THESIS

This is to certify that the

thesis entitled

A Comparison of the Quality of Selected Multiple-Choice Item Types within Medical School Examinations

presented by

Julie G. Nyquist

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Measurement and Educational Psychology

Robert L. Ebel

Major professor

Date May 20, 1981

O-7639

A COMPARISON OF THE QUALITY OF SELECTED MULTIPLE-CHOICE ITEM TYPES WITHIN MEDICAL SCHOOL EXAMINATIONS

By

Julie G. Nyquist

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling and Educational Psychology

1981

ABSTRACT

A COMPARISON OF THE QUALITY OF SELECTED
MULTIPLE-CHOICE ITEM TYPES WITHIN
MEDICAL SCHOOL EXAMINATIONS

by

Julie G. Nyquist

The purpose of this study was to compare groups of test items selected on the basis of format or item writing rule violation, for psychometric quality based on data collected from administration of regular classroom tests within medical education. Five item types were identified for study as were four relevant comparisons between them. All five item types are variants of the basic multiple-choice item format. The two formats chosen for comparison in this study were the traditional single-answer multiple-choice format and a multiple-answer multiple-choice format used primarily in medical education and known as type k on the National Board of Medical Examiners' examinations. The two rule violations are: use of negatively worded stems and use of incomplete stems, stems which do not state specifically the question to be answered.

A total of 718 test items were used in the study. All of the items came from the College of Human Medicine at Michigan State University and were taken from item pools composed of test items previously used on regular classroom test in each subject area. In the total study, 387 pairs of items were selected: 124 in the

single-answer multiple-choice vs. multiple-answer multiple-choice comparison, 51 in the single-answer multiple-choice vs. uninformative stem multiple-choice comparison, 69 in the multiple-answer multiple-choice vs. uninformative-stem multiple-answer multiple-choice comparison and 143 in the single-answer multiple-choice vs. negative-stem multiple-choice comparison. These 387 item pairs originated on 40 separate exams. Content was controlled by using pairs of items from the same test administration which were keyed to the same topic area. In the item selection process items with errors not being studied were eliminated. Also, the average number of options per item was controlled for comparisons including items with varying numbers of options.

Three item statistics were chosen as estimates of item quality: p, proportion getting the item correct; D, the upper-lower discrimination index, and rBis, the biserial item-total correlation coefficient. Univariate repeated measures ANOVAS were used to test the three hypotheses related to each of the four comparisons. The level of significance for the F-tests was set at .05. However, since a univariate analysis was used, the Bonferroni approach for correction of a possible inflation of alpha, the type one error, was applied. With this method the desired alpha level is divided by the number of comparisons (i.e., .05 is divided by 3) so the cutoff for significance was set at .017.

Results of all analyses based on item type performed to test the hypotheses of the study are summarized below:

> Hypothesis 1 - In the comparison of SA-MC and MA-MC items, the MA-MC items were found to be significantly more difficult and less discriminating than the SA-MC items.

Hypothesis II - In the comparison of SA-MC and US-MC items, the US-MC items were found to be significantly less difficult than the SA-MC items. There were no significant differences based on discrimination.

Hypothesis III - There were no significant differences in either difficulty or discrimination found when MA-MC and US-MA-MC items were compared.

Hypothesis IV - In the comparison of SA-MC and NS-MC items, the NS-MC items were found to be significantly less discriminating than the SA-MC items. There was no significant difference in difficulty between the two item types.

It was concluded that: Item format affected both average difficulty and discrimination; stem orientation, positive versus negative, affected item discrimination but not difficulty; and that informativeness of the item stem had some effect on item difficulty but no statistically significant effect on discrimination.

DEDICATION

To my daughter
Stephanie Kraig Olivero
Who brings joy to my life and
makes the struggle worthwhile

ACKNOWLEDGEMENTS

I am glad to finally have this opportunity to acknowledge all of the people who have given me support and assistance throughout my doctoral program. First and foremost I would like to thank the entire Educational Psychology faculty for their encouragement, their excellent teaching, and their faith in my ability. Special thanks beyond this go to Drs. Robert Ebel, Larry Lezotte, and Bill Mehrens for always being there, willing to listen to me and advise when appropriate.

I would like also to acknowledge those within medical education who helped provide my financial support and further helped me develop the direction for my professional career. Drs. Tom Parmeter, Jack Jones, Steve Downing, Marty Anderson, Dan English, and Verda Scheifley all have my deepest respect and regard for their contributions to my development as a professional, especially Tom. Thank you from the bottom of my heart, Dr. J. Thomas Parmeter, for caring enough to allow me the time and space to be a mother as well as become a medical educator.

I am also happy to acknowledge those who helped make the process of working toward a doctorate enjoyable. A warm thank you to my fellow doctoral candidates and friends, all of whom finished before I did but none of whom forsook me, Drs. Robert Griffore, John Molidor, and James Haf.

Finally, I would like to acknowledge the assistance of Harry Davis at the Medical College of Georgia for his extremely valuable help in

completing the analysis of my data. Thank you, Harry, for all of your time and persistent effort.

It is my profound hope that I am developing into a professional worthy of the caring support I have received from those mentioned above and many others as well.

TABLE OF CONTENTS

Chapter		Page
I.	THE PROBLEM	1
	Introduction	1
	Various Forms of Multiple-Choice Items	3
	Conventional Rules for Item Writing	5
	Some Highlights of Previous Studies	10
	Need for Further Research	13
	Context of the Study	14
	Specific Questions to be Answered	15
	Summary	16
II.	REVIEW OF RELATED RESEARCH	17
	Multiple-Answer Form	17
	Negative Stem Form	21
	Uninformative Stem Form	23
	Number of Options	29
	Other Rule Violations	31
	Summary	38
III.	PROCEDURES AND DESIGN	39
	Introduction	39
	Source of Items	39
	Item Selection and Content Control	43
	Measures of Item Quality	52
	Design	54
	Hypotheses and Analysis Methods	56
	Summary	61
IV.	RESULTS	62
	Introduction	62
	Results Based on Item Type	62
	Results of the SA-MC versus MA-MC Comparison	64
	Results of the SA-MC versus US-MC Comparison	66
	Results of the MA-MC versus US-MA-MC Comparison	68
	Results of the SA-MC versus NS-MC Comparison	70
	Results Based on Content Orientation	71
	Summary	75

Chapter Page

V. DISCUSSION AND CONCLUSIONS 76

Introduction 76
Discussion 76
Conclusions. 95
Suggestions for Future Research. 96

APPENDICES. 98

A. Source of Items - Breakdown by Exam of Origin. 98

B. Sample Item Pairs: SA-MC vs. MA-MC Comparison 100

BIBLIOGRAPHY 112

LIST OF TABLES

Table		Page
1.1	Frequency of Rule Violations in a Survey of Over 3,000 Items on Exams in the College of Human Medicine at Michigan State University	9
3.1	Number of Items used in Each of the Four Research Comparisons.	40
3.2	Source of Item Pairs by Content Area and Comparison . .	42
3.3	List of Topics used for Item Selection within the Focal Problem Areas and Separate Clinical Specialties	44
3.4	Illustration of Steps 3, 4 and 5 of the Item Selection Procedure using the SA-MC vs. MA-MC Comparison and the Spring 1976 Elevated BUN Exam as the Example. . .	47
3.5	Sample SA-MC vs. MA-MC Item Pairs from the Elevated BUN Spring 1976 Illustration.	49
4.1	Mean Values for All Relevant Groups in Comparisons Based on Item Types	63
4.2	Results for the SA-MC vs. MA-MC Comparison Two-Way Repeated Measures ANOVA, n =124	65
4.3	Results for the SA-MC vs. US-MC Comparison Two-Way Repeated Measures ANOVA, n = 51	67
4.4	Results for the MA-MC vs. US-MA-MC Comparison Two-Way Repeated Measures ANOVA, n = 69	69
4.5	Results for the SA-MC vs. NS-MC Comparison Two-Way Repeated Measures ANOVA, n = 143.	72
4.6	Mean Values for all Relevant Groups in Comparisons Based on Content Area	74
5.1	Sample SA-MC vs. US-MC Item Pairs	83
5.2	Sample NS-MC Items with Poorly Focused Stems (PFS Items)	89
5.3	SA-MC vs. NS-MC Comparison, Grouped Frequency Distribution Based on D and rBIS.	91

CHAPTER I

THE PROBLEM

INTRODUCTION

Teacher-made achievement tests are used widely in medical education as in other areas of education. In non-graded systems they are used to help determine whether students pass or fail; in programs using traditional letter grades the students' scores on any particular test may constitute anywhere from a small proportion to 100% of the total grade for a course. The quality of classroom achievement tests is therefore very important and of major concern to measurement specialists.

The most important consideration in achievement testing is the quality of the total test as reflected by reliability and validity. Since a test is composed of individual items the quality of the test will be influenced by the quality of each item. Technical quality of any item in terms of item difficulty and item discrimination cannot be computed until after the test is administered. Therefore measurement specialists look for factors which tend to consistently affect item quality. Knowledge of these factors is especially important in classroom achievement testing because there is no test tryout to eliminate poor items. The first time the test is given is often the last time as well; student performance is scored and grades given on the basis of this initial administration.

What factors consistently affect item quality? Item format can affect item quality. For instance, the vast majority of studies comparing single-answer multiple-choice items to true-false items have shown the

single-answer multiple-choice item to be generally more discriminating than true-false items even when the statistics were adjusted for differences in testing time per item. This does not mean that excellent true-false items cannot be written by experts. They can be and are. It does however indicate that item format can make a difference in typical usage. The primary purpose of this study is to answer the question, "Does typical use of the multiple-answer multiple-choice (MA-MC) format result in items which differ in terms of difficulty or discrimination from items written in the single-answer multiple-choice format?"

The other major factor thought to affect item quality is the violation of item writing principles. These principles are intended to guide item writers and when followed should result in high quality items (naturally only if combined with expert knowledge of the subject matter and understanding of the students to be tested). Violations may not always result in items of poor quality but because rule violations are often merely the visual evidence of poor quality thinking, poor quality items are often the result. The second purpose of this study is to address the question: "What effect does violation of either of the two rules specified have on the quality of items found on typical examinations within medical education, measured using average item difficulties and discrimination indices?" These two violations refer to items with negatively phrased stems and items with incomplete uninformative stems.

VARIOUS FORMS OF MULTIPLE-CHOICE TEST ITEMS

Many forms of multiple-choice test items have been used in the last fifty years. The five forms under study here are those most commonly found

within medical education. Below is an example of each form along with relevant directions:

Directions: Please select the ONE best answer for the item in the number corresponding to your choice on the answer sheet provided.

Single-Answer Multiple-Choice (SA-MC) Form - Example:

The diagnosis of intestinal amebiasis depends upon identification of

1. mucosal lesions by sigmoidoscopy
2. organism in the stool or tissue
3. increased antibody titers to E. histolytica
4. characteristic hepatic abscess
5. positive response to antiamebic drugs

Uninformative Stem Multiple-Choice (US-MC) Form - Example:

Trophozoites of E. histolytica

1. survive outside of the animal host, following passage in the stool
2. can readily be found in the stools of asymptomatic carriers
3. can live in the lumen or walls of the large intestine
4. usually produce gereralized inflammation of the intestinal wall

Negative Stem Multiple-Choice (NS-MC) Form - Example:

Antiviral chemotherapy is limited for all of the following reasons Except

1. much of the viral activity is dependent upon cell function
2. viral disease becomes evident only after extensive multiplication of the virus within the host
3. therapeutic disruption of viral multiplication is often accompanied by host cell death
4. the virus' proteins are the same as the host cell's proteins

Directions: For each of the questions or incomplete statements below, ONE or MORE of the answers or completions given is correct. On the answer sheet fill in the space under

1. if only A, B, and C are correct
2. if only A and C are correct
3. if only B and D are correct
4. if only D is correct
5. if all are correct

FILL IN ONLY ONE SPACE ON YOUR ANSWER SHEET FOR EACH QUESTION

Multiple-Answer Multiple-Choice (MA-MC) Form - Example:

Characteristics frequently encountered in virulent viruses include

A. ability to multiply well despite elevated host temperatures (fever)
B. poor induction of interferon
C. resistance to interferon inhibitory action
D. a DNA genome instead of RNA

Uninformative Stem Multiple-Answer Multiple-Choice (US-MA-MC) Form - Example:

Disseminated viral infections

A. involve only the target organ in viral multiplication
B. are caused by viruses which can find host cells (with appropriate receptor sites) only in the target organism
C. can be controlled by treatment with interferon at time of initial symptoms
D. may have several viremic stages

The MA-MC format was chosen for comparison because it is commonly used in medical education by classroom teachers in constructing tests used to make decisions about student performance. In a survey by the author of over 3000 items found on exams in the College of Human Medicine at Michigan State University, the average usage for this format was 17%; however, usage varied from under 5% to over 50% of the total items. What makes this particularly important is that these tests are graded on a pass-fail basis, using the same cut-off score for all Year One and Year Two exams. It is therefore important to know if items written in this format are comparable in difficulty and ability to discriminate to the more commonly used SA-MC items. This is a decision-oriented reason for studying this item format, to provide information useful in making specific educational decisions.

There is a second conclusion-oriented rationale for studying this type format. The idea of using a multiple-answer objective-type format has been discussed in the literature since Cronbach's article in 1939. Some authors have favored this general type while others have opposed its use. Research results (summarized in a later section) have been equivocal, partially due to the variety of formats included within this general item type. The MA-MC format as defined here provides one consistent format, used frequently at least in medical education, that can be studied in its natural environment. This particular format has been used frequently enough so that item data from items in many exams can be collected and used to compare this format to the single-answer format. This study should therefore be able to both provide information needed to make specific decisions and make a general contribution toward the further understanding of the performance of the multiple response type of multiple-choice item in comparison to the performance of the single response multiple-choice item.

CONVENTIONAL RULES FOR ITEM WRITING

Principles of Item Writing

There are three basic components of any multiple-choice test item: the item idea; the item stem (the statement posing the question); and the answer set (the possible answers). Each basic rule of item writing is directed toward one or more of these components. The rules are based on logic and common sense and in some cases backed up by

research findings. In general terms: the item idea should both precede and direct the writing of the test item. The item stem should clearly relate to the item idea, be concisely stated and direct the examinee to the answer set. The options or alternatives within the answer set should all be plausible possible answers to the question posed by the stem. Anything which gets in the way of clear communication of the intent of the item to the examinee, or inappropriately confuses or cues the examinee should be avoided.

The recommended rules for item writing have remained quite stable over time and consistent between various authors. This is reflected in the close similarity between Ebel's article, "Writing the Test Item," in the 1951 first edition of Educational Measurement, and the article of the same name by Wesman in the 1971 second edition. The set of rules used in the design of the present research reflects the overall body of suggestions and further reflects past research directed toward discovering the effect of rule violations on item quality. Following is a list of relevant item writing rules including the rule, the part of the item the rule is directed toward, and the relationship of the rule to the present study:

1. The item idea should be a single coherent thought. (Item idea) - Violation of this rule is not tested directly in this study.

2. The item should clearly relate to an objective of instruction. It should be worth testing and at the level of the students tested. (Item idea) - All the items used in this study were screened by medical faculty both before and after the tests were given. All items

with obvious content problems were eliminated from the item pools and from this study.

3. The stem should pose a specific question. (Item stem) - The basis for two research questions in this study.

4. The stem should be positively worded. (Item stem) - The basis for one research question in this study.

5. The item should be clearly and concisely worded so that the intent of the item is clear. (Item stem) - All items which were obviously confusing were eliminated during faculty review. However, items which had stems that relayed insufficient information were not eliminated and are being studied here.

6. One and only one correct answer should be included in the answer set. (Answer set) - Items with more than one correct answer were eliminated in the faculty review.

7. All responses should be grammatically consistent with the stem. (Answer set) - Items violating this rule were eliminated from the study.

8. Complex alternatives should be avoided as should the options "all of the above" and "none of the above." (Answer set) - Items violating this rule were eliminated from the study.

9. Avoid writing items where the correct answer is significantly longer and contains significantly more qualification than the incorrect responses. (Answer set) - Items violating this rule were eliminated from the study.

Rule Violations

Two rule violations were chosen for study in the present research: 1) items with negatively phrased stems, and 2) items with stems which are uninformative, i.e., stems which do not pose an answerable question.

Examples of negative stems:

- All of the following are characteristics of the nephrotic syndrome except:
- For treatment of all but one of the following poisons an emetic would be used. Mark the exception.
- Which of the following is not high in protein?

Examples of uninformative stems:

- Which of the following is true?
- Achalasia
- N. Meningitis

These two violations were selected primarily because of the relatively high level of incidence. In a survey of over 3,000 test items from exams given in the College of Human Medicine at Michigan State University, it was found that 40% of all single answer M-C items had at least one violation of item construction rules over 30% of the multiple-answer multiple-choice items had a violation. The two violations mentioned above were by far the most commonly occurring violations. Items displaying these two rule violations were found on every exam surveyed, although the prevalence varied from test to test. Table 1.1 indicates the overall frequency of occurrence of each of these violations for the single-answer and multiple-choice formats.

