher action with regard to each item**

After the teacher had the item analyzed, a choice of action was made. The teacher could save the item in an item bank, thereby affecting the final test reliability; the item could be deleted so that it did not affect the final test reliability; or the item could be modified to improve the item indices. If the teacher chose to modify the item and have it analyzed again, then he or she could request that the item violations be identified. The violation was not
immediately given to the teacher, instead the teacher was encouraged to find the violation from results of item indices and student response rates. If after modifying the item it still contained errors, the teacher could request to see what the stated violation was.

Summary

The MCTA simulation provided feedback on test items similar to the information a teacher would receive after administering test items to a secondary level class. The teacher would type in an unlimited number of test items one at a time to be analyzed for presence of item writing violations. Counts were made for the presence of repetition between the stem and one or more options, and for presence of negatives, pro specific determiners, and con specific determiners occurring in either the stem or the option. Specific procedures for identifying the violations follow.

Repetitions were identified when the first four letters in a word found in the stem were also found as the first four letters of a word in an option. When two or more repetitions were found as adjacent words in both the stem and option, the group was referred to as a phrase. The value of the phrase was weighted greater than the number of words in the phrase. When each option contained a repetition value greater than
zero, the minimum repetition value was subtracted from each option's repetition value. In this way, the effect of a repetition would not be inflated if it was necessary for each option to contain a word found in the stem.

Negatives were identified by matching a word in an item part (stem or option) with a list of words identified as negative or with any contraction ending in *'t* (Appendix C). The negative value contained the count of each different negative word found in an item part.

Pro and con specific determiners (proSD's and conSD's) were listed (Appendix A). Counts were made for each different proSD and conSD occurring in an item part. When both a proSD and conSD appeared in the same item part, the proSD and conSD values were adjusted depending on the item part being the stem, correct choices, or distracter.

Formulas were generated to identify the effect of the violations on base levels for item difficulty and item discrimination indices, and for test reliability. Although factual evidence regarding the extent of influence on item indices was limited, quantitative values were inferred from the apparent direction of violation effect. From the base level of 0.60, the item difficulty index was increased when an attracting cue was present, and was decreased when a repelling cue was present. Attracting cues consisted of repetition or proSD's in the correct choice, and negatives or
conSD's in the distracters. Repelling cues consisted of repetition or proSD's in the distracters, and negatives or conSD's in the correct choice. If negatives or specific determiners were found in the stem it was believed that they would tend to confuse the student, so the item difficulty index was decreased.

In every case the item discrimination index was decreased from the base level of 0.40 when a violation was present in the item. The effect of violations on test reliability was computed from a base level of 0.85. The change was computed as a function of the change in difficulty index.

Additionally, the choice attractiveness was determined as the percentage response anticipated for each option. The percentage remaining after the item difficulty index was subtracted from 1.00 was apportioned to each option in accordance with the presence of attracting or repelling violation cues.
In addressing the future role of computers in the instructional process, Molnar (1976b) defined the challenge for the instructional designer in the following manner:

The task of the instructional designer is to design interesting environments, build powerful artifacts and facilities and create rich information fields to cover a broad problem domain. The focus is on process and not product, on formative learning not repetition. The problem environments should be deep and reusable (p. 28).

It was the design intent that the Multiple Choice Test Analyzer (MCTA) would conform to Molnar's philosophy.

Participation in the MCTA occurred after classroom instruction on item writing and item indices. Teachers submitted their test items for analysis by the MCTA. The main procedures of the MCTA to store, analyze and interpret the items occurred only on the teacher's request. Therefore the experience a teacher received from the MCTA was not regulated by the program but rather it occurred as individual and self-directed learning.

As the result of experiencing the MCTA simulation in their training program, it was hypothesized that teachers would be better able to utilize test construction skills and
avoid common item writing violations which would result in constructing more valid and reliable classroom tests. It was assumed that improvement in item writing ability could be measured by a carefully constructed multiple choice test addressing application level objectives. It was also assumed that exposure to the simulation from one to two hours would be sufficient to produce a measurable difference.

The experimental design used to test the effectiveness of the MCTA was a four group design developed by Solomon (1949) as cited in Campbell and Stanley (1969). For the classes under investigation, teachers (preservice education students) were randomly assigned to one of the four groups. Two of the groups (groups 1 and 3) received a pretest, and two of the groups (groups 1 and 2) used the MCTA simulation. The investigator developed the 45 item instrument used as the pre-post test to measure teachers' ability to analyze effects of item writing violations on multiple choice test items. The posttest scores, obtained from all groups participating in the study, were used as the dependent variable. In addition to the experimental group status, the relationship to posttest scores of typing skill, number of items analyzed by the simulation and counts of the number of analyzed items containing repetitions, negatives, pro specific determiners and con specific determiners were computed using correlation procedures.
This chapter discusses the method of procedure for the study. The chapter begins with a description of the teacher's participation in the simulation, followed by the design of the study, administration of the experiment, development of the measurement instruments, and description of the dependent and independent variables.

Teacher's Participation in the Simulation

The MCTA was designed to supplement existing preservice instruction by evaluating an unlimited number of test items constructed by the preservice teacher. The items were analyzed according to the presence of selected item writing violations and group performance of secondary level students on those items was simulated. A description of the specific participation of teachers in the simulation follows.

When a teacher signed on the PLATO terminal for the first time, he or she completed an introductory lesson describing the operation of the keyboard with its special keys. This lesson took approximately 10 to 15 minutes to complete. Then, for purposes of this study, an exercise was conducted to determine the teacher's typing skill. After these preliminary procedures were completed, the teacher began the MCTA simulation.
At the beginning of the simulation, the teacher was routed to introductory material where an advance organizer was provided and an explanation of objectives for the lesson was described. Following this introduction, instructions for entering the items were given, and the teacher completed a practice activity to insure that he or she understood the use of the keys for entering an item. If the teacher were returning to the terminal after the initial session, these procedures could be bypassed. Finally the teacher came to the Multiple Choice Editor where he or she could enter an item, one part at a time. Figure 3 illustrates the teacher's view of the Editor after the entire item has been entered.

When the entire item was stored in the system, it was necessary to indicate the correct choice and instruct the HCTA to analyze the item. After the item was analyzed, the teacher received the estimated difficulty and discrimination indices for the item, and the reliability of a 50-item test containing the item analyzed. Finally, the teacher chose an outcome for the item. If the teacher felt the item was a good one, the item indices were saved to become part of the final group of test items used for determining test reliability. If the item was a poor one, the teacher would either delete it, or return to the Editor to modify it (Figure 4).
The type of test item which contains a beginning statement followed by several alternative options is

CHOICE A
completion

CHOICE B
matching

CHOICE C
multiple choice

CHOICE D
true-false

CHOICE E

Figure 3. Illustration of complete item entered in the Editor
TEST ITEM ANALYSIS

For a fifty item test, if your item were one of the fifty, the test would have a reliability of .85
an item difficulty level of .63
and an item discrimination level of .38

If you are satisfied with the analysis of your item
and would like to include it in your item bank, press
(Shift) Back

If you would like to modify your item to improve the item indices, press Next

If you would rather delete this item, and begin with a new one, press (Shift) Lab

To help you decide your action on this item, consider the options below:

Data Explanation of error on this item
Help Acceptable ranges
Lab Explanation of analysis terms
(Shift) Help Percent response to each choice

Figure 4. Illustration of initial analysis of item by MCTA
To help determine whether an item should be saved, modified or deleted, four help units were provided. The teacher could request a concise sentence description of item indices with the acceptable ranges for each index. For a more detailed description of the indices, more text in paragraph form was available at the teacher's request. Base levels for a "good" item were also provided. The third help unit provided an indication of choice attractiveness showing the percentage of students choosing each option. Each of these help units was available when the teacher took the initiative to request them. A notation of the help unit's availability was provided within the program, but teachers would not automatically see these units without specifically requesting the unit desired.

Lastly, a help unit describing violations found in the item could be provided conditionally. After the teacher had attempted to eliminate the violations by modifying the item, a request for a description of the violations present could be made. Directions for this unit were only available after an item modification had taken place. Teachers attempting to receive the information before modifying the item were told to try to remedy that item before the specific diagnosis was given. This method was used to encourage teachers to identify for themselves the violation using information commonly available to them as classroom teachers. In this
way, it was hoped that the unit would not be used in the simulation indiscriminately as might occur if the unit were always available.

These procedures for entering, storing, and analyzing items continued in a loop until the teacher requested a final summary of his or her progress. The final summary identified the number of items saved and the resultant test reliability. The summary also included a count of the number of items containing each violation (Figures 5 and 6).

Pretest of the MCTA

The MCTA was pilot tested with two groups, a graduate and an undergraduate class during the Spring term, 1977, and was critiqued by faculty who specialized in areas of measurement and evaluation. The 11 class members in Home Economics Education 515, Evaluation in Home Economics, Spring 1977, used the simulation as an integral part of their instruction on test construction. Each teacher tried ten items on the MCTA. The items then became part of a test development assignment for the class. After completing the simulation, the teachers answered several free response questions concerning the operation of the simulation (Appendix D). They indicated a generally positive reaction to the MCTA, and identified any problems they encountered.
Figure 5. Flow chart for teacher action in the MCTA simulation
Teacher enters an item one part (stem or choice) at a time

Identify the key choice and analyze the item

Show item indices for item difficulty, discrimination and test reliability

Go to teacher action

Figure 5. Continued
Teacher action

Is help needed to interpret item indices

No

Inform teacher to try to modify item before information is given

Yes

Show list of violations found

Has teacher tried to modify this item

No

Paragraph explanation of item indices and identification of base levels for "good" item

Identify acceptable ranges and concise definition of item indices

Show percent of students responding to each choice

Request list of violations found

Go to key

Figure 5. Continued
Modify the Item

Return to Editor with existing item and make modifications

Delete the Item

Store item indices for use in determining final test reliability

Save the Item

Clear storage in Editor and return to enter new item

Go to Editor

Go to final summary

Take some action on the item with respect to indices

Leave the Lesson

Figure 5. Continued
Calculate and print out final test reliability and summarize violations.

Want to try more items?

- Yes
  - Go to Editor

- No
  - Store teacher's summary variables for later use by investigator
  - Exit MCTA lesson

Figure 5. Continued
EVALUATION

During this session on PLATO, you analyzed 4 items. Of these, you saved 0 items. If you wrote a fifty item test with all items similar to the 0 items you saved, it would have a test reliability of .00.

Of the total 4 items you analyzed, the following item principles were violated:

- 0 items contained negative words such as never, no, not, none, and contractions in the stem or choices.
- 0 items contained attracting specific determiners such as but, except, frequently, generally, most, often, seldom, some, sometimes and usually in the stem or choices.
- 0 items contained repelling specific determiners such as absolutely, all, always, any, each, every, impossible, only, whenever, and wherever in the stem or choices.
- 0 items contained repetition of root words, words, or phrases between the stem and one or more of the choices.

Back to try more items
Shift + Stop to leave lesson.

Figure 6. Illustration of final MCTA summary
Subsequent revisions in storing the stem, and introductory verbal instructions were made.

After the MCTA was revised following the pilot test with graduate students, it was pilot tested with a group of eight Psychology 333, Educational Psychology volunteers. Forms for using human subjects were filed with the Psychology office, and the volunteers received one hour credit toward an increase in points on their final course grade. Approximately one to two weeks before participating in the pilot test, the psychology class had studied test construction so the volunteers were somewhat familiar with item writing skills. At the completion of the simulation, the volunteers also answered the free response questions concerning the MCTA. Their responses failed to indicate a need for further revision to the operation of the simulation.

Finally, the MCTA was individually demonstrated to five Iowa State University faculty members who are competent in the area of tests and measurements. They were encouraged to try several combinations of violations to determine the predicted effect on item indices. Acceptance of the simulation model by these experts was interpreted as establishing the usability of the model.
Research Design

The effect of the MCTA was investigated using a Solomon Four-Group experimental design. A pre-post test instrument used to measure teacher's knowledge of the effects of item writing violations was developed by the author. Separate pilot tests of the simulation and the instrument were made. Finally, the experiment was conducted twice, once with undergraduate students, and again with graduate students.

Undergraduate experiment

Students enrolled in Psychology 333 during the First Summer Session, 1977, hereafter referred to as the undergraduate educational psychology class, were the subjects for this experiment. The 52 preservice teachers were randomly assigned to four groups according to level and major. Groups 1 and 3 were administered the pretest, and groups 1 and 2 received the treatment (Table 2). Teachers receiving the treatment completed the one to one-and-one-half hour MCTA simulation independently within a five day period of time. After the treatment was completed, all four groups were administered the posttest.

The initial intent of this study was to investigate the effectiveness of the MCTA simulation in comparison to common classroom instruction. In the undergraduate class, the
common practice was to receive some initial instruction in test construction, and to spend a laboratory period in constructing items while a teaching assistant was present to help the teachers in constructing and evaluating their items.

However, the original structure of the design was altered due to the desire of the educational psychology instructor to have all preservice teachers use the MCTA as part of their class requirements. Therefore, the MCTA experience took the place of the laboratory and followed an introductory lesson by the instructor on writing test items, and computing and using item indices. The actual experiment included using the MCTA after the posttest for groups 3 and 4 with the resulting research design diagrammed in Table 2.

**Graduate experiment**

To determine the effectiveness of the MCTA with a more experienced group of teachers, 26 teachers who were enrolled in Home Economics Education 515, Evaluation in Home Economics, hereafter referred to as the graduate evaluation class, also participated in this study. The research design was the same Solomon Four-Group adaptation described in Table 2. Teachers were randomly assigned to four groups, stratified according to years since they received their bachelor's degree and amount of their graduate school experience. Groups 3 and 4 received the treatment after the posttest was administered.
Table 2. Solomon Four-Group experimental design as used in the study

<table>
<thead>
<tr>
<th>Group</th>
<th>Pretest</th>
<th>Treatment</th>
<th>Posttest</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>X</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0</td>
<td>2</td>
<td>X</td>
</tr>
</tbody>
</table>

¹Subjects are randomly assigned to groups.
The MCTA treatment followed instruction, which all class members received, by the home economics instructor on writing test items, computing and interpreting item indices. For the graduate level, more detailed instruction was given on concepts of difficulty, discrimination, and reliability. Teachers were encouraged to use items with the MCTA that they could later incorporate in a test development assignment.

Administering the experiment

At the beginning of the experiment, the investigator attended both classes and gave instructions to the subjects. The assignments of subjects to each group were made, and an introduction to the study was given. A four page sheet of instructions (Appendix E) was distributed. The first page of instructions described the location of the PLATO terminals and the teacher's expected preparation prior to using the simulation. They were instructed to write 15 test items to use with the MCTA simulation. Multiple choice items appropriate for the MCTA had at least three options, were verbal in nature (did not include numeric or chemical notation in the stems or options, and did not use dates or numbers for the options) and were written for secondary level students. The second page described procedures for signing-on the system. The last two pages outlined the use of specific keys to enter the items in the Multiple Choice
Editor. The description of keys also was included within the simulation.

A general introduction to the purpose of the study was given to the participants. They were told that the simulation checked for violations in test items. The specific violations the MCTA checked were not explained however. After the introduction, groups 1 and 3 remained to take the pretest, and groups 2 and 4 were dismissed. Approximately one week later, the investigator returned to administer the posttest to all participants.

At the conclusion of the experiment with the graduate level teachers, after all teachers had completed the simulation, the investigator returned to the classroom and discussed the simulation, the nature of the experiment and instruments used with the teachers. The teachers expressed continued interest in the simulation and the posttest. The discussions did indicate that several teachers believed the MCTA was providing true item indices, instead of postulated results.

Pre-Post Test Instrument

The instrument used to measure the effect of the MCTA simulation was developed by the author. It consisted of 45 multiple choice items designed to measure comprehension,
application and analysis of the effects of item writing violations on the difficulty and discrimination indices of multiple choice test items. Both the pretest and the posttest used the same instrument (Appendix F).

Several application and analysis level items contained illustrative partial or complete test items which teachers were to use in responding. In some cases, teachers were asked to choose the option a student might choose if he did not know the answer. At other times, teachers were to identify options which contained violations, or to predict the direction of change for the difficulty of an item when an option was reworded. To maintain content validity of the instrument, careful selection of illustrative test items was made. The content of the items was of a generally unfamiliar nature to teachers to insure that they were responding to the instrument's directions, rather than answering for content of the illustrative items.

The instrument measured three major content areas. The first 13 items tested the comprehension and application levels of aspects relating to item indices. The next 22 items measured identification of cues occurring in items and their attracting or repelling power. The final ten items asked the directional change in the difficulty index of items containing attracting or repelling cues over items containing no cues.
The instrument was checked for content validity by three measurement experts. Then it was pilot tested with 13 volunteer Psychology 333 students during the Spring 1977 term. For the Psychology 333 students who volunteered in the pilot testing, appropriate forms for using human subjects were filed with the Psychology office, and the volunteers received a one hour credit toward an increase in points on their final grade. The volunteers had previously received instruction on test construction so their participation could be viewed as what might be expected on posttest achievement scores by the control group.

Item analyses were performed on data from the 13 Psychology 333 volunteers. The reliability estimate as calculated by the Kuder-Richardson Formula 20 was 0.84. Item analysis data for the pilot testing of the instrument are presented in Table 3. Best items are determined by: a) a discrimination index (computed as item score correlation) equal to or greater than 0.20 unless the standard deviation is less than 0.20, in which case the discrimination index must be greater than 0.40, b) an item difficulty index between 0.30 and 0.70 and c) distracters receiving one or more responses when 50 respondents answered the item (Henne, 1976). Since only 13 respondents were used in the pilot test, best items were selected on the first two criteria.
Table 3. Item analysis data for instrument used in pilot testing

<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.85</td>
<td>0.31</td>
<td>0.36</td>
<td>1 1 11^2 0</td>
</tr>
<tr>
<td>2^5</td>
<td>0.38</td>
<td>0.73</td>
<td>0.49</td>
<td>4 1 5^2 3</td>
</tr>
<tr>
<td>3^3</td>
<td>0.92</td>
<td>0.33</td>
<td>0.27</td>
<td>0 0 1 12^2</td>
</tr>
<tr>
<td>4^5</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>3 1 6^2 2</td>
</tr>
<tr>
<td>5^1</td>
<td>0.77</td>
<td>0.34</td>
<td>0.42</td>
<td>2 10^2 1 0</td>
</tr>
<tr>
<td>6^3</td>
<td>0.85</td>
<td>0.36</td>
<td>-^-4</td>
<td>1 0 11^2 1</td>
</tr>
<tr>
<td>7^5</td>
<td>0.69</td>
<td>0.50</td>
<td>0.46</td>
<td>2 2 0 9^2</td>
</tr>
<tr>
<td>8^3</td>
<td>0.15</td>
<td>0.36</td>
<td>-^-4</td>
<td>2^2 3 2 6</td>
</tr>
<tr>
<td>9^5</td>
<td>0.62</td>
<td>0.59</td>
<td>0.49</td>
<td>0 8^2 4 1</td>
</tr>
<tr>
<td>10^5</td>
<td>0.46</td>
<td>0.64</td>
<td>0.50</td>
<td>6^2 6 0 1</td>
</tr>
<tr>
<td>11^5</td>
<td>0.54</td>
<td>0.69</td>
<td>0.50</td>
<td>1 1 4 7^2</td>
</tr>
</tbody>
</table>

^1Items that would probably meet item analysis criteria if respondent size equaled 50.

^2Indicates correct answer.

^3Items that need revision due to distracter analysis, discrimination index, or difficulty level.

^4Discrimination index less than 0.05.