TABLE 1.1

FREQUENCY OF RULE VIOLATIONS
IN A SURVEY OF OVER 3,000 ITEMS ON EXAMS
IN THE COLLEGE OF HUMAN MEDICINE
AT MICHIGAN STATE UNIVERSITY

RULE VIOLATION	% OF SA-MC	% MA-MC
No Violation	60%	70%
Negative Stem	20%	-
Uninformative Stem	10%	25%
All Other Errors	10%	5%
	100%	100%

SOME HIGHLIGHTS OF PREVIOUS STUDIES

A review of the literature indicates that there are three basic methods of selecting the items to be used in comparisons of item forms:

1) New items - all items for each format compared are new and untried
2) Recast items -
   a. Format A - old items are used which have been screened for appropriate levels of difficulty and high levels of discrimination.

      Format B - The items from format A are recast into format B and are untried.
   b. Format A - old items, pretried

      Format B - recast items, also pretried
3) Old items - items are used for comparison as they appeared on the original tests.

An early study by Eurich (1931) illustrates the use of all new items. Eurich compared four item types, essay, completion, SA-MC, and T-F. In constructing the tests a traditional essay-type examination was prepared to cover the salient points of the subject matter. Specimen answers were then written out in detail for each essay question. Then using the essay questions and detailed answers, completion, SA-MC, and T-F items were written independently. The four subtests were then administered to the student groups and reliabilities (Odd-Even) were then computed. This was done for two separate courses, an educational psychology course and a statistical methods course.

The results in terms of reliability coefficients:

		Essay	Completion	SA-MC	T-F
Exp I	As Constructed	.69	.72	.71	.41
	If made 60 min.	.79	.84	.88	.74
Exp II	As Constructed	.56	.80	.75	.69
	If made 60 min.	.69	.89	.90	.87

In both experiments, completion and SA-MC type were the most reliable. No significance data were available.

This method is perfectly legitimate and results in comparisons of expert written items of varying formats. A study of this nature answers the research question, "How do items written in format A and format B compare in terms of difficulty and discrimination if the items are written by experts and content is closely controlled."

A study done by Oosterhof and Glasnapp (1972) illustrates the use of the first type of recast items. The authors chose 40 single-answer M-C items from a pool of 100 items, based on relationship to course objectives and previously shown high discrimination indices. They then created 40 new true and 40 new false items from these items. The new true-false items were created in a very legitimate manner using the stem of the multiple-choice questions plus the correct answer as the true items and the stem plus the most discriminating distractor as the false items. They then administered these 2 groups of items, old proven multiple-choice and new untried T-F to 101 undergraduates in an introductory measurement course. The results indicated that the multiple-choice items differed significantly from the true-false items in both difficulty and discrimination. When corrected for guessing, the M-C items were less difficult and more discriminating. This design tends to favor the old provem items no

matter what their format, because the items are usually selected on the very basis later used to compare the two item formats. The design was therefore avoided.

Frisbie (1973) also used recast M-C items to compare the multiple-choice and true-false item types. However, he included an intermediate phase where the T-F items were tried-out and changes made in accordance with the item analysis data. Results using paired t-tests showed the eight M-C subtests to be significantly more reliable than their counterpart T-F subtests even when time was considered. ($t = 5.405$, $p < .001$) In this study it was determined that per unit of time students could answer 3 T-F items for each 2 M-C items attempted. Frisbie (1974) was basically a replication of the earlier study except that the T-F subtests were lengthened by a 3.2 ratio in comparison to the M-C subtests. Results were similar to the earlier study; however the T-F subtests were significantly less reliable as well as less difficult. This method of recasting items is an entirely legitimate method of selecting items for study and clearly addresses the question, "How do items in Format A versus Format B compare in terms of difficulty and reliability when content is controlled by recasting items from Format A into Format B?" The question which remains unanswered is how typical items written in Format A versus Format B would compare based on difficulty and discrimination.

A study by Mueller (1975) illustrates the approach of comparing items on the basis of data obtained from the item analyses of the original tests in which the items appeared. Mueller used the data from several administrations of a real estate licensing exam to compare the

difficulty and discrimination of items with substantive alternatives versus items including the options "all of the above," "none of the above," or some complex combination of alternatives such as "1 and 2 are correct." He found that use of these options affected both difficulty and discrimination. Items containing complex alternatives were the most difficult while items with the alternative "all of the above" keyed as correct were the least difficult. Discrimination, though less affected, showed items with only substantive alternatives to be most discriminating, while items with the alternative "none of the above" keyed as correct were least discriminating.

Of the three selection methods, this last approach is most conducive to the study of item formats as they are actually used within classroom tests. It was therefore the approach chosen for use in the selection of items to be included in this study.

NEED FOR FURTHER RESEARCH

As will be evident in the Review of Related Research a very small number of studies have addressed the questions posed in the present research. Further, the majority of studies intended to compare item formats have used the recasting method of item selection. Studies of this type have left unanswered the general question of how items of any two formats under study would compare in terms of average difficulty and discrimination, if all of the items used were written by actual classroom teachers for use in their own examinations.

What is proposed is a study in which the items and formats to be compared will come from actual tests used in the College of Human

Medicine at Michigan State University, at some time in the past several years. There is a need for evidence from actual classroom testing situations, using items written by typical medical school teachers, comparing specified item forms to determine format effectiveness and impact of rule violations. Evidence collected in this manner meshes well with that collected using professionally prepared or carefully recast items as the basis for comparing separate formats or item writing principles.

CONTEXT OF THE STUDY

All of the items used in this study came from the College of Human Medicine at Michigan State University and were taken from item pools composed of test items used in regular classroom tests on each subject area. Nine subject areas were used. Six of them were Focal Problem areas used with first and second year medical students as a means of testing specific basic science knowledge. A Focal Problem is a major presenting complaint which is used as a focus for studying related anatomy, physiology, microbiology, pharmacology, etc. The six Focal Problems used in this study were Altered Consciousness, Elevated BUN (blood urea nitrogen), Anemia, Chest Pain, Diarrhea, and Jaundice. The other three subject areas were the clinical science areas of Pediatrics, Internal Medicine, and Surgery. The clinical science items were all used on tests given as final exams in third or fourth year required clinical clerkships. Nine subject areas were used both to provide a cross-section of content areas within medical education and to provide a sufficiently large number of items of each

type under study to assure reasonable stability of the experimental results.

All of the items in the nine item pools had been screened by the faculty for serious content or wording problems both before and after the initial exam was given. All severely defective items were removed from the item pools and therefore not included in this study. Further, all items with item writing errors not under study were excluded from the selection process. The items used in each of the four comparisons were matched as closely as possible for number of options, content and exam of origin.

SPECIFIC QUESTIONS TO BE ANSWERED

All four research questions are identically stated, the only difference being the particular comparison specified.

The General Question: How do the two specified groups of items compare in terms of mean values for item discrimination and difficulty where: a) number of options, exam of origin, and content objective have been controlled, and b) the items and item statistics are taken from the original administration of actual classroom tests.

The Four Comparisons:

I. Single-Answer Multiple-Choice items without rule violations (SA-MC) vs.
Multiple-Answer Multiple-Choice items without rule violations (MA-MC)

II. Single-Answer Multiple-Choice items with positive stems (SA-MC) vs.
Single-Answer Multiple-Choice items with negative stems (NS-MC).

III. Single-Answer Multiple-Choice items with informative stems (SA-MC) vs.
Single-Answer Multiple-Choice items with uninformative stems (US-MC).

IV. Multiple-Answer Multiple-Choice items with informative stems (MA-MC) vs.
Multiple-Answer Multiple-Choice items with uninformative stems (US-MA-MC).

SUMMARY

The purpose of this study is twofold: first, to compare two item formats commonly used in medical education, and, second, to compare items which violate specific item writing principles to items without violations. Comparisons will be made on the basis of average difficulty and discrimination indices for the groups of items under study.

The two formats chosen for comparison in this study are the traditional single-answer multiple-choice format and a multiple-answer multiple-choice format used primarily in medical education and known as Type K on the National Board of Medical Examiners' examinations. The two rule violations chosen for study both deal with the manner in which the question is posed in the stem of a test item. The violations are: 1) use of negatively worded stems, and 2) use of incomplete stems, stems which do not state specifically the question to be answered. The first error results in items of the form, "All of the following diseases have killed virus vaccines available except. . ." and require the examinee to choose the exception. The second type are of the form "Which of the following is true?" and require the examinee to discern the intended question for himself after reading the options.

CHAPTER II

REVIEW OF RELATED RESEARCH

Many studies have been conducted to compare item types. The earliest studies comparing item types took place in the 1920's and were intended to compare the "new type items" to essay type questions. The study by Eurich (1931), described earlier, is typical in the choice of formats used for comparison: essay, completion, single-answer multiple choice, and true-false. Most of the more recent studies comparing item formats have concentrated on comparing the single-answer multiple-choice (SA-MC) format to either regular true-false (T-F) items or to some format which is a variant of either the SA-MC or T-F formats. The Frisbie studies (1973, 1974) detailed earlier, are recent examples of studies comparing the SA-MC and T-F formats. In the present study single-answer multiple-choice items will be compared to multiple-answer multiple-choice (MA-MC) items. Further, SA-MC and MA-MC items without rule violations will be compared to items with specified common violations.

MULTIPLE-ANSWER FORM

Study of this type of item format began over thirty years ago. An early positive mention appeared in an article by Cronbach in the Journal of Educational Psychology (Cronbach, 1939). In this article he advocated the use of an item format he called the multiple true-false format. The items are set up like regular single answer multiple-choice items, but each option is answered like a separate

true-false test item. Cronbach (1941) also conducted a study which compared the use of the multiple true-false format (all options marked with T or F) and the multiple multiple-choice format (only the true options are marked). The study disclosed little difference between these formats in terms of testing time, reliability, or validity. Already in this early study there are two variatons of the multiple-answer item format, neither of which is the MA-MC format used in the present research. Also, these item formats were compared only to each other and not to the single-answer format.

In a study reported by Albanese, Kent and Whitney (1977), the MA-MC format (national board type) was compared with two other multiple response item types using the same stems and options as the MA-MC items but answered like four separate T-F items. The MTF format was scored as 160 separate T-F items. In the MR type only those options that were felt to be true were marked and an additional option "none of the above" was also included. The MR items were scored as 40 individual multiple-choice items where credit was given only where the correct combination of true options was chosen. (The option "none of the above" was never a correct answer.) The findings were that the MR type was significantly more difficult than the MTF and the MA-MC even after difficulty level was corrected for chance, and that the MTF section was significantly more reliable than the MA-MC section only. The reliability results are not particularly surprising. It might be expected that a subtest of 160 T-F items would have a higher reliability than a 40 item MA-MC subtest. Unfortunately, although the authors mentioned inclusion of SA-MC items in

their design, they did not include these items in their analysis. Further, neither the MR nor MTF item type is practical for use in combination with SA-MC items, the MR type because it requires a different answer sheet and scoring procedure and the MTF because of probable problems with content balance. Therefore, although the results of the study are interesting they do not relate directly to the present research.

Dryden and Frisbie (1975) conducted a study designed to compare the multiple-answer multiple-choice format with the single-answer format. They took 64 multiple-answer items from a General Nursing Exam and converted these to SA-MC items. Their results showed that 25% more SA-MC items were answered per unit of time; and that SA-MC tended to be slightly more reliable; there was no significant difference in difficulty between the two types. However, most of their SA-MC items contained either complex alternatives like "a and b" or "all but c" or the option "all of the above." In addition, it appeared that the multiple-answer format was not one set format (like the traditional national board format) but a conglomeration of multiple-answer formats. Their use of this variety of format variations within both the single-answer and multiple-answer items makes it difficult to assess the meaning of their results and hazardous to generalize from them.

The piece of research that relates most directly to the study of the SA-MC item versus the MA-MC item was done in Canada by Shakun, et al., (1977). Part of their study was the comparison of the traditional single-answer multiple-answer format with the multiple-answer multiple-choice format (called Type k on the National Board of Medical

Examiners exams). Twenty items of each type were prepared by one of the authors to be parallel in the sense of testing the same content or basic concept. The items were then reviewed by the General Surgery Test Committee and revised where necessary. Eighteen items of each type, designated as the experimental items, were selected for inclusion in the 1976 General Surgery Certifying Exam of the Royal College. The items were then interspersed among other items of the same type in either Paper I or Paper II of the Exam (191 total SA-MC items; 111 total MA-MC items). The results for the experimental items showed the following:

	MA-MC (Type K)	SA-MC
p (percentage getting the item correct)	.69	.71
KR (reliability of subtest) 20	.52	.52

There was essentially no difference in the performance of the two item types. The major problem with this study is that 18 items is a very small number of items from which to generalize. If all items of each type in the total exam are considered a larger difference in average difficulty is noted; SA-MC, $p = .73$, MA-MC, $p = .65$. No test of significance was reported for these data, however. Also no item discrimination or subtest reliabilities were reported for the larger group of items. This study is a good example of the type of study which compares two item types on the basis of professionally written new items. All of the experimental items of both types were well written and prescreened, but not pretested for difficulty or discrimination.

The present research used comparisons similar to the one referred to above, using all of the items of both item types. In making the comparison groups all available items of each type from the specified item pools were used. As many item pairs as possible were matched, controlling for number of options, exam of origin and content objective. The research question here is: How do items of each type, as typically written and used, compare in terms of difficulty and discrimination?

NEGATIVE STEM FORM

The argument against the use of negatively worded stems is mainly logical. First, since examinees are used to thinking positively and are generally asked to choose correct answers, items calling for an incorrect response can be confusing (Mehrens and Lehmann, 1978). Second, since questions of this nature are rarely encountered in the real world, they lack practical relevance (Ebel, 1979). Finally, evidence collected from the study of true-false items indicates that negatively stated items can take longer to answer than positively phrased items (Wasen, 1961; Zern, 1967).

Research actually comparing the performance of items with negatively phrased stems to that of items with positive stems is somewhat limited. Terannova (1969) conducted a study using two option multiple-choice items to address the issue of negative versus positive stems. Three independent variables were used, positive and negative stems, frequency of change of direction set (0, 1, and 2 changes) and grade level of the student (5, 7, 9, and 11th grades). Six experimental instruments were prepared along with a common

instrument. The common instrument was given to all subjects while the experimental instruments were assigned randomly to the subjects. Two three-way analyses of covariance were used for analysis. The results indicated a significant difference in difficulty level (at the .95 level of confidence) based on stem type. Negative stem items were more difficult than positive stem items. There was no significant difference in the reliability of any of the experimental subtests.

Dudycha and Carpenter (1973) studied three variations of the SA-MC item type. One of these variations was the positive versus negative stem. (The other two were open-ended statement-type stem versus closed question type stems and the presence or absence of the option "none of the above".) Sixty-four items (all positive, closed, without the option "none of the above") were chosen from a group of 96 items on the basis of item analysis data and course instructor evaluation of content appropriateness. These items were then recast with great care to reflect all eight possible variations (i.e., positive, closed, without "none of the above" through negative, open, with "none of the above"). Sixteen separate experimental tests were written and administered randomly to 1124 students as the final exam in an introductory psychology course. Each test had one half negative and one half positive items. A 2x2x2 fixed-effects, repeated measures ANOVA was performed on item difficulties. The results indicated that negative stems were more difficult than positive stems ($F = 14.71$, $p < .001$). There were no interaction effects between the three factors studied. No significant differences were found between the average discrimination indices of items with positive versus negative stems.

Both of these studies provide evidence of a difference in difficulty between negative and positive stemmed items, when the negative items are recast from positive items of proven high quality. Both, however, leave unanswered the question addressed in the present study, that of the difference between positively phrased and negatively phrased items as they appear on typical classroom exams. The present study is also within a different content area and at a different level than the two previous studies, i.e., medical education, with students on a post-bachelor's degree level.

UNINFORMATIVE STEM FORM

The single-answer multiple-choice item with an uninformative stem can be looked at in several ways. It can be thought of as a conglomeration of separate true-false items or as a multiple-choice item which violates one or both of the following rules:

1. An item should test only one coherent thought.
2. The item stem should pose a direct question in a concise manner so that the meaning of the item is clear.

Ebel (1978) conducted a study which looked at this item type as a conglomerate of separate true-false items. First, a test consisting of 100 separate true-false items was prepared. Then these items were arranged to form a second test with 2 two-option items and 32 three-option items (23 with the stem "which statement is true" and 9 with the stem "which statement is false"). The two tests were then administered as Part I and Part II of a midterm examination of two classes of graduate students in education. The results of a comparison of the scores on the two parts showed: (1) that the test using separate true-false items was more reliable than the "grouped"

test for both groups, and (2) the individual "grouped" items were more difficult than the individual true-false items. No significance data were reported; however, this study points out two things: (1) a definite loss in information occurs if separate true-false items are combined to form MC items with uninformative stems, and (2) the resulting test would probably be less reliable.

In the present study, the M-C item with an informative stem is looked at as a regular single-answer M-C item which was written incorrectly. Some authors of classroom test items are under the mistaken impression that if one concept in an item is good, two is better, four better still. One solution would be to simply reverse the process used in Ebel's 1978 study and make each of these items into separate true-false items. However, this process could easily result in a different rule violation.

> Rule: The content being tested should be worth testing. It should be important knowledge.