^5Items that meet the item analysis criteria:
A difficulty index between 0.30 and 0.70 and a discrimination index greater than or equal to 0.20.
<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>132 0 0 0</td>
</tr>
<tr>
<td>135</td>
<td>0.38</td>
<td>0.61</td>
<td>0.49</td>
<td>7 0 52 1</td>
</tr>
<tr>
<td>141</td>
<td>0.62</td>
<td>0.13</td>
<td>0.49</td>
<td>82 2 3 0</td>
</tr>
<tr>
<td>155</td>
<td>0.69</td>
<td>0.70</td>
<td>0.46</td>
<td>3 92 1 0</td>
</tr>
<tr>
<td>165</td>
<td>0.38</td>
<td>0.56</td>
<td>0.49</td>
<td>2 52 6</td>
</tr>
<tr>
<td>171</td>
<td>0.23</td>
<td>0.78</td>
<td>0.42</td>
<td>8 2 32</td>
</tr>
<tr>
<td>181</td>
<td>0.23</td>
<td>0.78</td>
<td>0.42</td>
<td>32 2 8</td>
</tr>
<tr>
<td>193</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>132 0 0 0</td>
</tr>
<tr>
<td>203</td>
<td>0.92</td>
<td>0.29</td>
<td>0.27</td>
<td>0 1 0 122</td>
</tr>
<tr>
<td>215</td>
<td>0.62</td>
<td>0.25</td>
<td>0.49</td>
<td>1 82 1 3</td>
</tr>
<tr>
<td>225</td>
<td>0.69</td>
<td>0.36</td>
<td>0.46</td>
<td>92 2 2 0</td>
</tr>
<tr>
<td>235</td>
<td>0.46</td>
<td>0.57</td>
<td>0.50</td>
<td>1 4 62 2</td>
</tr>
<tr>
<td>243</td>
<td>0.77</td>
<td>0.53</td>
<td>0.42</td>
<td>0 102 0 3</td>
</tr>
<tr>
<td>251</td>
<td>0.54</td>
<td>0.13</td>
<td>0.50</td>
<td>0 4 72 2</td>
</tr>
<tr>
<td>263</td>
<td>0.85</td>
<td>0.36</td>
<td>0.36</td>
<td>2 112</td>
</tr>
<tr>
<td>273</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>132</td>
</tr>
<tr>
<td>285</td>
<td>0.69</td>
<td>0.65</td>
<td>0.46</td>
<td>92 4</td>
</tr>
<tr>
<td>Item number</td>
<td>Difficulty index</td>
<td>Discrimination index</td>
<td>Standard deviation</td>
<td>Distracter analysis</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>--------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>29³</td>
<td>0.77</td>
<td>0.15</td>
<td>0.42</td>
<td>0 10² 3</td>
</tr>
<tr>
<td>30³</td>
<td>0.62</td>
<td>0.15</td>
<td>0.49</td>
<td>3 8² 2</td>
</tr>
<tr>
<td>31³</td>
<td>0.31</td>
<td>0.15</td>
<td>0.49</td>
<td>8² 0 5</td>
</tr>
<tr>
<td>32⁵</td>
<td>0.46</td>
<td>0.23</td>
<td>0.46</td>
<td>7 2 4²</td>
</tr>
<tr>
<td>33¹</td>
<td>0.69</td>
<td>0.08</td>
<td>0.50</td>
<td>6² 4 3</td>
</tr>
<tr>
<td>34¹</td>
<td>0.23</td>
<td>0.72</td>
<td>0.46</td>
<td>2 2 9²</td>
</tr>
<tr>
<td>35¹</td>
<td>0.31</td>
<td>0.19</td>
<td>0.42</td>
<td>3² 4 6</td>
</tr>
<tr>
<td>36⁵</td>
<td>0.38</td>
<td>0.59</td>
<td>0.46</td>
<td>4² 9 0</td>
</tr>
<tr>
<td>37¹</td>
<td>0.23</td>
<td>0.68</td>
<td>0.49</td>
<td>8 5² 0</td>
</tr>
<tr>
<td>38¹</td>
<td>0.23</td>
<td>0.78</td>
<td>0.42</td>
<td>9 3² 1</td>
</tr>
<tr>
<td>39³</td>
<td>0.23</td>
<td>0.78</td>
<td>0.42</td>
<td>3² 10 0</td>
</tr>
<tr>
<td>40⁵</td>
<td>0.54</td>
<td>0.24</td>
<td>0.50</td>
<td>7² 6 0</td>
</tr>
<tr>
<td>41³</td>
<td>0.46</td>
<td>0.15</td>
<td>0.50</td>
<td>7 6² 0</td>
</tr>
<tr>
<td>42⁵</td>
<td>0.31</td>
<td>0.50</td>
<td>0.46</td>
<td>4 5 4²</td>
</tr>
<tr>
<td>43¹</td>
<td>0.38</td>
<td>0.08</td>
<td>0.49</td>
<td>5 5² 3</td>
</tr>
<tr>
<td>44³</td>
<td>0.54</td>
<td>0.15</td>
<td>0.50</td>
<td>2 4 7²</td>
</tr>
<tr>
<td>45⁵</td>
<td>0.08</td>
<td>0.38</td>
<td>0.27</td>
<td>1² 12 0</td>
</tr>
</tbody>
</table>
Of the 45 items, 18 were indicated as best items. At least one of the two criteria were met by an additional 12 items. Rewriting of items was not recommended until additional item analysis data are available as item analysis data tends to stabilize when n=100 or more.

Variables

The variables considered in the investigation included dependent, independent, and confounding types. The total posttest score served as the dependent variable. Counts for number of items analyzed and types of violations occurring in an item were obtained from the simulation. These counts along with the amount of time spent on the simulation were the independent variables. Variables which may have produced a confounding effect were typing skill and total class achievement measured by total class points.

Each teacher was given an individual sign-on to use with the MCQA program. Several measures were compiled on each teacher as he or she performed the simulation. These measures served as independent or covariate variables in the study. The compiled measures included time to complete one simulation, number of items analyzed, number of analyzed items containing negatives, analyzed items containing repetitions, analyzed items containing procedure's, analyzed
items containing conSD's, test reliability for the items saved, and a typing speed measure. Additionally, the pretest score was used to determine the teacher's prior knowledge.

Costar (1975) investigated the possibility of screening students prior to using a PLATO lesson to identify those who might benefit from PLATO from those who should use an alternative learning approach. An Intellectual Screening Assessment was developed for use as a preliminary screen to "bump off" of the PLATO system a student who may have difficulty with the system. In a similar way, typing speed was measured before teachers began the MCTA simulation, with the thought that teachers who could not type with some proficiency might not be suited for using this PLATO lesson. Typing may have been a hindering factor to teachers' achievement since the form for communicating with the PLATO system was through typing, and the nature of the MCTA required a substantial amount of typing as input to the system.

 Treatment of Data

For analysis of data, the two classes studied, undergraduate educational psychology and graduate evaluation, were treated as replication studies. The main hypothesis for the study was
There is no significant difference in posttest scores measuring the ability to write test items free from selected violations between preservice teachers using the computer simulation and those not using the simulation.

The analysis of variance statistic was used to test the hypothesis.

A partial correlation was performed to adjust for the confounding influence of typing skill while examining the relationship between simulation related variables and using the MCTA. Additional analyses included item analysis of the instrument and correlation of the number of items analyzed on the MCTA by total posttest score. Frequencies on remaining variables were also computed.

Summary

The investigation of the MCTA was conducted using an experimental research design. The item pre-post test instrument and the simulation were validated by experts in the field and through pilot testing. The experiment was conducted on two classes, one at the undergraduate and one at the graduate level using a Solomon Four-Group design. Data were collected from the test instrument and from compilations made while using the MCTA. Each class was treated for
analysis as a separate study. Appropriate statistics included analysis of variance, correlations and item analysis of the instrument.
FINDINGS AND DISCUSSION

The purpose of the present study was to investigate the effectiveness of the Multiple Choice Test Analyzer (MCTA) simulation for improving teacher's ability to write multiple choice test items. The MCTA effectiveness was measured by posttest scores on a 45 item multiple choice instrument developed by the investigator. Additional data were collected from the simulation on variables which may have influenced the effectiveness of the MCTA.

The main hypothesis for the study, stated in the null form, was:

There is no significant difference in posttest scores measuring the ability to write test items free from selected violations between preservice teachers using the computer simulation and those not using the simulation.

Subjects for the study were 45 undergraduate educational psychology students and 22 graduate evaluation students. Since the course objectives and composition of the classes differed, the investigations with each class were treated as replication studies. It was believed that the teachers had a wide variation in prior experience with test taking and test writing. Therefore, it was felt that a pretest should be administered to examine this difference and compare the
treatment to control group on the teacher's existing knowledge. Since only four simple item writing principles were involved in the simulation, a pretest might have alerted teachers to the major concepts addressed by the MCTA. Therefore, a Solomon Four-Group experimental design was used to test both the effect of the MCTA and of the pretest.

A two-way analysis of variance was calculated to test the hypothesis. The specific model statement used (Winer, 1971) was

\[ Y_{ijk} = \mu + P_i + M_j + PM_{ij} + \epsilon_{ijk} \]

where \( Y \) is the posttest score assigned the \( i \)th teacher by the \( j \)th treatment for the \( k \)th group

- \( \mu \) is the overall mean
- \( P \) is the pretest group
- \( M \) is the MCTA treatment
- \( PM \) is the interaction between the pretest group and the MCTA treatment
- \( \epsilon \) is the error

The level of significance selected for testing was 0.05. Additional statistical treatments included descriptive statistics for the simulation related variables, correlations on the simulation related and independent variables with posttest scores, and item analysis of the instrument. Since the two classes used in this investigation were treated separately as replication studies, the findings will be
presented in two parts, the undergraduate experiment and the graduate experiment.

Undergraduate experiment

Results of the analysis of variance for the effects of pretesting and participating in the MCTA simulation are presented in Table 4. Mean scores for each group are presented in Table 5. Inspection of the tables indicates that there was no difference in achievement of teachers as a result of participating in the MCTA simulation. Further, there were no differences in mean scores between teachers who took a pretest and those who did not.

Although the analysis of variance indicated no effect of the pretest on posttest scores, the mean scores for pretest and posttest indicated a gain after instruction. To determine if the gain in scores was significant, paired t-tests were performed on the groups. Table 6 indicates the results of these tests, and shows a significant difference in scores when all teachers pretested were considered together. Therefore, it appears that learning did occur for teachers when they received instruction in item writing.

The findings failed to support either the pretest or participation in the MCTA as making a difference on teacher achievement. While these findings implied that additional
Table 4. Analysis of variance for undergraduate educational psychology teachers

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean squares</th>
<th>F-ratio $^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTA</td>
<td>1</td>
<td>35.53</td>
<td>35.53</td>
<td>1.24</td>
</tr>
<tr>
<td>Pretest</td>
<td>1</td>
<td>3.16</td>
<td>3.16</td>
<td>0.11</td>
</tr>
<tr>
<td>MCTA x Pretest</td>
<td>1</td>
<td>4.21</td>
<td>4.21</td>
<td>0.15</td>
</tr>
<tr>
<td>Residual</td>
<td>41</td>
<td>1173.44</td>
<td>28.62</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>1217.91</td>
<td>27.68</td>
<td></td>
</tr>
</tbody>
</table>

$^1$The degrees of freedom for F are 1 and 41. Table value for F is 4.07 at 5%.
Table 5. Mean scores for undergraduate educational psychology teachers

<table>
<thead>
<tr>
<th>Groups</th>
<th>Number of students</th>
<th>Mean Pretest</th>
<th>Mean Posttest</th>
<th>Posttest score range</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>23</td>
<td>23.04</td>
<td>14-37</td>
<td></td>
</tr>
<tr>
<td>Non-participants</td>
<td>22</td>
<td>21.23</td>
<td>9-28</td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondents</td>
<td>21</td>
<td>19.10</td>
<td>21.81</td>
<td>9-28</td>
</tr>
<tr>
<td>Non-respondents</td>
<td>24</td>
<td>22.46</td>
<td>14-37</td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCTA-Pretest</td>
<td>10</td>
<td>18.50</td>
<td>22.40</td>
<td>18-27</td>
</tr>
<tr>
<td>MCTA-No pretest</td>
<td>13</td>
<td>23.54</td>
<td>14-27</td>
<td></td>
</tr>
<tr>
<td>No MCTA-Pretest</td>
<td>11</td>
<td>19.64</td>
<td>21.27</td>
<td>9-28</td>
</tr>
<tr>
<td>No MCTA-No pretest</td>
<td>11</td>
<td>21.18</td>
<td>15-27</td>
<td></td>
</tr>
</tbody>
</table>

Note: Maximum score = 45.
practice in item writing provided through the MCTA did not improve teacher's ability to write test items other variables may have influenced these outcomes. Some of these variables might include teacher achievement, prior experience with self-directed learning, classification by college year, and commitment to professional development. Additionally, refining the posttest instrument may provide more valid measurement of teacher achievement.

MCTA descriptive data
Teachers performed quite differently on the MCTA. It was expected that teachers would make varying errors in item writing and that the proportion of violations they made would differ. The teachers were initially instructed to try 15 items with the MCTA. Data indicated that a wide variation of ability and proportion of violation types existed. These data were difficult to interpret since the number of items tried varied from the 15 the teachers were instructed to try. Statistics were compiled for the following information: total time spent on MCTA, number of items analyzed, number of analyzed items containing negatives, number of analyzed items containing pros/IS's, number of items containing cons/SD's, number of items containing repetition, and typing skill. Means, standard deviations, and the minimum and maximum values for each variable are reported in Table 7.
Table 6. Paired t-tests of pretest and posttest for undergraduate class

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>t value</th>
<th>2-tailed probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined pretest</td>
<td>21</td>
<td>-2.29</td>
<td>0.033*</td>
</tr>
<tr>
<td>MCTA--Pretest</td>
<td>10</td>
<td>-2.02</td>
<td>0.074</td>
</tr>
<tr>
<td>No MCTA--Pretest</td>
<td>11</td>
<td>-1.14</td>
<td>0.283</td>
</tr>
</tbody>
</table>

*Indicates significance beyond 0.05 level.
Table 7. Mean values for undergraduate class on simulation related variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum Value</th>
<th>Maximum Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTA time in minutes</td>
<td>59.56</td>
<td>16.88</td>
<td>37.00</td>
<td>103.00</td>
</tr>
<tr>
<td>Items analyzed</td>
<td>12.96</td>
<td>5.70</td>
<td>4.00</td>
<td>28.00</td>
</tr>
<tr>
<td>Items saved</td>
<td>9.65</td>
<td>5.25</td>
<td>0.00</td>
<td>23.00</td>
</tr>
<tr>
<td>Items containing negatives</td>
<td>2.35</td>
<td>2.66</td>
<td>0.00</td>
<td>11.00</td>
</tr>
<tr>
<td>Items containing prosD's</td>
<td>2.00</td>
<td>2.58</td>
<td>0.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Items containing consD's</td>
<td>1.44</td>
<td>1.97</td>
<td>0.00</td>
<td>8.00</td>
</tr>
<tr>
<td>Items containing repetitions</td>
<td>3.00</td>
<td>2.58</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td>Typing skill in seconds</td>
<td>59.87</td>
<td>43.23</td>
<td>27.96</td>
<td>235.20</td>
</tr>
</tbody>
</table>

Note: Number of cases = 23.
From the table, it seems that teachers frequently violated the item writing principles being judged in the MCTA. Therefore, it appears that the content addressed by the MCTA is relevant and is not being met adequately at present.

The MCTA required more input to be typed into the terminal by the teacher than was expected when responding to other computer based programs. It was hypothesized that if a teacher's typing skill was unusually slow, he or she would not be able to enter as many items into the Editor and therefore would have a more limited experience on the MCTA. Therefore, a measure of typing skill was taken as the amount of time in seconds it took the teacher to make an exact copy of a statement printed on the screen. Partial correlations were computed to compare the relationships of the simulation related variables with the posttest score while controlling for the typing skill effect. Results of the correlations are presented in a 7x7 matrix in Table 8. Results indicated that posttest scores did not show a significant relationship with any of the variables tested. The number of analyzed items did show a positive significant relationship with the amount of time spent on the MCTA, and the amount of proSD's, conSD's, or repetitions found in the items. These relationships with number of items analyzed were expected phenomena.
For a final comparison, the teacher's total course points for work and tests in educational psychology were entered as the controlling variable in a series of partial correlations with posttest score and simulation related variables. The results are reported in a 9x9 matrix in Table 9. Inspection of the table revealed no relationship of posttest score with any of the variables tested. As expected, violations occurring in an item were highly correlated with number of items analyzed. Typing skill was negatively correlated with number of items analyzed. Typing skill was reported as the number of seconds required to type a line of text correctly; therefore, the negative direction of the correlation was expected.

**Mortality**

Problems with mortality of subjects were encountered with the undergraduate class. The investigator had little control over the class, and pretest and posttest instruments were not administered to all teachers at the same time. While class time was provided for these tests, it came at the end of the class period. Teachers were frequently dismissed one-half hour before the class period ended, so they refused to stay to take the tests. Mortality rate was high, 14%, because several teachers did not use the MCTA at their appointed time, or they were not assigned a time on the
Table 8. Partial correlation matrix for undergraduate teachers controlling for typing skill$^{1, 2}$

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>-16</td>
<td>02</td>
<td>19</td>
<td>-12</td>
<td>02</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>63**</td>
<td>34</td>
<td>27</td>
<td>39</td>
<td>58**</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td></td>
<td></td>
<td>60**</td>
<td>43*</td>
<td>72**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td></td>
<td></td>
<td>-06</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>53*</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>49*</td>
</tr>
<tr>
<td>VII</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^1$Roman numerals mean: I posttest score; II MCTA time; III items analyzed; IV negatives; V proSD's; VI conSD's; VII repetitions.

$^2$The decimal point has been omitted from the correlations on this table.

* Indicates significance at the 0.05 level.

**Indicates significance at the 0.01 level.
Table 9. Partial correlation matrix for undergraduate teachers controlling for total course points

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>-04</td>
<td>17</td>
<td>21</td>
<td>05</td>
<td>12</td>
<td>35</td>
<td>19</td>
<td>-27</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>54**</td>
<td>37</td>
<td>19</td>
<td>36</td>
<td>56**</td>
<td>34</td>
<td>06</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>39</td>
<td>62**</td>
<td>45*</td>
<td>73**</td>
<td>25</td>
<td>-49*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>07</td>
<td>-01</td>
<td>30</td>
<td>49*</td>
<td>-24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>13</td>
<td></td>
<td>51**</td>
<td>42*</td>
<td>-22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td></td>
<td></td>
<td></td>
<td>41</td>
<td>-17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>62**</td>
<td>-26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Roman numerals mean: I posttest score; II MCTA time; III items analyzed; IV negatives; V proSD's; VI conSD's; VII repetitions; VIII ratio of violation types per item; IX typing skill.

The decimal point has been omitted from the correlations on this table.

* Indicates significance at the 0.05 level.

** Indicates significance at the 0.01 level.
terminal because they were absent from class when assignments were made. Although the teachers did eventually participate in the MCTA as one of their class assignments, some did not complete the simulation prior to the posttest and thus they were dropped from the study.

Item analysis

An analysis was performed on the posttest. The reliability estimate calculated by the Kuder-Richardson Formula 20 was 0.68. The item analysis data are presented in Table 10. Best items and potentially good items were selected using the same criteria as for the pilot test. The data indicated 21 of the 45 items as best items. Further, 21 of the remaining items were considered potentially good items.

Graduate experiment

For consistency in analysis of data, the graduate experiment followed the same procedures as the undergraduate experiment. It was recognized that the cell sizes were small for most of the tests performed. Results of the analysis of variance for the effects of pretesting and participating in the MCTA are presented in Table 11. Mean scores for each group are presented in Table 12. Inspection of the tables
Table 10. Item analysis data for posttest instrument used in undergraduate educational psychology

<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>0.96</td>
<td>0.26</td>
<td>0.21</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>--</td>
<td>0.50</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>0.49</td>
<td>0.30</td>
<td>0.50</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.60</td>
<td>0.12</td>
<td>0.49</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>0.76</td>
<td>0.36</td>
<td>0.43</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>0.58</td>
<td>0.45</td>
<td>0.49</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>0.53</td>
<td>--</td>
<td>0.50</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>0.25</td>
<td>0.32</td>
<td>0.43</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>0.56</td>
<td>0.33</td>
<td>0.50</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>0.36</td>
<td>0.22</td>
<td>0.48</td>
<td>16</td>
</tr>
<tr>
<td>11</td>
<td>0.24</td>
<td>0.28</td>
<td>0.43</td>
<td>6</td>
</tr>
</tbody>
</table>

1Items that would probably meet item analysis criteria if respondent size equaled 50.

2Indicates correct answer.

3Items that need revision due to distracter analysis, discrimination index, or difficulty level.

4Discrimination index less than 0.05.