In general, when medical faculty edit and rewrite test items, it is more common for the faculty to find no concepts really worth testing in items of this type than to find four concepts all worth testing. Many items of this type seem to be the result of unclear or disorganized thinking on the part of the item author. However, some items with uninformative stems do have one basic item idea. In these cases, the error is that the stem does not pose a direct question, forcing examinees to look for the question among the options as well as looking for the answer to that question. In the first case, the item idea was unclear so no specific question was being asked by the item. In the second case, a specific question was intended but the stem did not

pose the question appropriately. For the purposes of the present study, these two variations are considered together as items with uninformative stems.

Before going on it is important to make the distinction between items which have uninformative stems, as described above, and items which merely have open ended stems. Open ended stems are stems which use an incomplete declarative sentence to pose a question instead of actually using an interrogative sentence. In this case, there is no loss of content, merely a difference in form.

Examples:

Question Format: Which of the following should be used for initial treatment of an anaphylactic reaction?

Incomplete Statement Format: An anaphylactic reaction should be treated initially by use of:

Uninformative Stem: An anaphylactic reaction.

In the present study, items using either the question or incomplete statement formats are classified as acceptably written items. The issue in the present study is the presence or absence of relevant information in the item stem, not the superficial form of that stem.

Four studies from the literature provide background for the present research, although none of them focus specifically on the question of major interest in the present study. Dunn and Goldstein (1959) studied the effect of each of four rule violations. One of their comparisons was between items with closed stems (questions) and items with open stems (incomplete statements). The authors stated that for convenience the open stem was designated as the "rule." The procedure for developing the experimental instruments was to rewrite acceptably

constructed items to reflect one or more rule violations. The items were obtained from Army Basic Military Subjects Tests. Four 100 item tests (A, B, C, and D) were constructed for their Series A experiment which combined the study of open versus closed stems with the study of the use of cues and specific determiners. Each test had 25 items reflecting each of the following combinations: open - no cue, question-no cue, question - cue, open - cue. No significant differences in difficulty or KR 20 reliability coefficients were found between groups of items with open versus closed stems.

The Dudycha and Carpenter (1973) study described earlier also investigated the open versus closed stem. In this study, the closed stem was used as the "rule." The open-ended experimental items were carefully recast from previously used well-written closed stem items so that there was no loss of information in the stem. The results indicated that items with open stems were more difficult than those with closed stems. ($F = 5.61$, $p < .05$) There was no difference found in average item discrimination between the two groups.

Both of these studies were focused primarily on the form of the stem, not the completeness of the content. Both ask the question, "Does a change in the form of the stem alone affect the average difficulty or discrimination indices for multiple-choice items?" Since these studies produced differing results the question is not answered definitively. Why the difference? One possible explanation might be in the type of stem found in the original items used for study. In the Dunn and Goldstein study, it appears that the original high quality items had open stems (incomplete statements) which were altered to question format

for comparison, while in the Dudycha and Carpenter study the opposite was true. This is a significant difference and could explain the results. Logically, it would seem that if the basis for comparison of two groups of items is going to be difficulty and item discrimination that either both sets of items or neither set should be selected on the basis of previous difficulty and discrimination values. This would involve either selection of original items on some other basis (i.e., using all available items, random selection, content appropriateness, etc.) or using a tryout to obtain item statistics for the recast items. This was not done in the above studies. These two studies serve as a stimulus, as indication of interest in studying the general topic of the effect of variations in the wording of positively phrased item stems. However, because their focus was on the form of the stem alone, they do not relate directly to the present research.

The last two studies reviewed did address the issue of informative versus uninformative stems and were both done by the same authors, Cynthia Board Schmeiser and Douglas Whitney. The first study (Board and Whitney, 1972) was designed to compare a group of well written test items to groups of items containing one of four violations of accepted item writing principles. Thirty items were chosen from a sixty item midterm exam in an Introduction to America Politics course at the University of Iowa. The items were chosen on the basis of their difficulty (40% - 70%) and discrimination indices (.3 or better) and their adaptability to being recast using the rule violations selected. Three 30 item experimental tests were written: Test I had the 30 well written midterm items. Test II had 13 items with "window dressing" and 17 items with incomplete stems, all recast from the original 30 items.

Test III was composed of 15 items with distractors made systematically longer or shorter than the correct answer and 15 items where the only option grammatically consistent with the stem was the correct answer. The authors did not state whether the original items were in the form of a question or an incomplete statement. The focus of the study was not the form as in the two previous studies but the content. When the items were recast, the stems were severely truncated to make the stems "grossly incomplete." All three of the experimental tests were administered along with a sixty item final exam. The results indicated that the items with incomplete stems were more difficulty than the original items ($F = 4.42$, $p < .05$). The analysis also indicated that compared to the subtest of original items the subtest of items with incomplete stems had a lower reliability using Feldt's approximation of the F test for comparing KR 20's ($W = 1.52$, $p < 05$).

The most recent study used a slightly different method. Schmeiser and Whitney (1975) first made an extensive search through teacher constructed exams at the University of Iowa's Evaluation and Examination Service, where they found a number of examinations in which at least one-fourth of the items contained incomplete stems. (This indicates that this item fault is a common error outside of the medical area as well as within medical education.) They chose for study purposes a sociology exam where 22 of the original 61 items had incomplete stems. They then rephrased these stems to make 22 comparison items. Following is an illustration of their method:

Incomplete stem: Free will

Rewritten stem: Which of the following statements characterize f will?

The items with incomplete stems were found to be significantly more difficult ($F = 23.88$, $p < .001$) but did not differ in average discrimination (using D, the upper-lower index).

The first study used the traditional method, criticized above, of taking proven good items and recasting them to produce poor items. The second study goes several steps further, first by locating items that already contained an error, second by using all available items in the recasting process. This study provides definite evidence that items with uninformative stems tend to be more difficult than items with complete-informative stems.

The present study contributes a different type of evidence about this item writing error than the past studies which all used recast items. A comparison will be made between approximately 50 items of each type (uninformative versus complete-informative stem) taken from actual classroom exams in medical education. The item pairs are matched for content objective and exam of origin.

NUMBER OF OPTIONS

The number of options in an item can affect the difficulty of the item with items with fewer options being easier than those with more options. Burmester and Olson (1966) recast 85 SA-MC items from a college natural science course into 85 true-false items. Items where student selection of incorrect answers was spread evenly across all distractors, were made true (using stem and correct answers). Items where one distractor was chosen most frequently were recast as false (using that distractor with the stem). The original SA-MC items had a mean difficulty (p value) of .57 and discrimination of .45 (Flanagan index). For the

recast items the figures were .71 and .41. Their conclusion was that the true-false items were easier but similar in discrimination. No significance data were given. Several other studies described earlier (Oosterhof and Glasnapp, 1972; Frisbie, 1973; Frisbie, 1974) further substantiate the claim that items with fewer options can differ significantly in difficulty. Further, two of these studies (Oosterhof and Glasnapp, 1972; Frisbie, 1974) also showed a significant difference in reliabilities between groups of items with two versus four options. These studies represent the extreme comparison based on differing numbers of options; i.e., true-false with two options versus multiple choice with four or more options. Several other studies have compared groups of items more similar in number of options.

Williams and Ebel (1957) using the vocabulary section of an Iowa Test of Educational Development recast 150 four-option SA-MC items into three option and two option items by dropping the least discriminating distractors. Following are their results:

	4 choice	3 choice	2 choice
Index of Difficulty (p - % getting item correct)	.50	.58	.68
Index of Discrimination (upper-lower)	.49	.47	.41

A fairly large decrease in difficulty with each option dropped was accompanied by a much smaller decrease in discrimination.

Several earlier studies conducted by Ruch & Stoddard (1925) and Ruch, De Graff, et al. (1926), also compared items with differing numbers of options with results similar to that of Williams and Ebel. In the Ruch and Stoddard study, two forms of a fifty-item history and social

science test were used, each item in each form having been prepared as a five-option, three-option and two-option multiple-choice item. Three groups of high school seniors, 135 students in each, were tested, one group being given both forms of the five-response test, one both forms of the three-response test and one both forms of the two-response test. Reliabilities were computed using the scores on the two forms. Both p-values and reliability coefficients increased with the increase in number of options. The Ruch, De Graff, et al., study was very similar with the exception of the addition of seven-option items. The results of this study showed a similar increase in p-values and reliabilities from the 2-option to 3-option to 5-option items. However, a decrement in reliability resulted when 7-option items were used.

All of the studies cited indicated that items differing in number of options also differ in resulting item statistics. For this reason, the average number of options per item has been controlled for all comparisons to be made in the present research.

OTHER RULE VIOLATIONS

Use of complex alternatives:

Items including the options "all of the above," "none of the above," or variations of the complex alternative "Both 1 and 2 are correct" were all eliminated from the present research. It is necessary to control for possible effects because of evidence which suggests that the presence of any of these options can affect the difficulty of the test item. Further, since items with these options appeared only sporadically it would be very difficult to control for their effect by

any means other than elimination. All seven of the relevant studies in the literature found some increase in difficulty with use of these options (Meuller, 1975; Dudycha & Carpenter, 1973; Williamson & Hopkins, 1967; Hughes & Trimble, 1965; Rimland, 1960; Boynton, 1950; and Wesman & Bennett, 1946). Five of these studies concentrated on the option "none of the above," while the other two looked at all three kinds of complex alternatives. Both the Meuller (1975) study and the Hughes and Trimble (1965) study compared items with each type of complex alternative to regular SA-MC items. The Meuller study compared large groups of items from actual exams used in a Real Estate Salesmen's Course. The results seem to indicate increased difficulty for items with combination alternatives, e.g., 1 and 2 are correct, as compared to items with only substanative alternatives. There were small differences in average discrimination indices for the varied groups of items with items containing only substanative options showing the highest average value. Since there was no report of significance tests, a summary of the results is provided below:

	ITEM TYPE			
	Substanative alternatives only	None of the above	All of the above	Combination alternatives
# of Items (k)	91	94	79	45
Average Difficulty (p)	.79	.74	.78	.64
Average Discrimination (r Pt. Bis.)	.30	.27	.27	.26

Hughes and Trimble (1965) used four option items originally written by the author and previously tried out in an Introductory Psychology course as control items. The items were recast to make comparison items by adding a fifth option; either "none of the above," "all of the

above," or "Both 1 and 2 are correct." Three experimental tests were made up:

Test 1 - 50 regular 4 option items (15 control-35 used in comparisons)

Test 2 - 15 regular control items; 35 items with "Both 1 and 2 above are correct"

Test 3 - 15 regular control items; 17 items including "none of the above;" 18 items including "all of the above."

The three tests were distributed to randomly selected subgroups of 26 students each. Results indicated in increased difficulty for the comparison items in both Test 2 and Test 3.

Test 1 (control		Mean = 27.19	
Test 2 (Both 1 and 2)		Mean = 22.05	
Test 3 (All or None)		Mean = 23.76	
F = 7.05, p .01			
Dunn's test - comparing Means	1 with 2	C = 3.60	p < .05
	1 with 3	C = 2.40	p < .05

There was no significant different in the reliabilities of the three experimental tests. One problem with estimating the meaning of the results of this study relates to the method of recasting items. The authors recast these items by increasing the number of options from 4 to 5. Therefore, the increase in difficulty could have been due to the content of the fifth option or to the mere presence of an extra option. The authors themselves admitted this was a problem in their procedures section but gave no explanation why it was done this way.

Despite the design problems in the second study both studies do provide evidence that the use of any of these kinds of alternatives can affect

mean item difficulty. Further evidence is provided by additional studies which included only the option "none of the above" in their design.

In the Dudycha and Carpenter study described earlier, sixteen experimental tests (64 items each) were written and administered randomly to 1,124 students as the final exam in an introductory psychology course. One-half of the items on each test included the option "none of the above," the other half had substanative alternatives only. The results indicated the items which included "none of the above" were more difficult ($F = 81.54$, $p < .001$). There were no significant interactions between the three factors tested (open vs. closed; negative vs. positive stems were the other factors). In reference to item discrimination, inclusion of the alternative "none of the above" decreased the discrimination ability of an item ($F = 17.19$, $p < .001$). Four earlier studies gave similar results in terms of difficulty but provided no evidence of a decrease in reliability or validity associated with the use of this option. Williamson & Hopkins (1967) used four standardized arithmetic tests with 345 fourth grade students recasting items in two of the tests to include "none of these" and recasting items in the other two tests to exclude that option. The results indicated a significantly higher mean difficulty for items including "none of these" on two of the four tests but no difference in the other two. Rimland (1960) used the Navy Arithmetic Test with 3,600 Navy recruits, recasting items to include the "right answer not given" option. Results showed a small but significant difference in difficulty with items including the "right answer not given" option being more difficult. Boynton (1950) studying the affect of the "none

of these" option with spelling items found an increase in difficulty. Finally, Wesman and Bennett (1946) conducted an exploratory study using only 20 vocabulary and 20 arithmetic items with 590 applicants as part of a test of admission to nursing schools. They found that 17 out of 20 of the vocabulary items increased in difficulty with the use of the "none of these" option while there was no difference in difficulty on the arithmetic section.

All seven of the studies mentioned substantiate at least to some degree the decision to eliminate items with complex alternatives to avoid the possibility of a variable confusing research results due to an uncontrolled effect on either difficulty or discrimination.

Grammar and Length Faults:

Grammar and length faults both refer to errors in the item options. An item has a grammar fault if the options are not grammatically consistent with the stem and a length fault if the correct answer is significantly longer than the incorrect options.

The relevant research studies (Dunn and Goldstein, 1959; McMorris, et al., 1972) both investigated the effects of each of these errors on average item difficulty and discrimination. The earlier study used Army enlistees as subjects and the Army Basic Military Subjects Test as the basis for recasting items to include faults while the later study used high school students and an American History Test. In both studies, faults were found to make the items easier; however, validity and reliability coefficients remained unchanged.

The Dunn and Goldstein (1959) study was discussed earlier. In the portion of the study which relates to these item writing errors, the authors used items which were recast to contain the fault in preparing four tests, E, F, G and H to form Series E in the experiment. Each test included 25 items belonging to each of the following groups: equal choices - good grammar; unequal choices - good grammar; equal choices - poor grammar; unequal choices - good grammar. An analysis was done comparing average difficulty and discrimination for the 4 groups of items. The results indicated that the errors resulted in easier items but had no effect on reliability or validity indices. On all four experimental tests, E, F, G and H, the four groups of items ranked in the same order based on difficulty: 1) Items which included both errors - easier, 2) Items containing a length error only, 3) Items containing a grammar error only, and 4) Items with no error - most difficult.

The McMorris, et al. (1972), study investigated three of the four errors looked at by Dunn and Goldstein. These were: length errors, grammar errors, and cue errors, where a cue to the correct answer is included in the stem. The authors first wrote, then pretested a set of well-written test items covering the New York Board of Regents objectives for American History. They then chose 42 items with difficulty indices between .25 and .85 which were recast to include one of the three errors. The resulting experimental instruments were two 42-item tests which were as identical as possible except for the faults. On each test, there were 21 items without faults and 7 items with each of the three errors. The tests were administered to 494 students with alternating students getting forms A & B. Within the

analysis, comparisons for each error were based on the 14 fault free items and their counterpart 14 faulted items. Differences in difficulty were tested by means of confidence intervals around the mean difference score (average number of students getting the faulted items correct minus average number of students getting the fault free items correct). The results indicated that both the length and grammar faults decreased the difficulty of the items at the 95% level of confidence. The analysis indicated no differences in the subgroups based on comparison of reliability and validity indices. Both of the studies reviewed provided evidence to indicate that inclusion of items with grammar or option length errors can affect overall difficulty of groups of items. These items were eliminated from consideration in item selection in an effort to avoid as many potentially confounding factors as possible.

Fault within NS-MC and MA-MC Types Relating to Content of Options

Elimination of items with this final error was based, not on past research but on logic as described below. This error is particularly to the negative stemmed and MA-MC item types. These are items where the correct answer and its direct opposite are both included in the set of alternatives thus automatically eliminating all other options. Recall that with single answer items with negative stems the examinee is looking for the exception. If two of the options contain statements that could never both be true at the same time (e.g., alkalosis and acidosis or hypothermia and hyperthermia) one must be the exception asked for. With the MA-MC item type an error of this kind could eliminate from one to three alternatives depending on which options were involved. The only error of this type actually found was where

the item author had included 2 correct options (1 & 3 or 2 & 4) and 2 incorrect options (1 & 3 or 2 & 4) in such a manner that all of the options are eliminated except option B (1 & 3 correct) and C (2 & 4 correct). With a positively worded SA-MC item this error cannot occur because a statement and its opposite are legitimate options. Both could be irrelevant to the question being asked or one of them could be the correct answer. The point is the examinee does not get an irrelevant cue to the correct answer. The effect of any error of this type on difficulty and discrimination is unknown so items with this error were eliminated.

SUMMARY

In this chapter the literature relevant to the general research questions of this study was reviewed. This included literature related to study of the multiple-answer format, and to the effect of selected item writing rule violations and the number of options per item. It is evident from this review that there are few past studies relating directly to the specific research questions of the current study.

In Chapter III the research procedures and design of the present study are detailed. Also the specific research hypotheses are stated.

CHAPTER III

PROCEDURES AND DESIGN

INTRODUCTION

The purpose of this research is to compare groups of test items, selected on the basis of format or item writing rule violation, for psychometric quality based on data collected from administration of regular classroom exams within medical education. Five item types have been identified for study as have four relevant comparisons between them. All five item types are variants of the basic multiple-choice item format. The measures used as a basis for the comparison of designated item groups are one estimate of item difficulty, p, proportion getting the item correct, and two estimates of item discrimination, the upper-lower index, D, and the bi-serial correlation coefficient, rBis.