5Items that meet the item analysis criteria: A difficulty index between 0.30 and 0.70 and a discrimination index greater than or equal to 0.20.
<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>0.42</td>
<td>0.08</td>
<td>0.49</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>0.11</td>
<td>0.41</td>
<td>0.31</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>0.69</td>
<td>0.48</td>
<td>0.46</td>
<td>31²</td>
</tr>
<tr>
<td>15</td>
<td>0.58</td>
<td>0.39</td>
<td>0.49</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>0.60</td>
<td>--*</td>
<td>0.49</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>0.40</td>
<td>0.25</td>
<td>0.49</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>0.36</td>
<td>0.46</td>
<td>0.48</td>
<td>16²</td>
</tr>
<tr>
<td>19</td>
<td>0.76</td>
<td>0.31</td>
<td>0.43</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>0.59</td>
<td>0.31</td>
<td>0.49</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>0.60</td>
<td>0.30</td>
<td>0.49</td>
<td>27²</td>
</tr>
<tr>
<td>22</td>
<td>0.67</td>
<td>0.40</td>
<td>0.47</td>
<td>30²</td>
</tr>
<tr>
<td>23</td>
<td>0.64</td>
<td>0.34</td>
<td>0.48</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>0.36</td>
<td>0.50</td>
<td>0.48</td>
<td>17</td>
</tr>
<tr>
<td>25</td>
<td>0.53</td>
<td>0.30</td>
<td>0.50</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>0.82</td>
<td>--*</td>
<td>0.38</td>
<td>7</td>
</tr>
<tr>
<td>27</td>
<td>0.87</td>
<td>0.20</td>
<td>0.34</td>
<td>5</td>
</tr>
<tr>
<td>28</td>
<td>0.64</td>
<td>--*</td>
<td>0.48</td>
<td>29²</td>
</tr>
</tbody>
</table>

*Response given when no option available.*
<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>293</td>
<td>0.78</td>
<td>0.42</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>301</td>
<td>0.47</td>
<td>0.18</td>
<td>0.50</td>
<td>17</td>
</tr>
<tr>
<td>315</td>
<td>0.33</td>
<td>0.27</td>
<td>0.47</td>
<td>15</td>
</tr>
<tr>
<td>321</td>
<td>0.27</td>
<td>0.44</td>
<td>31</td>
<td>12</td>
</tr>
<tr>
<td>331</td>
<td>0.60</td>
<td>0.10</td>
<td>0.49</td>
<td>27</td>
</tr>
<tr>
<td>341</td>
<td>0.71</td>
<td>0.29</td>
<td>0.45</td>
<td>7</td>
</tr>
<tr>
<td>355</td>
<td>0.27</td>
<td>0.32</td>
<td>0.44</td>
<td>12</td>
</tr>
<tr>
<td>361</td>
<td>0.11</td>
<td>0.28</td>
<td>0.32</td>
<td>5</td>
</tr>
<tr>
<td>371</td>
<td>0.13</td>
<td>0.34</td>
<td>0.36</td>
<td>6</td>
</tr>
<tr>
<td>381</td>
<td>0.13</td>
<td>0.34</td>
<td>0.37</td>
<td>6</td>
</tr>
<tr>
<td>391</td>
<td>0.09</td>
<td>0.53</td>
<td>0.28</td>
<td>4</td>
</tr>
<tr>
<td>405</td>
<td>0.43</td>
<td>0.33</td>
<td>0.50</td>
<td>19</td>
</tr>
<tr>
<td>415</td>
<td>0.51</td>
<td>0.34</td>
<td>0.50</td>
<td>18</td>
</tr>
<tr>
<td>425</td>
<td>0.58</td>
<td>0.34</td>
<td>0.49</td>
<td>6</td>
</tr>
<tr>
<td>431</td>
<td>0.47</td>
<td>0.14</td>
<td>0.50</td>
<td>18</td>
</tr>
<tr>
<td>445</td>
<td>0.58</td>
<td>0.47</td>
<td>0.49</td>
<td>6</td>
</tr>
<tr>
<td>451</td>
<td>0.22</td>
<td>0.27</td>
<td>0.42</td>
<td>10</td>
</tr>
</tbody>
</table>
Table 11. Analysis of variance for graduate evaluation class

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean squares</th>
<th>F-ratio$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTA</td>
<td>1</td>
<td>96.27</td>
<td>96.27</td>
<td>3.02</td>
</tr>
<tr>
<td>Pretest</td>
<td>1</td>
<td>0.82</td>
<td>0.82</td>
<td>0.03</td>
</tr>
<tr>
<td>MCTA x Pretest</td>
<td>1</td>
<td>10.08</td>
<td>10.08</td>
<td>0.32</td>
</tr>
<tr>
<td>Residual</td>
<td>18</td>
<td>574.42</td>
<td>31.91</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>680.77</td>
<td>32.42</td>
<td></td>
</tr>
</tbody>
</table>

$^1$The degrees of freedom for F are 1 and 18. Table value for F is 4.0 at 5%.
Table 12. Mean scores for graduate evaluation class

<table>
<thead>
<tr>
<th>Groups</th>
<th>Number of students</th>
<th>Mean Pretest</th>
<th>Posttest</th>
<th>Posttest score range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MCTA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participants</td>
<td>12</td>
<td>28.42</td>
<td></td>
<td>21-34</td>
</tr>
<tr>
<td>Non-participants</td>
<td>10</td>
<td>32.60</td>
<td></td>
<td>20-42</td>
</tr>
<tr>
<td><strong>Pretest</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondents</td>
<td>12</td>
<td>22.42</td>
<td>30.33</td>
<td>20-42</td>
</tr>
<tr>
<td>Non-respondents</td>
<td>10</td>
<td>30.30</td>
<td></td>
<td>27-37</td>
</tr>
<tr>
<td><strong>Interaction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCTA-Pretest</td>
<td>6</td>
<td>21.50</td>
<td>28.83</td>
<td>22-33</td>
</tr>
<tr>
<td>MCTA-No pretest</td>
<td>6</td>
<td>28.00</td>
<td></td>
<td>21-34</td>
</tr>
<tr>
<td>No MCTA-Pretest</td>
<td>6</td>
<td>23.33</td>
<td>31.83</td>
<td>20-42</td>
</tr>
<tr>
<td>No MCTA-No pretest</td>
<td>4</td>
<td>33.75</td>
<td></td>
<td>27-37</td>
</tr>
</tbody>
</table>

Note: Maximum score = 45.
indicated no difference in achievement of teachers as a result of participating in the MCTA simulation. Whether or not teachers took the pretest prior to instruction also failed to produce differences in mean scores. These findings tend to agree with the findings for the undergraduate experiment. The differences were approaching statistical significance however, and the mean scores associated with each cell for the graduate experiment indicated higher posttest scores by the non-MCTA participating teachers. The pretest control group also had higher pretest scores than the pretest MCTA group.

A possible explanation for the differences in posttest scores of the MCTA and control groups lies in the mortality rate during the experiment. For the graduate program, two teachers did not complete the experiment because they dropped the course. Both had been assigned to the no pretest—no MCTA group, thus reducing the group's cell size by 33%.

Although the analysis of variance indicated no effect of the pretest on posttest scores, the mean scores for pretest and posttest indicated a gain after instruction. To determine if the gain in scores was significant, paired t-tests were performed on the groups. Table 13 indicates the results of these tests, and shows a significant difference in scores for each group. Therefore it appeared that learning did occur for teachers.
Table 13. Paired t-tests of pretest and posttest for graduate teachers

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>t value</th>
<th>2-tailed probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined pretest</td>
<td>12</td>
<td>-4.66</td>
<td>0.001**</td>
</tr>
<tr>
<td>MCTA—Pretest</td>
<td>6</td>
<td>-3.84</td>
<td>0.012**</td>
</tr>
<tr>
<td>No MCTA—Pretest</td>
<td>6</td>
<td>-2.85</td>
<td>0.036*</td>
</tr>
</tbody>
</table>

* Indicates significance at 0.05 level.
** Indicates significance at 0.01 level.
The findings failed to support either the pretest or participation in the MCTA as making a difference on teacher achievement. While these findings implied that the additional practice in item writing provided through the MCTA did not improve teacher's ability to write test items, it must be remembered that the cell sizes for each group were unusually small, and that statistical significance would be more difficult to establish.

**MCTA descriptive data**

Built into the MCTA simulation was a facility to collect and report information on the following: total time spent on MCTA, number of items analyzed, number of analyzed items containing negatives, number of analyzed items containing probSD's, number of items containing conSD's, number of items containing repetition, and typing skill. Means, standard deviations, and minimum and maximum values for each variable are reported in Table 14. The table indicates that teachers frequently violate the item writing principles being judged by the MCTA. Therefore it appears that the content addressed by the MCTA is relevant to both preservice and experienced teachers and is not being met adequately through present teacher training practices.

Finally correlations were computed to identify the relationship between the teachers' posttest scores and the
Table 14. Mean values for graduate teachers on simulation related variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum value</th>
<th>Maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC T time in minutes</td>
<td>96.83</td>
<td>21.96</td>
<td>59.00</td>
<td>136.00</td>
</tr>
<tr>
<td>Items analyzed</td>
<td>19.42</td>
<td>3.81</td>
<td>16.00</td>
<td>30.00</td>
</tr>
<tr>
<td>Items saved</td>
<td>12.58</td>
<td>3.68</td>
<td>2.00</td>
<td>16.00</td>
</tr>
<tr>
<td>Items containing negatives</td>
<td>4.50</td>
<td>2.39</td>
<td>1.00</td>
<td>8.00</td>
</tr>
<tr>
<td>Items containing proSD's</td>
<td>1.67</td>
<td>1.56</td>
<td>0.00</td>
<td>5.00</td>
</tr>
<tr>
<td>Items containing conSD's</td>
<td>2.83</td>
<td>2.52</td>
<td>0.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Items containing repetitions</td>
<td>8.00</td>
<td>4.09</td>
<td>2.00</td>
<td>16.00</td>
</tr>
<tr>
<td>Typing skill in seconds</td>
<td>62.61</td>
<td>52.00</td>
<td>21.70</td>
<td>203.00</td>
</tr>
</tbody>
</table>

Note: Number of cases = 12.
simulation related variables. Unlike the undergraduate class, partial correlations which controlled for the effect of class achievement as measured by the total class points could not be performed. The way in which the MCTA was integrated into the graduate class, and the fact that a few of the items in the posttest were also used on a graded test given by the instructor prevented the use of this procedure. Pearson correlation coefficients were computed for each variable pair and are reported in the intercorrelation matrix in Table 15. The data indicate a significant relationship between the posttest score and the total class points. Total class points were also negatively related to the number of items containing consD's. The violation ratio, indicating the average number of violation types occurring in an item was positively related to number of items containing repetition violations. Typing skill did not appear to be related to any of the variables tested.

**Item analysis**

An item analysis was performed on the posttest. The Kuder-Richardson Formula 20 was used to calculate the reliability estimate at 0.77. Best items and potentially good items were selected using the same criteria as for the pilot study. The data indicated 14 of the 45 items as best items, and 20 additional items as potentially good items.
Table 15. Intercorrelation matrix of simulation related variables for graduate teachers¹,²

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>-06</td>
<td>-31</td>
<td>-12</td>
<td>50</td>
<td>-41</td>
<td>-18</td>
<td>02</td>
<td>44</td>
<td>78**</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>65*</td>
<td>39</td>
<td>-47</td>
<td>06</td>
<td>54</td>
<td>-01</td>
<td>-18</td>
<td>-14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>69**</td>
<td>57*</td>
<td>65*</td>
<td>19</td>
<td>00</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>-15</td>
<td>27</td>
<td>54</td>
<td>47</td>
<td>34</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>-02</td>
<td>-33</td>
<td>35</td>
<td>-17</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>52</td>
<td>54</td>
<td>-17</td>
<td>-71*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>70*</td>
<td>24</td>
<td>-24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIII</td>
<td>22</td>
<td>-13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IX</td>
<td>-00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹Roman numerals mean: I posttest score; II MCTA time; III items analyzed; IV negatives; V proSD's; VI conSD's; VII repetitions; VIII ratio of violation types per item; IX typing skill; X total course points.
²The decimal point has been omitted from the correlations on this table.
* Indicates significance at the 0.05 level.
** Indicates significance at the 0.01 level.
The item analysis data are presented in Table 16.

Comparisons of the posttest instrument item analysis for both classes revealed one item, #26, that did not meet the criterion for difficulty level or the criterion for discrimination level. Furthermore, distracter analysis for item #1 indicated that options "b" and "d" were never chosen by either class. In all, seven items appeared to need revision. They are items numbered 1, 7, 19, 26, 27, 29, and 30.

Summary

The present study dealt with the effectiveness of a computer simulation, the MCTA, for improving the multiple choice test writing skills of teachers. Subjects consisted of teachers enrolled in an undergraduate educational psychology class and a graduate evaluation class. Each class was treated as a separate study for implementation and analyses. A Solomon Four-Group experimental design was employed, and teachers were randomly assigned to each of the four groups.

Teacher's ability to write test items free from selected violations was measured by a 45 item multiple choice test developed by the investigator. The test served as a pretest as well as a posttest for the study. The posttest was
Table 16. Item analysis data for posttest instrument used in graduate evaluation

<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>--</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.68</td>
<td>0.62</td>
<td>0.47</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>0.55</td>
<td>0.48</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.50</td>
<td>--</td>
<td>0.50</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>0.77</td>
<td>0.44</td>
<td>0.42</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>0.86</td>
<td>0.33</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.86</td>
<td>0.09</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0.14</td>
<td>--</td>
<td>0.34</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0.57</td>
<td>0.50</td>
<td>0.49</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>0.76</td>
<td>0.22</td>
<td>0.43</td>
<td>2</td>
</tr>
</tbody>
</table>

1Items that would probably meet item analysis criteria if respondent size equaled 50.

2Indicates correct answer.

3Items that need revision due to distracter analysis, discrimination index, or difficulty level.

4Discrimination index less than 0.05.

5Items that meet the item analysis criteria: A difficulty index between 0.30 and 0.70 and a discrimination index greater than or equal to 0.20.
Table 16. Continued

<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>0.95</td>
<td>0.37</td>
<td>0.21</td>
<td>A 21 B 0 C 0 D 0</td>
</tr>
<tr>
<td>135</td>
<td>0.36</td>
<td>0.31</td>
<td>0.48</td>
<td>9 3 8 2</td>
</tr>
<tr>
<td>141</td>
<td>0.86</td>
<td>0.26</td>
<td>0.34</td>
<td>19 0 3 0</td>
</tr>
<tr>
<td>153</td>
<td>0.86</td>
<td>0.45</td>
<td>0.34</td>
<td>2 19 1 0</td>
</tr>
<tr>
<td>161</td>
<td>0.82</td>
<td>0.26</td>
<td>0.39</td>
<td>2 18 2 1 1</td>
</tr>
<tr>
<td>171</td>
<td>0.73</td>
<td>0.53</td>
<td>0.45</td>
<td>4 1 16 2</td>
</tr>
<tr>
<td>185</td>
<td>0.68</td>
<td>0.57</td>
<td>0.47</td>
<td>15 3 4</td>
</tr>
<tr>
<td>193</td>
<td>0.95</td>
<td>-</td>
<td>0.21</td>
<td>0 21 1 0</td>
</tr>
<tr>
<td>203</td>
<td>0.77</td>
<td>0.11</td>
<td>0.42</td>
<td>0 1 1 1 17 2</td>
</tr>
<tr>
<td>211</td>
<td>0.77</td>
<td>0.28</td>
<td>0.42</td>
<td>0 17 0 3</td>
</tr>
<tr>
<td>221</td>
<td>0.82</td>
<td>0.30</td>
<td>0.39</td>
<td>18 1 1 1 2</td>
</tr>
<tr>
<td>231</td>
<td>0.64</td>
<td>-</td>
<td>0.48</td>
<td>2 3 14 3</td>
</tr>
<tr>
<td>245</td>
<td>0.45</td>
<td>0.41</td>
<td>0.50</td>
<td>4 5 10 3</td>
</tr>
<tr>
<td>255</td>
<td>0.68</td>
<td>0.41</td>
<td>0.47</td>
<td>2 3 15 2</td>
</tr>
<tr>
<td>263</td>
<td>0.91</td>
<td>0.07</td>
<td>0.29</td>
<td>2 20 2</td>
</tr>
<tr>
<td>273</td>
<td>1.00</td>
<td>-</td>
<td>0.00</td>
<td>0 22 2</td>
</tr>
<tr>
<td>281</td>
<td>0.82</td>
<td>0.45</td>
<td>0.39</td>
<td>18 4</td>
</tr>
</tbody>
</table>

*Response given when no option available.*
Table 16. Continued

<table>
<thead>
<tr>
<th>Item number</th>
<th>Difficulty index</th>
<th>Discrimination index</th>
<th>Standard deviation</th>
<th>Distracter analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>291</td>
<td>0.86</td>
<td>0.38</td>
<td>0.34</td>
<td>0 192 3</td>
</tr>
<tr>
<td>301</td>
<td>0.86</td>
<td>0.55</td>
<td>0.34</td>
<td>3 192 0</td>
</tr>
<tr>
<td>311</td>
<td>0.23</td>
<td>0.28</td>
<td>0.42</td>
<td>5 0 17</td>
</tr>
<tr>
<td>321</td>
<td>0.59</td>
<td>0.59</td>
<td>0.49</td>
<td>8 1 13</td>
</tr>
<tr>
<td>331</td>
<td>0.86</td>
<td>0.21</td>
<td>0.34</td>
<td>192 0 3</td>
</tr>
<tr>
<td>341</td>
<td>0.91</td>
<td>0.30</td>
<td>0.29</td>
<td>0 2 202</td>
</tr>
<tr>
<td>351</td>
<td>0.27</td>
<td>0.35</td>
<td>0.45</td>
<td>6 2 14</td>
</tr>
<tr>
<td>365</td>
<td>0.59</td>
<td>0.25</td>
<td>0.49</td>
<td>132 7 1 16</td>
</tr>
<tr>
<td>375</td>
<td>0.59</td>
<td>0.31</td>
<td>0.49</td>
<td>9 132 0</td>
</tr>
<tr>
<td>385</td>
<td>0.50</td>
<td>0.47</td>
<td>0.50</td>
<td>11 112 0</td>
</tr>
<tr>
<td>395</td>
<td>0.45</td>
<td>0.46</td>
<td>0.50</td>
<td>102 11 1</td>
</tr>
<tr>
<td>405</td>
<td>0.45</td>
<td>0.28</td>
<td>0.50</td>
<td>102 10 2</td>
</tr>
<tr>
<td>411</td>
<td>0.23</td>
<td>0.30</td>
<td>0.42</td>
<td>17 52 0</td>
</tr>
<tr>
<td>423</td>
<td>0.73</td>
<td>0.15</td>
<td>0.45</td>
<td>3 3 162</td>
</tr>
<tr>
<td>435</td>
<td>0.55</td>
<td>0.46</td>
<td>0.50</td>
<td>10 122 0</td>
</tr>
<tr>
<td>443</td>
<td>0.82</td>
<td>0.15</td>
<td>0.39</td>
<td>2 2 182</td>
</tr>
<tr>
<td>451</td>
<td>0.50</td>
<td>0.43</td>
<td>0.50</td>
<td>112 8 3</td>
</tr>
</tbody>
</table>
administered to two of the groups and the posttest was administered to all four groups. Posttest reliability estimates for the two classes were 0.68 and 0.77.

Analysis of variance procedures indicated no significant differences between groups participating in the HCTA with those not participating. No significant differences were found between teachers responding to the pretest and those not responding.

For groups receiving the pretest, paired t-tests revealed significant differences between pretest and posttest scores. Inspection of the gains in mean scores shows that while both the undergraduate and the graduate classes had significant differences between their pretest and posttest scores, the practical significance of the undergraduate's gain (2.7 points) was questioned.

Inspection of the simulation related variables compiled while teachers participated in the MCTA revealed differences between the undergraduate and graduate classes. Undergraduate level teachers spent less time (59.56 minutes as compared to 96.83 minutes) on the MCTA than the graduate level teachers. While both classes were instructed to try 15 items on the MCTA, the undergraduate teachers tried fewer items. Since the MCTA was not able to identify the actual number of different items tried, further comparisons of the means of the two groups could not be made. The values
do indicate, however, that teachers continued to violate the item writing principles addressed by the MCTA.

Partial correlations between selected variables and the posttest score indicated that neither typing skill nor class achievement had a confounding influence on the posttest scores of undergraduate teachers. Since total class points in the graduate level class included scores on a few items also used in the posttest instrument, a partial correlation using total class points was not performed for that group.

Due to the sample sizes involved, the research findings should be interpreted conservatively. Maximum cell sizes for the undergraduate and graduate classes were 13 and 6 respectively. Cell sizes of 30 (120 total) subjects would strengthen the findings because as the sample size is increased, the treatment effect tends to stabilize (Borg & Gall, 1971).

Lastly, while the investigations of the two classes have been referred to as replication studies, it is difficult to generalize the findings over the two experiments. Several problems were encountered in the administration of the experiments. The investigator had little control over the undergraduate class, and the sequencing and times between pretest and posttest were not the same for all subjects. Mortality rate was high for both experiments. In the graduate experiment, all dropped subjects occurred in one
cell which changed the composition of the group considerably. Teacher commitment may have been higher for the graduate teachers due to their presence in the class, and the closer integration of the simulation to the major topics of the course.
Computer technology has influenced several aspects of educational practice. Computer assisted instruction has grown from the teacher replacement functions of tutorial and drill-and-practice programs, to more complex simulation and problem solving programs. The present challenge to instructional designers has been to develop environments which support heuristic learning strategies. Such environments encourage students to learn in self-directed and proactive ways.

The self-directed approach for learning systems dictates additional concern for evaluating such programs. Past studies have investigated techniques used in computer based learning such as framing sequences, graphic displays, presentation or flagging of errors, and response format types (Domenech, 1974; King, 1975; Mason, 1974). However, these techniques were often incorporated in less sophisticated learning environments. With the trend toward heuristic learning strategies, new appraisals for evaluating program effectiveness are necessary.