This chapter includes a description of: the source of the test items used, the item selection procedures, including a discussion of the method for control of content, the statistics used as measures of item quality, the design of the study, the hypotheses to be tested, and the statistical procedures to be used to test these hypotheses.

SOURCE OF ITEMS

A total of 718 test items were used in the present study. Table 3.1 provides a breakdown showing the number of items used in each of the four research comparisons. Since the pairings for each of the four comparisons were done independently, some of the individual SA-MC and MA-MC items are included in more than one comparison.

All of the items used come from the college of Human Medicine at Michigan

TABLE 3.1

NUMBER OF ITEMS USED IN EACH OF THE FOUR RESEARCH COMPARISONS

Comparison	SA-MC[a]	MA-MC	NS-MC	US-MC	US-MA-MC
SA-MC vs. MA-MC	124	124			
SA-MC vs. NS-MC	143		143		
SA-MC vs. US-MC	51			51	
MA-MC vs. US-MA-MC	___	69	___	___	69
TOTAL NUMBER	284	171	143	51	69

[a]The total number of SA-MC and MA-MC items is less than the number of each typed used in all comparisons because some of the SA-MC and MA-MC items are used in more than one comparison.

State University and were taken from item pools composed of test items previously used on regular classroom tests in each subject area. Of the nine subject areas used, six were focal problems areas: Altered Consciousness, Diarrhea, Anemia, Jaundice, Chest Pain, and Elevated BUN. Each focal problem is a major presenting complaint or symptom, used at Michigan State as the focus for a three to five week course of study of related basic, clinical and behavioral sciences. At the end of each of these mini-courses a content exam is administered. The focal problem items used in this study originated on these tests. The final three subject areas were the clinical science areas of Pediatrics, Internal Medicine and Surgery. The clinical science items used in this study all originated on tests given as final exams in these required clinical clerkships. Table 3.2 shows the source of all items pairs by content area and comparison. In the total study 387 pairs of items were selected: 124 in the SA-MC vs MA-MC comparison, 51 in the SA-MC vs US-MC comparison, 69 in the MA-MC vs US-MA-MC comparison and 143 in the SA-MC vs NS-MC comparison. This table further indicates the total number of item pairs used in the study from the focal problem area versus the clinical science areas. These 387 item pairs originated on 40 separate exams. Appendix A provides a complete listing by content area of the 40 exams used, along with the number of students tested and the number of item pairs selected from each exam for each of the four comparisons. The Pediatrics and Internal Medicine exams were administered and analyzed as combined exams throughout the entire period of this study so each exam appears under both content areas. However, although the exam items were combined into one test the preparation was completely separate, with the content being submitted to and reviewed by separate committees within the respective departments.

TABLE 3.2

SOURCE OF ITEM PAIRS BY CONTENT AREA AND COMPARISON

Content Area	Comparison SA-MC vs MA-MC	SA-MC vs US-MC	MA-MC vs US-MA-MC	SA-MC vs NS-MC
Focal Problems				
1. Altered Consciousness	12	3	11	8
2. Diarrhea	13	5	14	15
3. Anemia	7	7	7	19
4. Jaundice	13	6	10	10
5. Chest Pain	3	3	4	1
6. Elevated BUN	17	9	8	18
Subtotal Focal Problems	65	33	54	71
Clinical Science Areas				
7. Surgery	20	6	3	14
8. Pediatrics	17	9	5	26
9. Medicine	22	3	7	32
Subtotal Clinical Science	59	18	15	72
TOTAL	124	51	69	143

All of the items within the nine item pools had been screened by the faculty for serious content or wording problems, both before and after the initial administration of each exam. All severely defective items were removed from the item pools and therefore excluded before the item selection procedure for this study began. Further, all of the items included in each item pool were keyed to a topic area within the overall content area. This keying was later checked by another faculty member, group of faculty members, or a specially trained medical student. Table 3.3 displays the entire list of topics used by faculty in categorizing items for each of the content areas. The topic list was the same for all focal problems. It can also be noted that most of the topics are either basic science areas relating to the focal problems or subspecialties within the general clinical science areas.

ITEM SELECTION AND CONTENT CONTROL

Specific Selection Procedures - Six Focal Problem Areas

The College of Medicine at Michigan State University has a special Focal Problem Tract for first and second year medical students. Approximately 40 medical students in each class of 100 students participate. Over the two year period the students learn the same overall basic science material as students in the more traditional program; however, they learn it within the context of focal problems. A focal problem is a common general medical finding or symptom. Chest pain, abdominal pain, elevated BUN, anemia and jaundice are all examples. The students study these problems intensely, one at a time, for a three to five week period. Study is done independently and is based on a list of basic concepts and suggested references. At the end of the designated time the students take an achievement test (approximately 150 items) where every student

TABLE 3.3

LIST OF TOPICS USED FOR ITEM SELECTION WITHIN THE FOCAL PROBLEM AREAS AND SEPARATE CLINICAL SPECIALITIES

List for Focal Problem Areas	List for Pediatrics	List for Medicine	List for Surgery
Anatomy	Behavioral	Cardiology	Lumps
Behavioral Science	Cardiology	Endocrine/ Metabolic	Hernia
Biochemistry	Emergency Conditions	Gastro-intestinal	Abdominal Pain
Clinical Science	Endocrine/ Metabolic	Hematology	Abdominal Mass
Histology	Genetics/ Birth Defects	Immunology	GI Bleeding
Microbiology	Growth & Development	Infectious Diseases	Peripheral Vascular
Neurology	Gastrointestinal	Nephrology	Injuries & Burns
Pathology	Genitourinary	Neurology	Neck Masses
Pharmacology	Hematology	Oncology	General Procedures
Physiology	Immunology/Skin	Pulmonary	
	Infectious Diseases		
	Newborn		
	Nerves & Muscles		
	Respiratory & ENT		

is allowed sufficient time to answer every item. No correction for guessing formula is applied. Each test is very homogeneous in content because it is designated to measure the knowledge and understanding of concepts which all relate directly to one central medical finding. The items selected for use in this study had been screened by the faculty for obvious content or wording problems, but were not pretested or screened on the basis of item statistics. Each of the items was keyed to a specific topic area. Content was controlled by using pairs of items from the same test administration, which were further keyed to the same topic area. The initial keying was done by the physicians, basic scientists and behavioral scientists who wrote the items, but was also rechecked by the author. Item pairs were only made when the author's judgement and the item writer's judgement coincided. The final result is: A set of item pairs, each comprised of two items from the same test, one of each type to be compared, which were written independently and keyed to the same basic topic area. Selection of the pairs of items was based solely on item type, number of options and content similarity.

To illustrate this item selection and content control process the Focal Problem Elevated BUN and where appropriate the single answer multiple-choice versus multiple answer multiple-choice comparison are used as an example. The item selection and content control process is outlined below: 1) All available items in the Elevated BUN item pool were used. These items had been screened by the faculty for gross errors in content or wording, and were all keyed to one of the topic areas listed in Table 3.3. 2) Every item was then screened by the author for item writing errors not included in the present study and either eliminated

or classified into one of the five item types used in this study. 3) Each item was then listed according to test of origin, item type, keyed topic area and number of options. The item writers' keying of item to topic areas was checked by the author at this point and where disagreement existed the items were eliminated. 4) Every screened MA-MC item was matched as closely as possible to a five-option SA-MC item from the same test administration, in accordance with the keyed topic areas. In the rare cases where the MA-MC item could be matched to any of several SA-MC items the specific item was chosen using first, similarity of content, then if necessary, a random number generator, a total of four item pairs resulted for the SA-MC versus MA-MC comparison, all matched in terms of content designation, number of options and test of origin. Table 3.4 illustrates steps 3, 4 and 5 in this process using the Spring 1976 exam and the SA-MC vs MA-MC comparison. Table 3.5 shows the four item pairs resulting from this example of the selection process.

Specific Selection Procedures - Three Clinical Science Areas

All medical students at Michigan State University are required to participate in clinical clerkship experiences which range from six to twelve weeks in duration. As part of the experience in Pediatrics, Internal Medicine and Surgery the students are provided with a set of basic content objectives and a list of suggested reading materials. At the end of each clerkship the students take an examination (75-125 items) based on the relevant content materials. During the period of this study exams in Internal Medicine, Pediatrics and Surgery were administered three times a year to groups of approximately thirty students. Each item was keyed to a sub-specialty or problem area as listed in Table 3.3. As

TABLE 3.4

ILLUSTRATION OF STEPS 3, 4 AND 5 OF THE ITEM SELECTION PROCEDURE USING THE SA-MC vs. MA-MC COMPARISON AND THE SPRING 1976 ELEVATED BUN EXAM AS THE EXAMPLE

STEP 3: List each item according to test of origin, item type, keyed topic area and number of options.

EXAM OF ORIGIN: Elevated BUN - Spring 1976.

NUMBER OF OPTION: All items have five options.

ITEM TYPE			
SA-MC		MA-MC	
Item Number	Topic Area Number	Item Number	Topic Area Number
2	10	46	9
3	12	47	6
4	9	48	3
5	9	50	16
6	3	51	8
7	9	57	1
8	5	100	9
10	6		
12	8		
14	4		
18	9		
21	6		
23	9		
24	3		
25	10		
27	16		
34	3		
38	6		
39	11		

TABLE 3.4
(continued)

STEP 4: Since there were more SA-MC items, the MA-MC items were listed first. Then each was matched as closely as possible to a SA-MC item.

Topic Area Number	Item Type: MA-MC Item Number	MA-MC Item Content Description	SA-MC Item Number	SA-MC Item Content Description
1	57		none	
3	48	amino acids	6	amino acids
			24	amino acids
6	47		21	
8	51		none	
9	46	antibiotic-absorption	4	antibiotic-aminoglycosides
	100	antibiotic-sensitivity	5	antibiotic-side effects
			7	antibiotic-toxicity
			18	antibiotic-sensitivity
			23	antibiotic-absorption
			38	general-toxicity

STEP 5: Resultant Item Pairs.

Topic Number	MA-MC Item Number	SA-MC Item Number	How Selected
3	48	6	random number
6	47	21	only possible pair
9	46	23	content similarity
	100	18	content similarity

TABLE 3.5

SAMPLE SA-MC vs MA-MC ITEM PAIRS FROM THE ELEVATED BUN SPRING 1976 ILLUSTRATION

(1) Which of the following serum amino acids transports amonia to the liver

1. aspartate
2. glutamine
3. histidine
4. serine
5. tyrosine

Which of the following amino acids can be directly synthesized via intermediates from the glycolytic pathway or pentose phosphate pathway

A. aspartate
B. glutamine
C. phenylalanine
D. serine

(2) Which of the following is the most common causative organism in acute bladder and kidney infections in patients in whom no obstruction exists and neither antimicrobial agents nor instrumentation have been used

1. Enterobacter
2. Escherichia coli
3. Klebsiella
4. Proteus mirabilis
5. Pseudomonas aeruginosa

Which of the following tests would be helpful in differentiating typical E. coli from E. aerogenes

A. indole test
B. methyl red test
C. Voges-Proskauer reaction
D. citrate test

(3) To prolong the absorption time, repository penicillin is administered

1. intramuscularly
2. intravenously
3. intrathecally
4. orally
5. subcutaneously

TABLE 3.5
(continued)

Which of the following are stabile in gastric acid and undergo good absorption after oral administration

A. penicillin G
B. oxacillin
C. methicillin
D. ampicillin

(4) The incidence of hypersensitivity reactions to cephalosporins is higher in patients who have shown allergic manifestations following the administration of

1. gentamicin
2. penicillin
3. polymyxin
4. sulfonamide derivatives
5. tetracycline

A patient with a urinary tract infection and known sensitivity to penicillin could be treated with

A. ampicillin
B. methicillin
C. cephalexin
D. lincomycin

with the Focal Problem Exams the items were initially classified by the item writers, then later checked by a different faculty member or group of faculty during the test review process, as well as being rechecked by the author during the process of item selection. The total selection process was identical to that described in the Elevated BUN example. The result was also the same: Pairs of items from the same test, one of each format, which were written independently and keyed to the same basic sub-content area within that medical speciality.

Content Control

Since the primary purpose of the item selection procedure outlined above was to control content this vital topic has already been discussed at some length. However, a few additional remarks are necessary.

The present study makes use of independently written items from regular classroom tests. The purpose of content control, therefore was to assure that each item within a pair had been designed to test similar content. The important factor in each case was the keying of the items to the topic areas used to guide item writers. (Table 3.3). The procedure used to assure as much accuracy as possible in item keying was outlined earlier. Summarized, there were three steps: 1) Initial keying, 2) Check by other faculty, and 3) Recheck by the author. This double check should be sufficient to assure content similarity. However, one final checking procedure was used. The procedure was followed as outlined below:

1) Since the primary comparison in the study was the SA-MC vs MA-MC comparison, this was the one chosen for use in this final item pairing check.

2) Three topic areas were sought, two focal problem topics and one clinical subspecialty. These were chosen on the basis of having the largest number of item pairs in the SA-MC vs MA-MC comparison. Selected were the focal problem topics, pharmacology and clinical science and the clinical subspecialty, Cardiology.

3) All of these SA-MA item pairs were collected and listed. Appendix B contains the entire listing: Clinical Science, 12 pairs, Pharmacology, 11 pairs and Cardiology, 8 pairs.

4) The completed lists were given to two physicians who were asked to do two things. First, place a check beside all items which were in his opinion correctly keyed to each topic area. Second, to rate each item pair in terms of content similarity: very different, different, similar, or very similar.

The result of this final check showed that both reviewers felt that all of the item pairs were keyed to the correct area. Further, both reviewers rated all sample pairs as being either similar or very similar in content. This result increases the confidence in the appropriateness of the item pairs used throughout the study.

MEASURES OF ITEM QUALITY

The reliability and validity of test results depends on the properties of the individual items which make up the test. The total test has no properties which cannot be derived from those of the single items or the relationships between them. (Magnusson, 1967) The present study concentrates on the quality of the single test items, using difficulty and discrimination indices as measures of indivùal item quality. In his 1939 article, Flanagan outlined the primary and

secondary bases for judging test items, stating that the primary considerations are item difficulty (percentage of persons getting the item correct) and item validity (the extent to which an item will predict the criteria, i.e. predict total test score).

The specific test statistic chosen to estimate item difficulty was simply p, the percentage of examineed who marked the item correctly. Two statistics were chosen to estimate item discrimination: D, the upper-lower index, computed using the upper and lower 27% of the examinees based on total test score; and the biserial correlation coefficient. The rationale for using two discrimination indices is based on their relationship to item difficulty. The biserial correlation coefficient was chosen because it is independent of the difficulty of the item. Pyrczak (1973) published the results of a study intended to measure the validity of the Diserimination Index (Biserial) as a measure of item quality. Biserial correlation coefficients were computed for each of 27 items on Form A and Form B (parallel form) of a nonspeeded arithmetic-reasoning test administered to 364 teacher education students. The resulting discrimination indices were compared to the average of the ratings of three judges to determine validity. The judges based their ratings on nine criteria for item quality, rating each item on both the presence of each fault and their opinion concerning how seriously this fault should affect validity. The validity coefficients for the discrimination indices were .544 and .558 for Forms A and B respectively. Both values were significant at the .01 level. Pyrczak's conclusion was that discrimination indices that are relatively free of the influence of item difficulty are at least moderately valid indicators of item quality. Using p along with the biserial provides two relatively independent measures of item quality.

The Upper-Lower Index provides information slightly different from either of the other two indices. The difficulty level of an item influences of value of D, with the highest values only possible in the intermediate range of difficulties. Since the two indices function differently when item difficulties are low, and since a high percentage of the test items to be used in the study had low difficulty levels, the decision was made to use both indices as estimators of item quality. One final note concerning the use of the biserial coefficient is necessary. If an item has a very low or very high difficulty level, the value for the biserial will occasionally exceed 1.00. Since in reality it is impossible to achieve a correlation of more than 100% the upward limit for the biserial was set at .99 and the lower limit at -.99.

DESIGN

To avoid sources of internal invalidity in any experimental design, it is necessary to attempt to make sure that the two groups of subjects (in this case items) do not differ in any way other than on the experimental treatment (in this case item format). (Campbell & Stanley, 1963) Other possible sources of difference should either be included within the design of the study or controlled for, either during the study or in the statistical analysis. With this in mind the variable of content area (focal problem vs. clinical science) was included in the statistical analysis of each comparison. The reasoning behind the decision to include this factor in the analysis was based on the differing orientation of the two overall groups of test constructors. The focal problem exams were used as criterion referenced exams with an absolute cut-off for passing of 70 percent correct whereas the clinical science exams were used as norm-referenced exams with the passing score

dependent or the performance of the specific group being tested. Since it is not certain what effect this difference in orientation may have on difficulty or discrimination values for items of any particular item format this factor was included in analysis.