The purpose of the present study was to develop a computer program for educators which facilitated heuristic learning strategies. The Multiple Choice Test Analyzer (MCTA) was developed as a supplementary aid to classroom instruction on writing multiple choice test items. The MCTA
encouraged proactive learning in preservice teachers by allowing variation in the types of test items used with the simulation. Few restrictions were placed on the content of the items, number of options per item or the number of items tried.

The actions taken by the simulation were teacher directed. Rather than specifically identifying the good and poor items, the simulation provided item indices of difficulty and discrimination and an indication of test reliability. Help units were available to assist the teacher in interpreting these indices. The teacher then determined whether to save the item for the total test, modify it, or delete it and try a new item. Only after the teacher attempted to modify the item and then requested the outcome of the analysis was the actual violation type given.

The main hypothesis for the study, stated in the null form, was

There is no significant difference in posttast scores measuring the ability to write test items free from selected violations between preservice teachers using the computer simulation and those not using the simulator.

To test this hypothesis, a 45 item instrument was developed by the investigator. The instrument served as both the pretest and the posttest for the study.
The study was conducted with two classes, an undergraduate level educational psychology class and a graduate level evaluation class. The study was administered separately for the two classes during the First Summer Session 1977. Results of the experiment were treated as replication studies.

Teachers enrolled in the classes were randomly assigned to one of four groups. Two of the groups received the treatment, using the MCTA simulation, and other two groups served as the control. A Solomon Four-Group experimental design was used, administering a pretest to one of the treatment and one of the control groups. Analyses of variance were performed on the posttest scores of teachers in each group to determine the effect of the MCTA. Results of the analyses supported the null hypothesis. No significant differences were found between teachers participating in the MCTA and those who received only classroom instruction. Likewise, no significant differences were found between teachers who had taken a pretest prior to the experiment and those who had not.

Descriptive data were compiled on several variables associated with the MCTA. Findings indicated that most teachers spent one to one-and-one-half hours at the terminal, and that the violations checked by the MCTA (repetitions, negatives, pro and con specific determiners) frequently
occurred in the test items submitted by the teachers.

Finally, correlation coefficients were computed to determine any relationship between simulation related variables and the posttest scores. Data indicated no significant relationships between posttest scores and other variables tested. Neither the total course points, nor the typing skill of the teacher appeared to have a confounding effect on the relationship between the simulation related variables and the posttest score.

While the outcomes of this study failed to support the effectiveness of the MCTA simulation, they probably do not suggest omitting the simulation from instruction. Factors which could have influenced these results include intervening variables not addressed in this study, and limitations to the implementation of the study. Some of the intervening variables which might have affected the teacher's performance are prior experience with self-directed learning, classification by college year and commitment to professional development. Limitations in class sample size, mortality, time available for instruction, and the degree to which the MCTA was incorporated within the participating classes might have adversely influenced the study. The cell sizes for each study were small, with a maximum size of 13 for the undergraduate and 6 for the graduate class groups. Secondly, the control by the investigator over the administration of
the study was limited. Mortality rate was high and not evenly distributed over the groups. Further investigation of the problem is highly recommended.

If the study were to be replicated, several refinements are suggested. It appeared from the data on pretest scores that teachers' prior knowledge of test development did not have a significant effect on the outcome of the study. Further, the pretest did not appear to alert some groups and to give them an advantage in learning the item writing concepts. It is suggested that controlling for pretested groups is no longer necessary. Using a design other than the Solomon Four-Group seems justified, and would allow for larger cell sizes and more degrees of freedom given similar sized classes.

Recommendations for further study are made in the following areas: introduction and utilization of the simulation, refinement of the instrument, and refinement of the simulation.

Introduction and Utilization of the Simulation

For the experiment, a brief introduction to the MCTA was given. The investigator spent approximately 15 minutes describing the study and PLATO operation. A large portion of that time was spent in assigning group and terminal
assignments, and in explaining procedures to get onto the terminal.

To encourage the type of self-directed learning expected in the simulation, a more extensive introduction should be given. Teachers must realize that the MCTA is an integral part of their course instruction. The use of the simulation and the information gathered could be used in class discussions and as an initial step in constructing evaluative instruments. A follow-up discussion to explain possible simulation outcomes would be helpful, not only to understand the program, but also to reinforce the concepts involved.

Refinement of the Instrument

To assure that the instrument is measuring the anticipated outcomes from participation in the MCTA, further analysis should be made. Responses to test items might be compared between control and treatment group. Discrimination of items might be examined by experimental group rather than the customary high- and low-scorers. Further validation of the instrument might include comparing the reliability between test score and actual performance when writing items.
Refinement of the Simulation

The PLATO system provides several reporting functions. Counts can be made on the number of times teachers respond in certain ways. The MCTA contained several units which the teacher could choose to help determine his or her action on an analyzed item. Studying the ways that teachers utilize these units might provide insight into the quality of self-directed experience the teacher receives. One feature of the MCTA provided the actual reporting of violations found only after the teacher had made an attempt to modify the item. Pilot study results indicated that some teachers were not aware of the existence of the feature while others found the feature very useful. Further investigation of the use of this feature, or the types of teachers choosing to seek such assistance might be helpful for future development.

Okey and Majer (1975) found no significant differences in achievement while using a PLATO terminal when small groups rather than individual students used the program. They did find a significant difference in the time taken to complete the program however, and concluded that groups of size three and four were the most efficient. Investigation into the effect peer interaction has on learning through the MCTA could be studied.
Additional development of the simulation may include adapting the program for use with a printer. If such a facility were available, the teacher could instruct the computer to print out the items that have been saved as good items. In this way, the teacher would leave the terminal with a printed copy of the revised items.

Continued development of the MCTA features could involve expanding the number of violations tested. Possible additional violations might include effect of length of option, response pattern for correct choices and repetition of words between the choices.

There is recent evidence of interest in computer test generating systems. Many advances have been made in computer pooling of test items with accompanying item analyses when available for future test generation. Such systems could provide for sequencing and individualizing tests to specific pupil needs (Denney, 1973; Emerson, 1974; Hazlett, 1973; Prosser & Jensen, 1971; Spineti, 1974). The algorithm used in the MCTA could be adapted for use in conjunction with these systems to identify violations in the items and thereby strengthen the test generating process.


Costin, F. Three choice vs. four choice items: Implications for reliability and validity of objective achievement tests. Educational and Psychological Measurement, 1972, 32, 1035-1038.


Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, 1974, 6, 208-212. (a)


ACKNOWLEDGMENTS

The author would like to express appreciation to several people for instruction, guidance and encouragement received during graduate study. Very special thanks are extended to Dr. Rex Thomas for the continued instruction, guidance, support and encouragement he has given throughout the graduate program and as committee co-chairperson.

Appreciation is expressed to Dr. Alyce M. Fanslow for her help as committee member and teacher, and especially for her advice throughout the development of the theoretical framework and measurement instruments.

In addition, the author is grateful to committee members Dr. Geraldine Montag and Dr. Donald Schuster for their interest, direction and counsel.

Students and instructors who participated in the pilot testing and the experiment are acknowledged for their cooperation and comments. The author also wishes to acknowledge the preliminary work done on this topic by Michael Prentice in conjunction with a special topics course in Computer Science.

Special recognition is due Dr. Eleanor Kohlmann, initial advisor and co-chairperson for her support and concern, and Dr. Ruth F. Hughes for her willingness and guidance in duties of advisor and co-chairperson after Dr. Kohlmann's death.
Thanks are extended to Pete Boyser, graduate assistant assigned to the PLATO project, for his help in programming efficiency.

Finally, the author is grateful to family and friends who were willing to listen, stand by, and encourage whenever needed.
APPENDIX A: PRO SPECIFIC DETERMINERS AND CON
SPECIFIC DETERMINERS USED IN MCTA ARRAYS

**ProSD's**

But

Except

Frequently

Generally

Most

Often

Seldom

Some

Sometimes

Usually

**ConSD's**

Absolutely

All

Always

Any

Each

Every

Impossible

Only

whenever

Wherever
APPENDIX B: IGNORE FOR REPETITION ARRAY USED IN MCTA

A
An
And
At
Be
Been
By
For
Had
Has
Have
If
In
Is
Of
On
One
That
The
This
To
You
APPENDIX C: NEGATIVE ARRAY USED IN MCTA

Never
No
Not
None
n't
APPENDIX D: OPEN ENDED QUESTIONS ADMINISTERED TO PILOT GROUP
Thank you for taking the time to use the Multiple Choice Test Analyzer lesson on the PLATO terminal. It would help me to improve the program if you would give me your reactions to its operation. If you would, please answer the following questions and return them to the designated box in 2068 Curtiss before you leave. Thank you.

1. What is your overall reaction to the program?

2. Did you have any mechanical difficulties with the program? If so, please describe the difficulties and in what section of the lesson they occurred.

3. Were the instructions easy to understand? If not, please identify any sections that confused you.

4. How much time did you spend on the lesson:
   a. in preparation before going to the terminal?
   b. while using the terminal?

5. How many different items did you try on the simulation?

6. The simulation checked for four item writing violations. Can you list them?

7. After each item is analyzed, you were given an option to modify the item. If you did modify the item and had it re-analyzed, you then had the opportunity to see what specific errors still existed in your item if you pressed the DATA key. Were you aware of the DATA key option? Did you feel that it was appropriate after you first tried to modify your item?

8. Did the results of the simulation support knowledge from your class work? Explain areas of support and/or nonsupport.
APPENDIX E: INSTRUCTIONS FOR USING THE PLATO TERMINAL
AND MCTA SIMULATION
PLATO Lesson on Multiple Choice Item Writing

Preparation before going to the PLATO terminal

For the PLATO lesson on test item writing, you will be asked to write several items which then will be analyzed for certain grammar related errors. Before going to the terminal, write the 15 items you plan to use and take them with you. Please follow these suggestions in writing your items.

All items are to be of multiple choice format.
Include 3 to 5 choices for each item.
The items should be written for junior high or higher level students.
The items must be in verbal form. That is, items which use scientific equations or have numbers or dates as choices, are not appropriate with this lesson.

Scheduled time

Make a note of your scheduled time to use PLATO here. The terminals are located in 206B Curtiss.

PLATO name

terminal number 27-
day/date

time

The attached sheet contains information you need to sign on the terminal. The "General Instructions" give you sign on and lesson information. The "Multiple Choice Question Editor" summarizes the keys you will need to use the Editor, and to finish the lesson. These instructions are also contained within the PLATO lesson itself.