Several other factors were also identified as being possible extraneous sources of difference. Therefore, these were controlled to the extent feasible and in the manner described below:

1. Violations of item writing principles not under study: All items containing any violations of item writing principles not under study were excluded from the comparison groups. These violations included:
   a) items that included any complex alternatives or the alternatives "all" or "none of the above";
   b) items lacking grammatical consistency between the stem and all options;
   c) items where the correct answer was significantly longer and contained significantly more qualification than the distractors; and
   d) negative-stemmed or multiple-answer items where the correct answer and its direct opposite were both included in the answer set.
2. Number of options: Since all multiple-answer multiple-choice items (National Board Type K) have five options, these items were compared only to one another or to five-option SA-MC items. For the comparisons involving only single-answer items the average number of options was controlled so that this figure would be very similar for both item types in each comparison made.

3. Item content: The same basic procedure was used to control content for all comparison made in the study. In general, the procedure involved selection of pairs of items from the same exam which shared the same content objective. The procedure used was described in detail in an earlier section.

HYPOTHESES AND ANALYSIS METHODS

IA H_0: There is no difference in the mean difficulty of test items based on item format (single-answer multiple choice (SA-MC) versus multiple-answer multiple-choice (MA-MC)).

H_1: There is a difference in the mean difficulty of test items based on item format (single-answer multiple choice (SA-MC versus multiple-answer multiple-choice (MA-MC)).

IB H_0: There is no difference in the mean discrimination (using D, upper-lower index) of test items based on item format (single-answer multiple choice (SA-MC) versus multiple-answer multiple choice (MA-MC)).

H_1: There is a difference in the mean discrimination (using D, upper-lower index) of test items based on item format (single-answer multiple choice (SA-MC) versus multiple-answer multiple choice (MA-MC)).

IC H_0: There is no difference in the mean discrimination (based on rBis, biserial correlation coefficient) of test items based on item format (single-answer multiple choice (SA-MC) versus multiple-answer multiple choice (MA-MC)).

H_1: There is a difference in the mean discrimination (based on rBis, biserial correlation coefficient) of test items based on item format (single-answer multiple choice (SA-MC versus multiple-

answer multiple choice (MA-MC)).

IIA H_0: There is no difference in the mean difficulty of regular one correct answer multiple choice test items based on completeness of the stem (single-answer multiple-choice items without rule violation (SA-MC) versus single-answer multiple-choice items with uninformative stems (US-MC)).

H_1: There is a difference in the mean difficulty of regular one correct answer multiple choice test items based on completeness of the stem (single-answer multiple-choice items without rule violation (SA-MC) versus single-answer multiple-choice items with uninformative stems (US-MC)).

IIB H_0: There is no difference in the mean discrimination (using D, the upper-lower index) in regular one correct answer multiple choice test items based on completeness of the stem (single-answer multiple-choice items without rule violation (SA-MC) versus single-answer multiple-choice items with uninformative stems (US-MC)).

H_1: There is a difference in the mean discrimination (using D, the upper-lower index) in regular one correct answer multiple choice test items based on completeness of the stem (single-answer multiple-choice items without rule violation (SA-MC) versus single-answer multiple-choice items with uninformative stems (US-MC)).

IIC H_0: There is no difference in the mean discrimination (using rBis, the biserial correlation coefficient) in regular one correct answer multiple choice test items based on completeness of the stem (single-answer multiple-choice items without rule violation

(SA-MC) versus single-answer multiple-choice items with uninformative stems (US-MC)).

H_1: There is a difference in the mean discrimination (using rBis, the biserial correlation coefficient) in regular one correct answer multiple choice test items based on completeness of the stem (single-answer multiple-choice items without rule violation (SA-MC) versus single-answer multiple-choice items with uninformative stems (US-MC)).

IIIA H_0: There is no difference in the mean difficulty of type k multiple choice items based on completeness of the stem (multiple-answer multiple-choice (MA-MC) versus multiple-answer multiple-choice items with uninformative stems (US-MA-MC)).

H_1 There is a difference in the mean difficulty of type k multiple choice items based on completeness of the stem (multiple-answer multiple-choice (MA-MC) versus multiple-answer multiple-choice items with uninformative stems (US-MA-MC)).

IIIB H_0: There is no difference in the mean discrimination (using D, the upper-lower index) of type k multiple choice items based on completeness of the stem (multiple-answer multiple-choice (MA-MC) versus multiple-answer multiple-choice items with uninformative stems (US-MA-MC)).

H_1: There is a difference in the mean discrmination (using D, the upper-lower index) of type k multiple choice items based on completeness of the stem (multiple-answer multiple-choice (MA-MC) versus multiple-answer multiple-choice items with uninformative stems (US-MA-MC)).

IIIC H_0: There is no difference in the mean discrimination (using rBis,

the biserial correlation coefficient) of type k multiple choice items based on completeness of the stem (multiple-answer multiple-choice (MA-MC) versus multiple-answer multiple-choice items with uninformative stems (US-MA-MC)).

H_1: There is a difference in the mean discrimination (using rBis, the biserial correlation coefficient) of type k multiple choice items based on completeness of the stem (multiple-answer multiple-choice (MA-MC) versus multiple-answer multipel-choice items with uninformative stems (US-MA-MC)).

IVA H_0: There is no difference in the mean difficulty of regular one correct answer multiple-choice test items based on stem orientation (single-answer multiple-choice items with positive stems (SA-MC) versus single-answer multiple-choice items with negative stems (NS-MC)).

H_1: There is a difference in the mean difficulty of regular one correct answer multiple-choice test items based on stem orientation (single-answer multiple-choice items with positive stems (SA-MC) versus single-answer multiple-choice items with negative stems (NS-MC)).

IVB H_0: There is no difference in the mean discrimination (using D, the upper-lower index) of regular one-correct answer multiple-choice test items based on stem orientation (single-answer multiple-choice items with positive stems (SA-MC) versus single-answer multiple-choice items with negative stems (NS-MC)).

H_1: There is a difference in the mean discrimination (using D, the upper-lower index) of regular one-correct answer multiple-choice test items based on stem orientation (single-answer

multiple choice items with positive stems (SA-MC) versus single-answer multiple-choice items with negative stems (NS-MC)).

IVC H_0: There is no difference in the mean discrimination (using rBis, the biserial correlation coefficient) of regular one-correct-answer multiple-choice test items based on stem orientation (single-answer multiple-choice items with positive stems (SA-MC) versus single-answer multiple-choice items with negative stems (NS-MC)).

Univariate repeated measures ANOVAS will be used to test the three hypotheses related to each comparison. The level of significance for the F-tests is set at .05. However, since a univariate approach will be used, the Bonferroni approach (Harris, 1975) for correction of the possibility of an inflated alpha will also be used. With this method the desired alpha level is divided by the number of comparisons (i.e. .05 divided by 3) so the cut-off for significance is set at .017. This approach is conservative because it assumes the highest possible inflation of alpha, which would occur only in cases where the dependent variables are completely independent and unrelated. Since a strong positive correlation would be expected between the three measures used, especially between the two discrimination measures the Bonferroni approach should more than control for any inflation of alpha, the type one error, resulting from making multiple comparisons. Although this model is robust for normality, it was tested and the findings indicated that the distributions for the three statistics fell well within the acceptable range. The r to z transformation of the biserial was also tested but was found not to fit the model closely at all. Therefore consideration of its use was dropped.

SUMMARY

This study was designed to test the effect of differing item type on the average item difficulty and discrimination when the groups of items being compared are composed of item pairs matched for content and test of origin.

Five item types were identified within exams in the College of Human Medicine at Michigan State University: Single-answer multiple-choice, national board type k multiple-answer multiple choice (MA-MC), single-answer multiple choice with uninformative stems (US-MC), type k with uninformative stems (US-MA-MC), and single-answer multiple choice with negative stems. Four comparisons between item types were selected as being most relevant: SA-MC vs. MA-MC, SA-MC vs. US-MC, MA-MC vs. US-MA-MC, and SA-MA vs. NS-MC. Groups of item pairs for each comparison were selected independently and on the basis of content similarity. Three item statistics were chosen as estimates of item quality p, proportion getting the item correct; D, the upper-lower discrimination index, and rBis, the biserial item-total correlation coefficient. Within each comparison a repeated measures ANOVA will be used to test the specific hypotheses relating to each of the three measures of item. quality.

Chapter IV presents the results and data analyses performed for this study.

CHAPTER IV

RESULTS

INTRODUCTION

This Chapter presents the results of the statistical analyses performed to test the hypotheses of this study. Results are presented and compared concerning the mean values for difficulty and discrimination for each item type within the four comparisons being studied. Overall results will be presented followed by the results for each individual hypothesis. Finally, additional results relating to the effect of content orientation (focal problem versus clinical science) on item difficulty and discrimination will be reported.

RESULTS BASED ON ITEM TYPE

All of the research hypotheses in this study relate to comparisons based on item type. Table 4.1 displays the means and standard deviations for all three measures of item quality for both groups in each of the four comparisons under study. This table also displays the average number of options for each item type for the two relevant comparisons. To control for the possible effect on item statistics of differing average number of options, an attempt was made to assure that the average number of options was very similar for each group of items in the SA-MC vs. US-MC and SA-MC vs. NS-MC comparisons. As noted on the table these average values were very close. No averages were reported for the other two comparisons because all of the items included in both comparisons had five options.

TABLE 4.1

MEAN VALUES FOR ALL RELEVANT GROUPS
IN COMPARISONS BASED ON ITEM TYPE

COMPARISONS (Mean number of options per item)	ITEM TYPES (Number of items)	MEAN VALUES (Standard Deviations) DIFFICULTY p value	DISCRIMINATION Upper-Lower Index	DISCRIMINATION r Bis
I. Based on item format: SA-MC versus MA-MC	SA-MC	.740	.263	.378
	n=124	(.217)	(.226)	(.279)
	MA-MC	.632	.156	.185
	n=124	(.237)	(.298)	(.329)
II. Based on stem quality in single answer items: Informative versus Uninformative	SA-MC	.709	.262	.361
	n=51	(.246)	(.245)	(.284)
Mean No. of Options(S.D.)	US-MC	.788	.181	.289
SA-MC 4.57(.567)	n=51	(.191)	(.232)	(.314)
US-MC 4.51(.538)				
III. Based on stem quality in multiple ans. items: Informative versus Uninformative	MA-MC	.670	.199	.257
	n=69	(.208)	(.265)	(.327)
	US-MA-MC	.694	.159	.268
	n=69	(.229)	(.264)	(.322)
IV. Based on stem orientation: Positive versus Negative	SA-MC	.729	.256	.371
	n=143	(.203)	(.230)	(.282)
Mean No. of Options(S.D.)	NS-MC	.697	.159	.251
SA-MC 4.78(.414)	n=143	(.235)	(.219)	(.300)
NS-MC 4.83(.381)				

RESULTS OF THE SA-MC VERSUS MA-MC COMPARISON

Five-option single-answer multiple choice items were matched on the basis of content similarity with multiple-answer items from the same test administration. The resulting 124 item pairs were compared for item quality on the basis of difficulty (p) and discrimination (Upper-Lower Index and rBis). Analysis was performed using a two-way repeated measures analysis of variance with the item type as one independent variable and content orientation (focal problem versus clinical science) as the second. Alpha was set at .05. However, since three univariate analyses were performed there was a risk of an inflated type one error. Therefore, the Bonferroni Approach for correction was used. In this approach the alpha level is simply divided by the number of comparisons (.05 / 3) making the cutoff for significance $p \leq .017$.

The results of the three ANOVAs relating to hypotheses IA, IB, and IC are displayed in Table 4.2. Hypothesis IA asks: Are the mean difficulties for the SA-MC and MA-MC item types the same or different? The results indicate that the MA-MC item type (Mean p = .632) was significantly more difficult at the .001 level than the SA-MC type (Mean p = .74) thus resulting in a rejection of the null hypothesis (IAH_0) in favor of the alternative hypothesis (IAH_1).

Hypothesis IB asks: Are the mean discrimination values (based on the Upper-Lower Index) for the SA-MC and MA-MC item types the same or different? The results indicate that the SA-MC item type (Mean D = .263) had a significantly higher mean value for discrimination when measured using the Upper-Lower Index than the MA-MC item type (Mean D = .156). This brings a rejection of the null hypothesis (IBH_0) in

TABLE 4.2

RESULTS FOR THE SA-MC VS. MA-MC COMPARISON
TWO-WAY REPEATED MEASURES ANOVA, n=124

RESULTS FOR DIFFICULTY (p= proportion getting the item correct)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.578	1	.578	10.173*	.002
Error	6.935	122	.057		
SAMC vs. MAMC	.774	1	.774	18.496*	.000
Interaction	.001	1	.001	.020	.887
Error	5.103	122	.042		

RESULTS FOR DISCRIMINATION (D= Upper-Lower Index)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.040	1	.040	.448	.504
Error	10.765	122	.088		
SAMC vs. MAMC	.584	1	.584	11.090*	.001
Interaction	.129	1	.129	2.460	.119
Error	6.412	122	.053		

RESULTS FOR DISCRIMINATION (rBis= biserial correlation coefficient)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	1.081	1	.081	.636	.427
Error	15.567	122	.128		
SAMC vs. MAMC	2.134	1	2.134	35.250*	.000
Interaction	.014	1	.014	.226	.635
Error	7.385	122	.061		

* Significant at the .05 level, using the Bonferroni Approach to set the cut-off for p of F at .017.

favor of the alternative hypothesis (IBH_1).

Hypothesis IC asks: Are the mean discrimination values based on rBis the same or different for the SA-MC and MA-MC item types? The results indicate that the SA-MC item type (Mean rBis = .378) had a significantly higher mean value for discrimination when measured using the biserial correlation coefficient than the MA-MC type (Mean rBis = .185). This results in a rejection of the null hypothesis (ICH_0) in favor of the alternative hypothesis (ICH_1).

Reporting of all results relating to comparisons based on content orientation or any possible interactions between content orientation and any item type will be defered until a later section of this Chapter.

RESULTS OF THE SA-MC VERSUS US-MC COMPARISON

Single-answer multiple choice items without rule violations were matched on the basis of content similarity with single-answer items with uninformative stems. The average number of options per item was controlled as closely as possible resulting in a mean of 4.57 options per SA-MC item and 4.51 options per US-MC item. Fifty-one item pairs resulted. These were compared in the same manner as the SA-MC vs. MA-MC pairs.

The results of the three ANOVAs relating to Hypotheses IIA, IIB, and IIC are displayed in Table 4.3. Hypothesis IIA asks the question: Are the mean difficulties for the SA-MC and US-MC item types the same or different? The results indicate that the SA-MC item group (Mean p = .709) was significantly more difficult than the US-MA group (Mean p = .788) with an alpha of .016. This results in a rejection of the null hypothesis of no difference ($IIAH_0$) in favor of the alternative

TABLE 4.3

RESULTS FOR THE SA-MC VS. US-MC COMPARISON
TWO-WAY REPEATED MEASURES ANOVA, n=51

RESULTS FOR DIFFICULTY (p=proportion getting the item correct)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.123	1	.123	2.489	.121
Error	2.415	49	.049		
SAMC vs. USMC	.268	1	.268	6.228*	.016
Interaction	.205	1	.205	4.776	.034
Error	2.107	49	.043		

RESULTS FOR DISCRIMINATION (D= Upper-Lower Index)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.0001	1	.0001	.002	.965
Error	2.750	49	.056		
SAMC vs. USMC	.215	1	.215	3.665	.061
Interaction	.058	1	.058	.995	.324
Error	2.875	49	.059		

RESULTS FOR DISCRIMINATION (rBis= biserial correlation coefficient)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.003	1	.003	.033	.856
Error	4.500	49	.092		
SAMC vs. USMC	.224	1	.224	2.560	.116
Interaction	.176	1	.176	2.009	.163
Error	4.291	49	.088		

* Significant ar the .05 level, using the Bonferroni Approach to set the cut-off for p of F at .017.

hypothesis ($IIAH_1$).

Hypothesis IIB asks the question: Are the mean values for discrimination (based on the Upper-Lower Index) for the SA-MC and US-MC item types the same or different? The results showed that the mean value for the SA-MC item type (Mean D = .262) was higher than the mean value for the US-MC item type (Mean D = .181) but that the difference was not significant at the .05 level. Therefore, the null hypotheses ($IIBH_0$) could not be rejected.

Hypothesis IIC asks the question: Are the mean values for discrimination (based on the biserial correlation coefficient) for the SA-MC and US-MC item types the same or different? The results showed that although the mean discrimination value for the SA-MC type (Mean rBis = .361) was higher than that for the US-MC type (Mean rBis = .289) that this difference was not significant at the .05 level. Therefore, the null hypothesis ($IICH_0$) was not rejected.

RESULTS OF THE MA-MC VERSUS US-MA-MC COMPARISON

Multiple-answer multiple choice items without rule violations were matched on the basis of content similarity to multiple-answer items with uninformative stems. The sixty-nine item pairs which resulted were compared in the same manner as the SA-MC vs. MA-MC pairs.