Problems

If you have any difficulties with the lesson, please contact Dr. Rex Thomas at 294-2219 or Cheryl Hausafus at 294-7012.
Two PLATO terminals are located in 206B Curtiss. If you have any difficulties with the terminals please notify Dr. Rex Thomas at 294-2219 or Cheryl Hausafus at 294-7012.

General Instructions for PLATO terminal

Signing on When you first sit down at the PLATO terminal, press the NEXT key on the right side of the keyboard. You will see a welcome sign asking your name. Type in your first initial and last name without any spaces. DO NOT USE CAPITAL LETTERS. Then press NEXT.

Course name You will be asked to enter your course. It is testitem. Type in the course name and press SHIFT-STOP. (Hold down the SHIFT key and continue holding it while you press the STOP key.)

Password If a password is requested, choose any word you wish to be your password and enter it when asked. Only X's will appear on the screen as you type in your password so no one will be able to read it. Remember the password you choose. Each time you enter PLATO you must give it. Press NEXT after your password.

Lessons: If you have not used the PLATO terminal before, do lesson "a", Introduction to PLATO. It will take about 15 minutes. It will teach you how to operate a PLATO terminal using such keys as NEXT, COPY, EDIT and ERASE.

Multiple Choice Test When you are familiar with the general operations of PLATO, proceed to lesson "b", the Multiple Choice Test Analyzer. The instructions for the lesson are on the following page and are also contained within the lesson.

Signing off Before you leave the terminal, make sure you have signed off properly. Press SHIFT-STOP two or more times until the Press NEXT to begin appears on the screen.
Using the Multiple Choice Question Editor

Command arrow. To enter an item into PLATO, keys must be pressed in a specific order. First, you must indicate which part of the item you are entering, the stem "s", or one of the choices "a", "b", "c", "d", or "e" by pressing s, a, b, c, d or e at the command arrow at the top right of the screen.

Multiple lines per item part

The arrow should then move to the item part you indicated. Type in that part of the item example. You may use up to four lines for each item part. To proceed to the next line in an item part, press NEXT. To return to a previous line, press BACK.

Storing the item part

When you have completed typing an item part, you must store that part before going on. Press SHIFT-BACK to store each part. The command arrow at the top of the screen will then wait for your next instruction.

Deleting an item part

After an item part has been stored if you wish to delete that part, press SHIFT-LAB and indicate which part to delete.

Analyzing the item

After you have entered your entire item, and are ready to have it analyzed, press SHIFT-DATA. As a result of the analysis, you may wish to modify the item or try other items. Instructions for each option will appear on the screen.

Completing the lesson

When you have entered all your test items, you may receive a summary of item violations you made. Press SHIFT-NEXT to go to the summary.

Signing off

Before you leave the terminal, make sure you have signed off properly. Press SHIFT-STOP two or more times until the Press NEXT to begin appears on the screen.
APPENDIX F: PRE-POST TEST INSTRUMENT
Analyzing Test Item Construction

This test is in four parts. Each part contains specific directions which should be fully understood before answering items in that part. Be sure to read the directions for each before proceeding to that part. Mark your answers to each item by blackening the space corresponding to your answer on the answer sheet provided. Enter your name as directed on the top left of the answer sheet and your Social Security number in the first nine columns of the identification number in the top right of the answer sheet.

--- PART I ---

For each item, select the choice which best answers the question.

Section A

1. The type of test item which contains a beginning statement followed by several alternative options is
   a. completion
   b. matching
   c. multiple-choice
   d. true-false

2. The characteristic of a test which produces consistent results when administered to similar groups of students is
   a. correlation
   b. discrimination
   c. reliability
   d. content validity

3. Item difficulty is indicated by the percent of
   a. students who think the item is hard
   b. students who answer the item incorrectly
   c. top half of the class who answer the item wrong
   d. total class who answer the item correctly

4. The characteristic of a test which produces consistent results when administered to the same group of students at two different times is called
   a. correlation
   b. discrimination
   c. reliability
   d. content validity

5. A comparison of the number of better students answering an item correctly to poorer students answering the item correctly is
   a. difficulty
   b. discrimination
   c. reliability
   d. content validity
6. To determine if an item is appropriate or a poor item, examine
   a. percent response for each choice
   b. item reliability
   c. both difficulty and discrimination
   d. both correlation and content validity

7. To determine if all the distractors are viable, examine each option's level of
   a. difficulty
   b. discrimination
   c. reliability
   d. response

8. Measures of item difficulty are useful in
   a. evaluating attainment of instructional objectives
   b. comparing the performance of better and poorer students
   c. determining the appropriateness of distractors
   d. establishing the appropriateness of item content for students

9. Measures of item discrimination are useful in
   a. evaluating attainment of instructional objectives
   b. comparing the performance of better and poorer students
   c. determining the appropriateness of distractors
   d. establishing the appropriateness of item content for students

Section B

To answer items 10-13, examine the following response patterns expressed as the percent of students answering each choice in an item.

<table>
<thead>
<tr>
<th>Item A</th>
<th>Item B</th>
<th>Item C</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. 0</td>
<td>a. 13</td>
<td>a. 85 answer</td>
</tr>
<tr>
<td>b. 40 answer</td>
<td>b. 14</td>
<td>b. 7</td>
</tr>
<tr>
<td>c. 32</td>
<td>c. 60 answer</td>
<td>c. 4</td>
</tr>
<tr>
<td>d. 28</td>
<td>d. 13</td>
<td>d. 4</td>
</tr>
</tbody>
</table>

10. Of the three items above, which had a difficulty index of .40?
   a. Item A
   b. Item B
   c. Item C
   d. insufficient data to determine

11. Which item had a discrimination index of .15?
   a. Item A
   b. Item B
   c. Item C
   d. insufficient data to determine

12. Which item appears to have the most viable distractors?
   a. Item A
   b. Item B
   c. Item C
   d. insufficient data to determine

13. Which item above had the largest difficulty index?
   a. Item A
   b. Item B
   c. Item C
   d. insufficient data to determine
Section C

14. If a word from one of the following lists were used in an item option, which would most likely cause students to choose it?
   a. generally, often, usually
   b. any, each, only
   c. all, always, every
   d. never, none, not

15. Which of the following lists of words would be more appropriate to use as determiners in item construction?
   a. generally, often, usually
   b. major, primary, secondary
   c. all, always, every
   d. never, none, not

---PART II---

Part II contains sample items and parts of items which you must analyze for item construction. Answer each item according to the directions given at the beginning of the section. Do not attempt to answer items for the subject knowledge being tested.

Section A

For the following item distractors, choose the distractor which would least likely give students verbal cues in choosing it.

16. The primary diet of the aardvark is
   a. grasses and all types of leaves
   b. green grasses and leaves
   c. only grasses and leaves

17. The primary diet of the aardvark
   a. contains fruits and berries exclusively
   b. generally consists of fruits and berries
   c. includes fresh fruits and berries

18. The primary diet of the aardvark
   a. consists of freshwater fish
   b. frequently includes birds
   c. never includes termites

Section B

For the following item choices, identify the underlined word in each choice that violates item writing principles by giving verbal cues. Select "d" if none of the underlined words causes a violation.

Stem: The purpose of an abrasive substance is to

Choices:

19. grind down all softer substances
   a.  
   b.  
   c.  
   d. 

20. cut or drill through denser substances
   a. 
   b. 
   c. 
   d. 

21. polish each of the cereal grains
   a. 
   b. 
   c. 
   d. 

Section C

Assuming students did not know the correct answers to the following items and used verbal cues for their selection, which choice would they most likely select?

22. According to the article by Scott on the effects of A and B, which of the following is true?
   a. generally A occurs before B
   b. every time A occurs, B follows
   c. A always occurs before B
   d. A and B never occur in the same situation

23. Acetanilid is a drug which is
   a. seldom used to inhibit infection
   b. administered by injection only
   c. usually used as a diuretic
   d. no longer used in the medical profession

24. An example of a viscous fluid is
   a. mineral oil
   b. French dressing
   c. lighter fluid
   d. distilled water

25. The Yahgan Indians are of special interest to anthropologists because they
   a. move quickly but don't run
   b. only fashion their weapons from hardened clay
   c. make their clothing from woven human hair
   d. always eat vegetarian meals

Section D

For each pair of item stems, select the stem which is more appropriate.

26. a. It is not a good practice to seat students according to
   b. It is a poor practice to seat students according to

27. a. A household accident is often caused by
   b. The primary cause of household accidents is

28. a. A basic tenet of the theory of cosmogony is
   b. Which of the following beliefs concerning cosmogony is always true?
Section E

For each choice below, determine if the choice contains
a. attracting words which would influence students to choose it
b. repelling words which would influence students to avoid choosing it
c. words that neither attract nor repel students from choosing it

Stem: When arranging furniture you should
Choices:
29. not have to think about the size of the pieces
30. use only upholstered or wood pieces in one room
31. keep the purpose of the furniture pieces in mind
32. plan the areas of greatest use first

Stem: An archipelago is a
Choices:
33. refuge often sought by prisoners
34. group of scattered islands
35. storage room with arched entrances

---PART II---PART III---PART IV---

For each item, select the choice which best answers the question.

36. When attracting words are present in the correct choice, it causes the
difficulty index to
   a. increase
   b. decrease
   c. remain the same

37. When attracting words are present in a distractor, it causes the difficulty
   index to
   a. increase
   b. decrease
   c. remain the same

38. When repelling words are present in the correct choice, it causes the
difficulty index to
   a. increase
   b. decrease
   c. remain the same

39. When repelling words are present in a distractor, it causes the difficulty
   index to
   a. increase
   b. decrease
   c. remain the same
For the following items, analyze each for its item construction. Do not attempt to evaluate items for the subject matter being tested.

For each of the partial items below, either alternate answers or alternate distractors are provided. If the choices were changed from the "before" to the "after" wording, what would happen to the item difficulty index?

a. increase
b. decrease
c. remain the same

Distractor pairs:

40. Red tide is caused by
   Distractor Before: red scum which comes in with the tide
   After: scum and bacteria brought in with the current

41. The location of the story in the opera La Boheme takes place
   Distractor Before: in the northern part of present day Czechoslovakia
   After: in the northern part of Bohemia

42. A seahorse's eggs are laid in
   Distractor Before: a sandy nest
   After: a coral nest

Answer pairs:

43. The kiwi bird's nest can be found
   Answer Before: usually on a river bank
   After: on a river bank

44. Koala bears live in
   Answer Before: bushes in Australia
   After: trees in Australia

45. To give recompense means to
   Answer Before: pay for in kind
   After: make compensation for