The results of the three ANOVAs relating to Hypotheses IIIA, IIIB, and IIIC are displayed on Table 4.4. Hypothesis IIIA concerned the relationship between the two item types based on p, the percentage getting each item correct (difficulty) asking whether the mean difficulty levels are the same or different? The results showed the two means to be

TABLE 4.4

RESULTS* FOR THE MA-MC VS. US-MA-MCCOMPARISON
TWO-WAY REPEATED MEASURES ANOVA, n=69

RESULTS FOR DIFFICULTY (p= proportion getting the item correct)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.216	1	.216	3.514	.065
Error	4.127	67	.062		
MAMC vs US-MAMC	.014	1	.014	.445	.507
Interaction	.000	1	.000	.0002	.989
Error	2.173	67	.032		

RESULTS FOR DISCRIMINATION (D= Upper-Lower Index)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.029	1	.029	.470	.495
Error	4.133	67	.062		
MAMC vs US-MAMC	.047	1	.047	.590	.445
Interaction	.001	1	.001	.018	.893
Error	5.345	67	.080		

RESULTS FOR DISCRIMINATION (rBis= biserial correlation coefficient)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.105	1	.105	1.052	.309
Error	6.663	67	.099		
MAMC vs US-MAMC	.005	1	.005	.043	.836
Interaction	.001	1	.001	.010	.921
Error	7.575	67	.113		

* None of the results were significant at the .05 level, using a cut-off of .017 for p of F determined using the Bonferroni Approach.

very similar with the mean difficulty equal to .670 for the MA-MC type and .694 for the US-MA-MC type. Since there was no significant difference at the .05 level, the null hypothesis ($IIIAH_0$) was not rejected.

Hypothesis IIIB concerns the relationship between the two item types based on D, the Upper-Lower Index for item discrimination, asking whether the mean values for discrimination are the same or different? The results showed the two means to be quite similar with an average D of .199 for the MA-MC type and .159 for the US-MA-MC type. Since there was no significant difference at the .05 level, the null hypothesis ($IIIBH_0$) was not rejected.

Hypothesis IIIC concerns the relationship between the two item types based on rBis, the biserial item-total correlation coefficient, asking whether the mean values for discrimination are the same or different? The results showed the two means to be almost identical with an average rBis of .257 for the MA-MC item type and .268 for the US-MA-MC type. Since there was no significant difference at the .05 level, the null hypothesis ($IIICH_0$) was not rejected.

RESULTS FOR THE SA-MC VERSUS NS-MC COMPARISON

Single- answer multiple choice items with positive stem orientation were matched on the basis of content similarity with single-answer items with negatively stated stems. The average number of options per item was controlled as closely as possible resulting in a mean of 4.78 options per SA-MC item and 4.83 options per NS-MC item. One-hundred and forty-three item pairs resulted. These were compared in the same manner as the SA-MC vs. MA-MC pairs.

The results of the three ANOVAs relating to hypotheses IVA, IVB, and IVC are displayed in Table 4.5. Hypothesis IVA asks the question: Are the mean difficulties for the SA-MC and NS-MC item types the same or different? The results indicate that the NS-MC item type (Mean p = .697) was slightly more difficult than the SA-MC item type (Mean p = .729). However since this difference was not significant at the .05 level the null hypothesis ($IVAH_0$) was not rejected.

Hypothesis IVB asks the question: Are the mean discrimination values (based on the Upper-Lower Index) for the SA-MC and NS-MC item types the same or different? The results indicate that the SA-MC item type (Mean D = .256) had a significantly higher mean value for discrimination when measured using the Upper-Lower Index than the NS-MC item type (Mean D = .159). This results in a rejection of the null hypothesis ($IVBH_0$) in favor of the alternative hypothesis ($IVBH_1$).

Hypothesis IVC asks the question: Are the mean discrimination values based on rBis the same or different for the SA-MC and NS-MC item types? The results indicate that the SA-MC item type (Mean rBis = .371) had a significantly higher average value for discrimination when measured using the biserial correlation coefficient than the NS-MC type (Mean rBis = .251). This results in a rejection of the null hypothesis ($IVCH_0$) in favor of the alternative hypothesis ($IVCH_1$).

RESULTS BASED ON CONTENT ORIENTATION

All of the research hypotheses in this study relate to comparisons based on differences in item type. As stated earlier the analysis for each hypothesis was performed using a two-way repeated measures ANOVA with

TABLE 4.5

RESULTS FOR THE SA-MC VS. NS-MC COMPARISON
TWO-WAY REPEATED MEASURES ANOVA, n =143

RESULTS FOR DIFFICULTY (p= proportion getting the item correct)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	1.366	1	1.366	26.196*	.000
Error	7.353	141	.052		
SAMC vs. NSMC	.076	1	.076	2.161	.144
Interaction	.038	1	.038	1.077	.301
Error	4.947	141	.035		

RESULTS FOR DISCRIMINATION (D= Upper-Lower Index)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.142	1	.142	2.558	.112
Error	7.830	141	.056		
SAMC vs. NSMC	.671	1	.671	15.005*	.000
Interaction	.053	1	.053	1.179	.279
Error	6.305	141	.045		

RESULTS FOR DISCRIMINATION (rBis= biserial correlation coefficient)

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p of F
Content	.899	1	.899	9.361*	.003
Error	13.537	141	.096		
SAMC vs. NSMC	1.028	1	1.028	15.188*	.000
Interaction	.038	1	.038	.566	.453
Error	9.543	141	.068		

* Significant at the .05 level, using the Bonferroni Approach to set the cut-off for p of F at .017.

one of the independent variables being content orientation. This variable was included in the analysis primarily to determine if content orientation interacts with item type, in any of the relevant comparisons. This type of interaction might occur if performance of any specific item type varied systematically, either with the type of content or with increased student experience with that item type. (Focal Problems are studied by year 1 and 2 students while Clinical Sciences are studied by year 3 and 4 students.) In either case an interaction effect would be demonstrated by a statistically significant F test for the interaction term. Three analyses were performed for each of the major research hypotheses (I - IV). There were no significant interactions found (at the .05 level) in any of the analyses. Therefore, it is assumed that the two variables, item type and content orientation, act independently in relation to item difficulty and discrimination.

Table 4.6 displays the means and standard deviations for all three measures of item quality, for both content areas in each of the four comparisons under study. Results related to average item difficulties showed that in all cases the Clinical Science items had a higher average level of difficulty, with this difference significant at the .05 level for both the SA-MC vs. MA-MC and SA-MC vs. NS-MC comparisons.

Results related to average item discrimination showed that the mean values were generally quite similar for the two content orientations, especially those for the Upper-Lower Index. However, one significant difference was found within comparison IV (SA-MC and NS-MC items). The average value for rBis was higher for Focal Problem items (Mean rBis = .368) than for Clinical Science items (Mean rBis = .256).

TABLE 4.6

MEAN VALUES FOR ALL RELEVANT GROUPS
IN COMPARISONS BASED ON CONTENT AREA

COMPARISONS	CONTENT AREAS (Number of items)	MEAN VALUES (Standard Deviations) DIFFICULTY p value	DISCRIMINATION Upper-Lower Index	DISCRIMINATION r Bis
I. SA-MC versus MA-MC	Focal Problem n=130	.732[a] (.232)	.198 (.240)	.299 (.314)
	Clinical Science n=118	.635[a] (.224)	.223 (.300)	.262 (.326)
II. SA-MC versus US-MC	Focal Problem n=66	.774 (.202)	.222 (.236)	.329 (.300)
	Clinical Science n=36	.702 (.254)	.220 (.252)	.318 (.305)
III. MA-MC versus US-MA-MC	Focal Problem n=108	.703 (.224)	.171 (.254)	.277 (.340)
	Clinical Science n=30	.607 (.181)	.206 (.301)	.210 (.255)
IV. SA-MC versus NS-MC	Focal Problem n=142	.783[b] (.197)	.230 (.223)	.368[c] (.314)
	Clinical Science n=144	.644[b] (.220)	.185 (.234)	.256[c] (.268)

a,b,c The differences between these mean values were significant at the .05 level using the Bonferroni Approach to adjust the cutoff for p of F to .017.

SUMMARY

Results of all analyses performed based on item type and used to test the hypotheses of the study, have been presented in this chapter and are summarized below:

> Hypothesis I - In the comparison of SA-MC and MA-MC items, the MA-MC items were found to be significantly more difficult and less discriminating than the SA-MC items.
>
> Hypothesis II - In the comparison of SA-MC and US-MC items, the US-MC items were found to be significantly less difficult than the SA-MC items. There were no significant differences in mean item discrimination.
>
> Hypothesis III - There were no significant differences in either difficulty or discrimination found when MA-MC and US-MA-MC items were compared.
>
> Hypothesis IV - In the comparison of SA-MC and NS-MC items, the NS-MC items were found to be significantly less discriminating than the SA-MC items. There was no significant difference in difficulty between the two item types.

Results based on content orientation were also reported in this chapter and indicated that items on the clinical science exams tended to be more difficult than items on the focal problem exams. This difference reached the .05 level of significance within the SA-MC vs. MA-MC and SA-MC vs. NS-MC comparisons. There did not appear to be any consistent difference in the mean discrimination values between the two content orientations, although there was one significant difference based on rBis within the SA-MC vs. NS-MC comparison. This difference showed the focal problem items to be more discriminating.

In Chapter V these findings will discussed, conclusions will be drawn, and suggestions for future research will be made.

CHAPTER V

DISCUSSION AND CONCLUSIONS

INTRODUCTION

In this chapter the results reported in Chapter IV will be discussed and related where relevant, to past research. Each comparison will be discussed seperately followed by a general discussion and summary. Conclusions will then be drawn and suggestions for future research made.

DISCUSSION

SA-MC versus MA-MC Comparison

The purpose of Hypotheses IA-IC was to help answer the research question: How do the two item types (SA-MC and MA-MC), as typically written and used within medical education, compare on the basis of average item difficulties and discrimination. One reason for asking this question was a concern that the MA-MC (national board type k) item type might be more complex without being any more effective than the more straight forward single-answer multiple choice type. The results of testing Hypotheses IA-IC provide evidence to support this concern. The MA-MC type was found to be significantly more difficult (based on p) while having a significantly lower average value for discrimination (based on both the upper-lower index and rBis) as displayed in Table 4.2.

Past research comparing a variety of multiple-answer formats to one another and to the single-answer format showed few significant differences. The study having the most relevance to the present study

(Skakun et. al., 1977) showed no significant differences between the SA-MC and MA-MC formats based on 18 matched pairs of Surgery items written by the authors for a certification examination. Two differences between this study and the present study may account for the differing results. First, the items in the Skakun study were written by professional item writers while those in this study were written by regular classroom teachers. Second, 18 items is a very small sample providing little power to detect differences.

Careful examination of the results of the present study and of all items used in the SA-MC vs. MA-MC comparison disclosed several hypotheses which might explain the results. The first is the straight-forward hypothesis that the results are due to an inherent difference in complexity between the two formats. This difference in complexity is easily demonstrated since the SA-MC format always calls for one and only one correct alternative, while the MA-MC format can have anywhere from one to four correct alternatives arranged to form five different combinations. (Recall that the choices are: Option 1 if A,B and C are correct, 2 if A and C are correct, 3 if B and D are correct, 4 if only D is correct, and 5 if A,B,C and D are correct.) It is reasonable to hypothesize that this inherent complexity contributed to the increased difficulty of the format and may have affected discrimination.

The second hypothesis relates to the content of the keyed correct responses of the two formats. Well-written SA-MC items generally have the best possible answer as the keyed correct response, the students need only to identify this response. In comparison MA-MC items sometimes include in the correct response the best possible answer plus

one or more other responses which are correct less frequently or only under certain circumstances, so students are forced to make finer distinctions between what is and what is not correct. An example may help to clarify. Example:

> The clinical symptoms of hepatitis are in some ways similar to and <u>often</u> confused with :
>
> *A. infectious mononucleosis
> *B. herpes simplex infections
> *C. toxoplasmosis
> *D. idiosyncratic drug reactions
>
> (p = .33, D = .20, rBis = .30)

In this item only 33% of the students chose the correct response (5, all are correct) while 55% of the students chose option 4 (D only) and all students included D as part of their chosen response. Clearly the students were most familiar with the similarity of symptoms between hepatitis and idiosyncratic drug reactions while most did not judge the connection between the symptoms of hepatitis and the other problems significant enough to mark them as correct responses.

Items calling for this type of decision making can be either appropriate or inappropriate. This would depend primarily on whether the total item, including all options, was at the correct level for the students being tested. Logically, items of increased difficulty but with appropriate content should show higher item discriminations, while those at an inappropriate level should discriminate poorly. The hypothesis is that many items contain options at an inappropriate level so that both difficulty and discrimination are affected.

The inappropriate level of some items, is due at least partially to the fact that teachers sometimes include alternatives that contain informa-

tion unfamiliar to most or all students being tested. This can happen intentionally, based on the questionable notion that including some very difficult or unfamiliar material "challenges" students. It can also happen unintentionally. An instructor needing just one more option may "settle for" something less than optimal. Further, many instructors, especially physicians, tend to write items using their own experiences instead of refering directly to the materials available to students. This practice of including unfamiliar material in test items is questionable at best, but does not pose much problem to the student whose task is to choose the one best response. He will probably simply ignore the unfamiliar option. However, when the task is to choose the appropriate combination of correct alternatives an option with unfamiliar material poses a greater problem. If the student has no rational basis for judging the correctness of the option he is forced to guess. This type of guessing interferes with the item's ability to measure student learning.

Further, it is contended that this problem is exacerbated by use of confusing wording in some stems. An example will help illustrate this:

> In a patient with major motor seizures, status epilepticus <u>may be</u> precipitated by:
>
> *A. abrupt withdrawl of anticonvulsant drugs
> *B. brain tumor
> *C. brain injury
> D. hypoglycemia
>
> (p = .51, D = .00, rBis = -.13)

An examination of the item analysis for this item showed that while 51% of the students chose the correct option (1, A,B and C correct), the other 49% chose option 5 (all correct). This indicates either that

many students were not familiar with the relationship (or lack of) between hypoglycemia and onset of status epilepticus or that they interpreted the phrase "may be precipitated" differently than the item writer intended.

Careful reading of the stems shows that both this item and the earlier example contain phrases which could cause confusion. The underlined phrase in each item requires a value judgement; how often is often enough and is the opposite of may be, never or seldom? Use of nebulous wording like this within a multiple correct answer format can detract from the accurate measurement of student knowledge. My hypothesis is, that use of this type of terminology combined with the problem of varying degrees of correctness among options makes the students' task more difficut, resulting in increased item difficulty and in some cases negatively affecting item discrimination.

To summarize, the results of testing Hypotheses IA-IC indicated that the MA-MC items are more difficult and less discriminating. Two hypotheses were offered as possible explanations: 1) The format itself is more complex which results in items of higher difficulty but lower discrimination; and 2) The format is used poorly by classroom teachers resulting in items which are overly difficult and less discriminating due to (a) confusing wording in the stem and/or (b) inclusion of options with unfamiliar information or presenting distinctions too fine for students to discern. A final possibility is that the difference in difficulty and discrimination between the SA-MC and MA-MC formats is due to a combination of all of the factors listed above.

SA-MC versus US-MC Comparison

Hypotheses IIA - IIC addressed the question: What effect does completeness (informativeness) of the item stem of multiple-choice items have on item difficulty and discrimination? In these hypotheses single-answer items with complete, informative stems were compared with those with uninformative stems. The results of the analysis, displayed in Table 4.3, indicated that while the US-MC items were less difficult than the SA-MC items, there was no significant difference in the average value of either discrimination index.

How does this result compare to the results of past research? Of the studies discussed earlier in the review of the literature only one shares enough similarity with the present study to merit discussion here. In this study (Schmeiser and Whitney, 1975) a teacher-made sociology exam was identified in which 22 out of 61 items had incomplete stems. The items containing this error were then rewritten by the authors to make the stems appropriately complete. Two groups of items resulted which were compared on the basis of item statistics. The results of this analysis (detailed earlier) indicated that the items with less complete stems were more difficult but equally discriminating.

On the surface this result seems to contradict the results of the present study. However, the two studies are not really addressing the same research question. In the Schmeiser and Whitney study the comparison items were identical except for the changes, made by the authors, in the wording of the stems. In the present study, while the items were intended to test the same content or objective, the comparison items were written independently and were not identical. Comparison of

the conclusions which could be drawn on the basis of the results of the respective studies will make this difference clear. An appropriate conclusion which could be drawn from the Schmeiser and Whitney research would be: When measurement specialists rewrite items that have uninformative stems, so that the stems are as informative as possible, the resulting items are easier than the original items but no more discriminating. This was a logical result; when an item which poses no identifiable question in the stem is rewritten so that the question asked is clearer, the rewritten item ought to be less difficult. Since the present study has a different focus the results should not be expected to be identical. A conclusion based on the results of the present study would be: When items with uninformative stems are matched with independently written informative-stemmed items that were intended to test the same content or objective, the items with uninformative stems are easier but not significantly less discriminating.

What possible explanations are there for the present result? Careful examination of the items included in the SA-MC vs. US-MC comparison reinforced the earlier statement that many US-MC items lack a clear focus, i.e. many of these items do not appear to be based on a definite single item idea. Table 5.1 shows two sample SA-MC vs. US-MC item pairs. In the first pair both items were intended to test student knowledge about seizures in childhood. It can be seen that the US-MC item is very general in focus (grand mal seizures) and that the options are quite heterogeneous, while the focus of the SA-MC item is specific (pharmacologic treatment of petit mal seizures under specified circumstances) and the options very homogeneous. A similar comparison can be made between the two items in the second pair. Both items were

TABLE 5.1

SAMPLE SA-MC VS. US-MC ITEM PAIRS

1. US-MC Item

A grand mal seizure (major motor seizure)

1. seldom causes unconsciousness
2. never has an aura preceding it
*3. may follow a temporal lobe seizure
4. is usually construed to mean movement of the arm and leg on one side

SA-MC Item

When a child with petit mal seizures shows no response to phenobarbital, the drug of choice is now

*1. ethosuximide (Zarontin^{R})
2. trimethadione (Tridione^{R})
3. acetazolamide (Diamox^{R})
4. paramethadione (Paradione^{R})
5. phensuximide (Milontin^{R})

2. US-MC Item

The space of Disse in the liver

1. contains both the plasma and cillular elements of blood
2. receives bilirubin glucuronide on its way from the hepatocyte to the bile ducts
3. connects directly with the central veins
*4. receives proteins formed in the hepatocytes on their way to the intravascular compartment

SA-MC Item

The liver arises as a diverticulum of

1. esophagus
2. midgut
*3. foregut
4. hindgut

testing student knowledge of the anatomy of the liver. However, while the SA-MC item asked a specific question with only one item idea behind it, the US-MC item was less specific with the options being a collection of facts (or falsehoods) related in some way to a particular part of the liver.

These examples are shown here to raise the possibility that since many of the US-MC items have stems which are quite general and options which are quite heterogeneous, average item difficulty may have been affected. One of the precepts of item writing (Wesman, 1971; Ebel, 1979) states that items with heterogeneous options (options with meanings that are widely divergent) tend to be less difficult than items with homogeneous options (options with meanings that are very similar.) Since a high percentage of the US-MC items used heterogeneous options, while most SA-MC items had more homogeneous options, this is suggested as a possible explanation for the results showing US-MC items to be significantly less difficult than the SA-MC items.

A second possible explanation for these results relates to the statement made in Chapter II, that very often the type of item classified as having an uninformative stem, was found on careful examination by medical faculty, to have been testing several concepts, none of which were important concepts. In other words many of these items are basically trivial in nature. This is offered here as a possible explanation why this item type had a lower average difficulty than the SA-MC group. However, this logic would also lead to the expectation that the US-MC items would also be less discriminating. While it was the case that the average value for both D and rBis was lower for the US-MC type,

neither difference was significant. It is possible that future research using more refined techniques for identification of US-MC items and for pairing these with appropriate SA-MC items might be able to detect significant differences.

MA-MC versus US-MA-MC Comparison

Hypotheses IIIA - IIIC addressed the same basic question asked in Hypothesis II: What effect does completeness (informativeness) of the item stem have on item difficulty and discrimination? However, the focus of Hypothesis III was on the MA format, comparing multiple-answer items with complete, informative stems to those judged to have uninformative stems. The results, shown in Table 4.4, indicated that there were no significant differences between the two item types, based on either difficulty or discrimination.

There are two logical alternative explanations for this lack of significant results. First is the possibility that the completeness of the stem has in fact, no effect on either the difficulty or discrimination of multiple-answer items. An alternative explanation relates to the general problems of the MA format outlined earlier in the discussion of the SA-MC vs. MA-MC comparison. In this discussion several previously unlooked for item writing errors were proposed as problems in the multiple-answer format. Since these problems, i.e. specific types of confusing wording in the item stem and use of inappropriate alternatives, were not considered in the item selection process for the MA-MC vs. US-MA-MC comparison, it is doubtful that items classified as "well-written" MA-MC items were actually error free. Therefore, the results of the present comparison could have been contaminated by items of both

types, MA-MC and US-MA-MC, containing errors other than the one under study. Future research will be better able to take these errors into consideration when studying differences between variations of the multiple-answer format. Under those more controlled conditions the most common errors within the multiple-answer format can be identified and studied.

SA-MC versus NS-MC Comparison

Hypotheses IVA - IVC were directed toward the comparison of items based on stem orientation, positive versus negative. The results, shown in Table 4.5, indicated that while there was no significant difference in difficulty between the SA-MC and NS-MC item types, the SA-MC type had a significantly higher average value for discrimination, based on both the upper-lower and biserial indexes.

This result differs somewhat from those of the earlier studies described in the review of the literature. Recall that two studies were described, (Terannova, 1969; Dudycha and Carpenter, 1973) both of which used professionally recast items to test for differences. Each study found items with negative stems to be more difficult than those with positive stems, but found no significant differences in discrimination. In the present study, although the difference in difficulty was not statistically significant the negative-stemmed items were found to be somewhat more difficult. The more notable difference was that the present study showed a significant difference in average discrimination while the earlier studies did not.

As discussed in the section relating to the SA-MC vs. US-MC comparison, the difference in research approach between use of recast items in past studies and use of independently written items in this study goes far to explain differing results. Further, there were other differences between these past studies and the present study. The Terannova study looked only at two-option m-c items while the average number of options per item in the present study was 4.8 which could contribute to differences. In the Dudycha and Carpenter study the initial items used in the recasting process were chosen on the basis of middle difficulty and high values for discrimination, the same bases later used to compare item groups. Use of the same basis for both selection and comparison casts some doubt on the validity of this result and adds to the differences between their study and this one.

The present study provides evidence that the NS-MC item type is less effective than the positive stem (SA-MC) item type. What factors might contribute to this decreased effectiveness? The first factor which might logically have an effect on item discrimination is the negative orientation itself and the resultant fact that the keyed correct response is actually a wrong answer. Not only could this present problems for examinees but it is proposed here that it presents even more difficulties for instructor item writers and can affect item quality because of the relationship between the keyed correct response and item quality.

With the positive orientation where the item calls for the best correct response, the quality of the item depends on: 1) the item idea and its expression in the stem, 2) the correct response, and 3) the appropriateness of the incorrect or less correct alternatives. However, with

the negative orientation where the item calls for the incorrect or least correct response, the order of importance changes to become: 1) the item idea and its expression in the stem, 2) the wrong response, and 3) the appropriateness of the correct alternatives. It is possible that it is easier to write high quality items (based on item discrimination) when the quality is based more on the item writer's knowledge of the most correct answer than when the quality depends more on the writer's choosing the best wrong answer. In other words, typical instructors in medical education may be less able to write appropriate NS-MC items which results in lower average discrimination values.

A second possible factor contributing to this difference in discrimination is the type of questions asked within the two item types. On careful examination of the NS-MC items used in this comparison, the negative orientation was found to be a rather stereotypic approach. Approximately 50% of the NS-MC items had stems which were very general and often poorly focused, i.e. the specific item idea is unclear. In this approach, for simplicity refered to as the PFS (poorly focused stem) approach, the stems took one of the following three general forms:

1) All of the following statements concerning ..... are true Except:
2) Each (All) of the following may be (are) associated with ..... Except:
3) All of the following are characteristics of ..... Except:

Table 5.2 provides a specific sample item of each form. The PFS approach appears to be unique to the NS-MC item type, and its prevalence results in a very large number of items which are rather superficial since they ask only for recognition of facts. Looking again at the examples it can also be seen that instead of asking a direct question with a specific

TABLE 5.2

SAMPLE NS-MC ITEMS WITH POORLY FOCUSED STEMS (PFS ITEMS)

1) All of the following statements concerning echoencephalograms are true Except:

   1. it is an invasive procedure with many associated potential hazards
   2. midline displacement of 3mm is considered abnormal
   3. it is difficult to distinguish epidural from subdural hematomas with this test
   4. it is most beneficial when used in conjunction with other tests
   5. it is not useful in posterior fossa tumor detection

2) All of the following are associated with pityriasis rosea Except:

   1) acute, self-limited disease
   2) allergic origin
   3) herald (patch) lesion
   4) rare before age one year
   5) corticosteriods are effective when severe pruritus is present

3) All of the following are characteristics of penicillin G Except:

   1. rapid renal excretion
   2. instability in gastric acid
   3. rather poor penetration into CSF
   4. high incidence of adverse and toxic reactions
   5. narrow antimicrobial spectrum

correct answer, this approach uses a general stem which provides the student with less direction and requires the student to choose the incorrect response from among a list of statements. It is not known whether this approach results in items with difficulties or discrimination values that differ from other NS-MC items, but because of its prevalence and uniqueness within this item type it is offered as a possible factor.

To further explore the SA-MC vs. NS-MC comparison and the possible effect of the PFS approach, a grouped frequency distribution was prepared using values for both D and rBis. This distribution is displayed in Table 5.3. Using this distribution the 26 best items (based on D) and the 53 worst items (based on rBis) were selected and scrutinized. Of the 26 best items, 16 SA-MC and 10 NS-MC, the only commonality found was that all except one of the items had well-focused stems asking a direct question. Only one of the ten NS-MC items in this group used the PFS approach criticized above. For 50 of the 53 worst items the most probable problem could be identified. These are listed below:

Problems	Number of SA-MC items	Number of NS-MC items
Poorly focused stem	0	14
No one missed the item	7	9
Student confusion between 2 options	5	3
Item content appeared too difficult or inappropriate	6	3
Stem wording was confusing	0	3
Problem could not be identified	2	1
	20	33

This exploratory analysis of the items used, supports the idea that heavy use of the PFS approach may have a negative effect on discrimination.

TABLE 5.3

SA-MC VS. NS-MC COMPARISON, GROUPED FREQUENCY DISTRIBUTION BASED ON D AND rBIS

	Discrimination Index			
	D		rBis	
Range of Values	Number of SA-MC Items	Number of NS-MC Items	Number of SA-MC Items	Number of NS-MC Items
-1.00 - .00	28	47	20	33
.01 - .10	5	10	8	10
.11 - .20	29	24	15	18
.21 - .30	28	33	19	24
.31 - .40	21	8	14	13
.41 - .50	16	11	16	16
.51 - 1.00	16	10	51	29
	143	143	143	143

Only one of the 10 best negatively oriented items had stems of this type while 14 of the 33 worst NS-MC items had stems of this type.

In summary, the results of the SA-MC vs. NS-MC comparison indicated that the SA-MC items are more discriminating than the NS-MC items but not significantly less difficult. Two factors were suggested which might contribute to the lower average discrimination of the NS-MC type: 1) the task of choosing an effective wrong answer to be the keyed correct response may be more difficult than choosing an effective correct answer, and 2) many NS-MC items use poorly focused stems (PFS).

Content Orientation

All of the comparisons in this study were analyzed using two-way repeated measures analysis of variance with item type an one independent variable and content orientation as the other. The results of these analyses, in Tables 4.2 - 4.5, showed no significant interactions between content orientation and any item type, for either difficulty or discrimination.

The results of these analyses also indicated that on the average the clinical science items were more difficult than the focal problem items, with this difference being statistically significant within two of the four comparisons, SA-MC vs. MA-MC and SA-MC vs. NS-MC. This result is consistent with the differing grading method used, at the time of this study, within the focal problem curriculum versus that used in the three clinical clerkships. Within both content orientations a pass-fail grading system was used. However, different methods were used to determine the cutoff score for passing. The focal problem exams were criterion-referenced, using an absolute cutoff score for passing, which varied in different focal problems from 68 -72 percent correct. The clinical

science exams were norm-referenced, the passing score for each exam was dependent on the performance of the group tested. The cutoff point was generally set at 1.5 - 2.0 standard deviations below the mean score.

Since the focal problem exams used cutoff scores of approximately 70% correct, the average difficulty value would be expected to be above .70. The actual range for the mean difficulties in the four comparisons in this study was .70 - .78. Since the clinical science exams were norm-referenced, the average difficulty would be expected to be closer to the ideal level of difficulty for trying to maximize discrimination. When four or five-option items are used this level would be between .625 and .667. The actual range for the mean difficulties in the four comparisons was .61 - .70.

The results relating to item discrimination indicated that in seven out of eight comparisons there was no significant difference between focal problem and clinical science items. The one significant difference was within the SA-MC vs. NS-MC comparison and showed that the focal problem items had a significantly higher value for discrimination, based on the biserial coefficient. This result does not seem to have any practical significance since it was an isolated result and not even consistent with the result for the other discrimination index within the same comparison.

General Discussion and Summary

The majority of the discussion thus far has concentrated on the negative aspects of the MA-MC, US-MC, US-MA-MC and NS-MC item types. At this point, a brief statement needs to be made about the positive aspects of the SA-MC item type. The major advantage that the SA-MC item type

appears to have is that it uses the most direct approach to asking questions. In this type of item the instructor (item writer) is asking only one question for which he is requesting only one best answer. The stem and alternatives are separate parts of the item; before seeing the options the student knows exactly what is required and could (if he has sufficient knowledge) answer the item without ever seeing the options. None of the other item types studied have these characteristics. The multiple-answer types have varying numbers of correct alternatives and cannot be answered by reading the stem alone, even if the stem asks a specific question, because the student does not know how many answers he needs. Items of the NS-MC type also have more than one "right" answer. The students cannot directly answer the question posed in the stem because the keyed correct response is always a "wrong" answer. A further asset of the single-answer format is that when instructors use this format they more frequently use complete well-focused stems. Recall from Table 1.1 that of all single-answer items only about 10% had poorly focused stems (US-MC items), while 25% of the multiple-answer items had this problem (US-MA-MC items). Also, as discussed earlier, approximately 50% of the negatively oriented items used general stems which lacked clear focus (PFS items). Clearly, the SA-MC item type has strengths which help to account for the very high average values for discrimination associated with this item type throughout the study. Part of the difference, at least for discrimination, must be attributed to these strengths as well as being attributed to the very real weaknesses of the other item types.

To summarize the discussion section, several key points are outlined below:

1) SA-MC items were significantly less difficult and more discriminating than MA-MC items. The MA-MC format is inherently more complicated which may have contributed to differences. Also the possibility exists that instructors are less able to use the format correctly and effectively. This results in items which have confusing wording in the stem and/or include inappropriate options, which can have a negative effect on item statistics.

2) SA-MC items were significantly more difficult than US-MC items. The two explanations offered were: a) that many items using this item form are testing trivial content so are not as difficult, and b) that many items of this type have very heterogeneous alternatives, which may result in lower average difficulties.

3) SA-MC items were significantly more discriminating than the NS-MC items. Many of the NS-MC items used very general item stems which seemed to result in a higher number of very poor items (based on D and rBis) and a lower number of very good items. This probably contributed to the lower average discrimination values for the NS-MC item type.

4) Items on the criterion-referenced focal problem exams had significantly lower average difficulty when compared to items on the norm-referenced clinical clerkship exams.

CONCLUSIONS

Following are the conclusions of this study:

1. Item format affected both average difficulty and discrimination.

   a. Items in the single-answer multiple-choice (SA-MC) format had higher average values for discrimination than items in the multiple-answer multiple-choice (MA-MC) format.

   b. The SA-MC items were less difficult than the MA-MC items.

2. Stem orientation, positive versus negative, affected item discrimination but not difficulty. Items with positive stems were more discriminating than those with negative stems.

3. Informativeness of the item stem had some effect on item difficulty but no statistically significant effect on discrimination. Within the single-answer format, items with complete informative stems (SA-MC)

were more difficult than items with incomplete stems (US-MC). Within the multiple-answer format there were no differences.

SUGGESTIONS FOR FUTURE RESEARCH

This study was to a certain extent exploratory, attempting first to compare the MA and SA formats within medical education, then to point out areas of difference which might give direction to future research. This study showed large differences in difficulty and discrimination between the SA-MC and MA-MC items and offered some possible explanations. This should lead to additional research efforts in this area in the future. It would be premature to draw major conclusions or make broad generalizations on the basis of this one study. It must be considered one link in a whole chain of related research. It is to be hoped that this study will prompt future research efforts in the study of varying item types within medical education.

One direction this research might take would be replication under more thoroughly controlled conditions. It would be helpful to refine the techniques used to select item pairs, increasing the certainty that both items in each pair really were intended to measure the same knowledge or skill. It should also be possible to refine the techniques for choosing well-written MA-MC items for future studies. Since this format has been studied much less frequently than the single-answer format less was known concerning what kinds of item writing errors might be common within this format. This study has added to that body of knowledge so that future studies can have increased precision.

This study was exploratory in another way as well, trying to determine whether differences in item statistics similar to those found in studies

using professionally written or recast items. would also be found when independently written teacher-made items were used to compare item types. Some of the results found were similar, but there were enough instances of difference to help stimulate interest in further exploration of the performance of selected item types when all items are written by classroom instructors.

Throughout the results and discussion of this study one dimension of the test item seemed to come repeatedly to the forefront, the item idea upon which the item is based. In the present study uninformative stems were chosen as the focus for two comparisons. This focus is once removed from what seems now to be the more appropriate focus on the item idea. The question would not be, was the stem informative, but, can a clear singular item idea be identified as having been the basis for both the stem and options? Research exploring this important area would not be easy to design or carry out, but may have great potential in the ongoing search for better understanding of the principles underlying the writing of valid and reliable test items.

Finally, it must be noted that although the MA-MC format had poor item statistics in comparison to the SA-MC format, this type can still have a valid place in the overall testing process in medical education. Since most physicians obtain licensure by use of the National Board of Medical Examiners exams, which contain this item format, medical students need practice with its use. What item writers can gain from this study is the caution that this format can perform poorly, along with some specific suggestions concerning possible problems with the format (e.g. confusing wording in the stems and use of options with inappropriate content) that might be avoided with special care.

APPENDICES

APPENDIX A

SOURCE OF ITEM PAIRS - BREAKDOWN BY EXAM OF ORIGIN

APPENDIX A

SOURCE OF ITEM PAIRS - BREAKDOWN BY EXAM OF ORIGIN

Content Area-Exam Dates	No. of Students	Number of Item Pairs per Comparison: SA-MC vs. MA-MC	SA-MC vs. US-MC	MA-MC vs. US-MA-MC	SA-MC vs. NS-MC
Altered Consciousness					
Spring '78	28	4	0	6	4
Spring '77	47	3	1	1	1
Spring '76	40	4	1	2	1
Winter '76	47	0	1	1	1
Spring '75	12	1	0	1	1
		12	3	11	8
Diarrhea					
Winter '78	33	2	3	3	1
Winter '77	36	9	0	4	9
Spring '75	44	1	2	3	4
Fall '74	19	1	0	4	1
		13	5	14	15
Anemia					
Winter '78	29	3	3	2	7
Winter '77	43	0	0	0	4
Winter '76	40	2	2	3	1
Winter '75	39	0	0	0	5
Winter '74	16	2	2	2	2
		7	7	7	19
Jaundice					
Winter '78	21	8	3	6	3
Winter '77	36	5	2	1	3
Winter '76	40	0	1	3	4
		13	6	10	10
Chest Pain					
Fall '77	29	1	1	0	1
Fall '76	38	0	1	0	0
Spring '75	34	2	1	4	0
		3	3	4	1
Elevated BUN					
Winter '78	34	8	5	4	9
Winter '77	39	4	4	0	8
Spring '76	32	5	0	4	1
		17	9	8	18
Surgery					
Spring '78	23	8	3	3	7
Fall '77	31	4	2	0	4
Fall '76	13	8	1	0	3
		20	6	3	14

APPENDIX A (Continued)

SOURCE OF ITEM PAIRS - BREAKDOWN BY EXAM OF ORIGIN

Content Area-Exam Dates	No. of Students	Number of Item Pairs per Comparison: SA-MC vs. MA-MC	SA-MC vs. US-MC	MA-MC vs. US-MA-MC	SA-MC vs. NS-MC
Pediatrics					
Spring '78	35	1	1	1	2
Winter '78	41	0	0	0	2
Fall '77	34	1	0	0	1
Winter '77	53	1	1	1	2
Fall '76	34	3	0	0	1
Spring '76	32	2	0	2	0
Winter '76	40	2	0	0	2
Fall '75	26	2	0	1	2
Spring '75	18	3	1	0	0
Winter '75	36	0	0	0	3
Fall '74	30	2	1	0	4
Spring '74	23	0	0	0	2
Winter '74	35	0	3	0	4
Fall '73	39	0	2	0	1
		17	9	5	26
Internal Medicine					
Spring '78	35	0	0	0	3
Winter '78	41	1	0	0	0
Fall '77	34	1	0	0	4
Winter '77	53	3	1	2	5
Fall '76	34	5	0	0	0
Spring '76	32	1	0	0	6
Winter '76	40	0	0	0	2
Fall '75	26	2	1	0	1
Spring '75	18	3	0	2	2
Winter '75	36	1	0	1	1
Fall '74	30	4	1	1	3
Spring '74	23	0	0	0	1
Winter '74	35	0	0	1	3
Fall '73	39	1	0	0	1
		22	3	7	32

APPENDIX B

SAMPLE ITEM PAIRS SA-MC vs. MA-MC COMPARISON

APPENDIX B

SAMPLE ITEM PAIRS SA-MC vs. MA-MC COMPARISON

Clinical Science

1. A 32 year old female with intermittent diarrhea, constipation, and abdominal cramps of 18 months duration has a GI series of x-rays which reveal an abnormal terminal ileum and cecum with narrowing of the lumen and edema of the bowel wall. Tissue removed at surgery reveals inflammation in the mucosa, muscularis, and subserosa. The findings are most characteristic of

   1. ulcerative colitis
   2. amebic dysentery
   3. regional enteritis
   4. diverticulosis
   5. non-tropical sprue

   Diarrhea can be contracted via

   A. contaminated food
   B. contaminated water
   C. respiratory tract infection
   D. infected animals

2. The treatment of choice in functional diarrhea (irritable bowel syndrome) is

   1. cholestyramine
   2. diphenoxylate
   3. sodium carboxymethyl cellulose
   4. morphine
   5. propantheline

   Colonic motility in the irritable bowel syndrome differs from normal in that there is an increased responsiveness

   A. cholecystokinin (endogenous or exogenous)
   B. parasympathomimetic drugs
   C. meals
   D. cyclic AMP

3. The drug of choice for treating status epilepticus is

   1. diphenylhydantoin
   2. succinylcholine
   3. oxazepam
   4. thiopental
   5. diazpam

APPENDIX B
(continued)

Types of vascular headache include

A. migraine headache
B. hypertensive headache
C. cluster headache
D. headache caused by fever

4. The most sensitive test for detection of the hepatitis B surface antigen (HBsAg) is

1. counterimmunoelectrophoresis (CIEP)
2. gel diffusion
3. hemagglutination inhibition
4. radioimmunoassay (RIA)
5. complement fixation

Chronic active hepatitis is more likely if bridging necrosis is found. Other cues to this complication are

A. elevated SGOT and bilirubin at 2 months after acute symptoms
B. persistent hepatomegaly
C. the patient feels well
D. the HBsAg remains positive at 4 months

5. The ability of the neonatal liver to handle bilirubin may be enhanced by treatment of either the pregnant mother or the neonate with

1. salicylates
2. enterovioform
3. sulfonamides
4. phenobarbital
5. interferon

If serial studies of amniotic bile pigments demonstrate a severely affected fetus, which of the following should be tried to prevent its death from erythroblastosis fetalis

A. induce premature labor if at least 34 weeks old
B. inject the uterus with phenobarbital
C. attempt intrauterine transfusion if infant is too immature for delivery
D. begin phototherapy prior to delivery

6. A disease, generally idiopathic, which is characterized by widespread crescent formation throughout all glomeruli, a poor prognosis and increased renal failure leading to uremia is

1. acute pyelonephritis
2. membranous glomerulenephritis
3. rapidly progressive glomerulonephritis
4. necrotizing papillitis
5. proliferative glomerulonephritis

APPENDIX B

(continued)

Complications of acute pyelonephritis include

A. renal carbuncle
B. renal papillary necrosis
C. perinephric abscess
D. chronic pyelonephritis

7. Your patient is a 22-year-old who complains of a unilaterial throbbing headache that has awakened him from sleep for the past 3 nights. The pain is accompanied by lacrimation, and intense rhinorrhea. The entire episode subsides spontaneously in approximately 40 minutes. Your diagnosis is

1. tension headache
2. classic migraine headache
3. depression headache
4. cluster headache
5. headache caused by berry aneurysm

In a patient with major motor seizures, status epilepticus may be precipitated by

A. abrupt withdrawal of anticonvulsant medication
B. brain tumor
C. brain injury
D. hypoglycemia

8. A 36 year old man with recurrent episodes of abdominal paid due to peptic ulcer disease is admitted because of weakness progressive for 3 months. He had an episode of black stools 4 months previously during one of his bouts of active peptic ulcer symtpoms. History is negative for other symptoms. Physical examination reveals a pulse rate of 70, blood pressure of 120/65, pallor of conjunctiva and mucous membranes. The remainder of the physical examination is normal.

RBC 3.2 mil/mm^3
Hgb 6.2 gm/100 ml
Hct 24%
Retic Count 1.7%

Calculation of the red cell indices suggest that the anemia is

1. normocytic, normochromic
2. microcytic, hypochromic
3. macrocytic, hypochromic
4. macrocytic, normochromic
5. microcytic, normochromic

APPENDIX B

(continued)

A 70 year old man is admitted to the hospital because of increasing fatigue and exertional shortness of breath which has been developing over a period of several months. He is found to have macrocytosis on the peripheral blood smear and the bone marrow is megaloblastic in type.

Causes of this clinical picture would include

A. folic acid deficiency
B. chronic liver disease
C. vitamin V-12 deficiency
C. iron deficiency

9. The most consistent indication of an infiltrative liver lesion such as carcinoma or granuloma may be

1. greatly increased SGOT activity
2. hyperbilirubinemia
3. reticulocytosis
4. elevated serum alkaline phosphatase activity
5. increased LDG activity

The clinical symptoms of hepatitis are in some ways similar to and often confused with

A. infectious mononucleosis
B. herpes simplex infections
C. toxoplasmosis
D. idiosyncratic drug reactions

10. A 30 year-old woman is admitted to the hospital with a history of diarrhea for over 10 years. This is characterized by 8-10 bulky, greasy foul smelling stools/24 hrs. with 2-3 at night. Her appetite is good yet she had lost 15 lbs; her weight has been stable for several years. She complains of bloating, mild abdominal cramping and a good deal of flatulence. For 3 months she has had persistent mid-back pain. She considers herself "nervous". She is 5' tall and weights 98 lbs. Her tongue is rather smooth, her abdomen moderately distended and tympanic and there is tenderness over D_{12}. The P. E. is otherwise normal. The Hb is 10 G/dl, the Hct 36% and RBC $4 \times 10^6/mm^3$.

In writing admission orders one of the earliest diagnostic goals would be to determine if she has

1. abnormal stool flora
2. gastric hypersecretion
3. parasitic infestation
4. steatorrhea
5. blood in her stool

APPENDIX B
(continued)

Large volume (stool weight > 300 G/day) diarrhea is seen in

A. cholera
B. irritable bowel syndrome
C. celiac sprue
D. Crohn's disease of the colon

11. With anemia of chronic disease the plasma iron (Fe) and total iron binding capacity (T.I.B.C.) would most likely show which of the following

1. normal Fe, increased T.I.B.C.
2. normal Fe, normal T.I.B.C.
3. normal Fe, decreased T.I.B.C.
4. decreased Fe, increased T.I.B.C.
5. decreased Fe, decreased T.I.B.C.

Findings in pernicious anemia include

A. weakness
B. macrocytic erythrocytes
C. sore tongue
D. numbness and tingling in the extremities

12. Diffuse, symmetrical slowing of the EEG would most likely be associated with

1. a hemorrhage
2. an infarction
3. a supratentorial tumor
4. an abscess
5. a metabolic disorder

Pneumoencephalography is used for viewing the

A. cerebral arteries
B. posterior fossa contents and cisterns
C. cerebral lymphatics
D. ventricles

Pharmacology

1. Which of the following lists gives the correct ranking of the barbiturates in order of increasing duration of action

1. thiopental, phenobarbital, pentobarbital
2. pentobarbital, thiopental, phenobarbital
3. pentobarbital, phenobarbital, thiopental
4. thiopental, pentobarbital, phenobarbital
5. phenobarbital, pentobarbital, thiopental

APPENDIX B
(continued)

The major anticonvulsant drugs are believed to work by

A. suppressing abnormally discharging foci
B. carbonic anhydrase inhibition
C. reducing excitability of curcuit neurons
D. alterations in the acid-base balance

2. One of the important mechanisms of action of nitroglycerin which results in the relief of anginal pain is

1. a direct depressant effect on myocardial metabolism causing a reduced myocardial oxygen consumption
2. a direct effect on hemoglobin oxygen binding resulting in an increase in release of oxygen from the red blood cells in the myocardial capillary network
3. an increase in cardiac output resulting in an increase in coronary blood flow
4. direct effect on the cardiac pacemaker resulting in slowing of the heart rate
5. a drop in systemic blood pressure resulting in a decrease in cardiac work

Pharmacologic effects of nitroglycerin include

A. peripheral venous vasoconstriction
B. smooth muscle relaxation
C. increase in myocardial oxygen consumption
D. peripheral arterial dilitation

3. Mrs. Johnson is a 40-year-old woman in renal failure. She has an extra-renal infection caused by organisms susceptible to all of tetracyclines. Whicg of the following would be the best agent to use

1. chlortetracycline
2. oxytetracycline
3. doxycycline
4. minocycline
5. demecycline

Which of the following, in "fixed-dose" combination preparations, have a place in the approved management of infection

A. penicillin G - streptomycin
B. penicillin G - chlortetracycline
C. kanamycin - methicillin
D. trimethoprim - sulfamethoxazole

APPENDIX B
(continued)

4. An antibiotic has the following characteristics: Bactericidal; Resistance is infrequent; Activity is sharply limited to gram negative bacteria; Surface active agent leads to disorientation of the lipoprotein membrane. This antibiotic is

   1. penicillin
   2. gentamicin
   3. tetracycline
   4. polymyxin B.
   5. cephalosporin

   Penicillin G in repository form can safely be given by which route (s)

   A. sub q
   B. intrathecal
   C. IV
   D. IM

5. To prolong the absorption time, repository penicillin is administered

   1. intramuscularly
   2. intravenously
   3. intrathecally
   4. orally
   5. subcutaneously

   A patient with a urinary tract infection and known sensitivity to penicillin could be treated with

   A. ampicillin
   B. methicillin
   C. cephalexin
   D. lincomycin

6. When the urine is acid, the clearance of a drug is found to be less than the rate of glomerular filtration. However, when the urine is alkaline, the drug clearance is greater than the rate of glomerular filtration. The drug is a

   1. strong organic base
   2. weak organic base
   3. strong organic acid
   4. weak organic acid
   5. nonelectrolyte

   The rapid renal clearance of a drug is favored if the drug

   A. has low solubility in water
   B. reduces renal blood flow
   C. has a high degree of binding to plasma protein
   D. has low solubility in lipid

APPENDIX B
(continued)

7. Which of the following drugs finds its major usefulness in petit mal epilepsy

   1. dephenylhydantoin (Dilantin)
   2. phenobarbital (Luminal)
   3. primidone (Mysoline)
   4. trimethadione (Tridione)
   5. phenacemide (Phenurone)

   Which of the following would be indicated in status epilepticus

   A. morphine
   B. ethosuximide (Zarontin)
   C. succinylcholine (Anectine)
   D. diazepam (Valium)

8. The incidence of hypersensitivity reactions to cephalosporins is higher in patients who have shown allergic manifestations following the administration of

   1. gentamicin
   2. penicillin
   3. polymyxin
   4. sulfonamide derivatives
   5. tetracycline

   Which of the following are stabile in gastric acid and undergo good absorption after oral administration

   A. penicillin G
   B. oxacillin
   C. methicillin
   D. ampicillin

9. Which of the following parasympathically innervated functions is most sensitive to low doses of atropine

   1. salivary secretions
   2. vagal effects on the heart
   3. micturition
   4. gastric secretion
   5. accomodation of the eye

   Which of the following are antimuscarinic agents

   A. scopolamine
   B. atropine
   C. propantheline (Probanthine)
   D. morphine

APPENDIX B
(continued)

10. The antibiotic of choice for treating salmonella is

    1. carbenicillin
    2. erythromycin
    3. tetracycline
    4. ampicillin
    5. oxacillin

    Which of the following might be useful for treating a Pseudomonas infection

    A. Polymyxin B
    B. Gentamicin
    C. Carbonicillin
    D. Tetracycline

11. The most important and serious side effect of the use of gentamicin is

    1. elevation of blood urea nitrogen
    2. overgrowth of Candida on oral administration
    3. enterocolitis
    4. cardiac arrhythmias
    5. ototoxicity

    Gentamicin is most important in the treatment of serious infection, including those caused by

    A. pseudomonas aeruginosa
    B. Enterobacter
    C. Klebsiella
    D. Strep pneumoniae

Cardiology - Pediatrics

1. A continuous murmur (with diastolic accentuation) is heard over the primary aortic area in a 3-year-old-boy. The murmur is obliterated by compression of the neck veins on the right side and by assumption of the supine position. The most likely diagnosis is

    1. venous hum
    2. patent ductus arteriosus
    3. ventricular septal defect with prolapsed aortic valve cusp
    4. aortic aneurysm
    5. aortic stenosis and insufficiency

APPENDIX B
(continued)

Functional or "Innocent" heart murmurs occur frequently in children. Appropriate management should include which of the following

A. routine radiographs and ECG's on all children with vibratory murmurs
B. refer for further cardiac evaluation and possible cardiac catheterization if the murmur becomes louder when the patient has a fever
C. antibiotic prophylaxis before dental work
D. emphasize to parents that no restriction of activity is needed for a child with a functional murmur

2. A blood pressure cuff that is too small gives

1. false low readings
2. false high readings
3. slightly lower readings than usual
4. markedly lower readings than usual
5. accurate readings

A 12-year-old boy, on routine physical examination in the office, shows a blood pressure of 140/90. Indicated steps in his work-up and management should include

A. checking the pressure in his other arm
B. checking the pressure in his leg
C. evaluating the size of his arm relative to the size of the blood pressure cuff
D. rechecking the pressure after 15 minutes of quiet rest

3. An infant with early cyanosis, progressive cardiac enlargement and pulmonary plethora should be suspected of having

1. patent ductus arteriosus
2. coarctation of the aorta
3. vascular ring
4. complete transposition of the great arteries
5. Ebstein's anamoly

Cardiac failure in the first few weeks of life may be due to

A. coarctation of the aorta
B. paroxysmal atrial tachycardia
C. transposition of the great vessels
D. cerebral arteriovenous fistula

APPENDIX B
(continued)

Cardiology - Internal Medicine

1. A 50-year-old male arrives in the emergency department with a history of dyspnea for 4 hours. He has a history of recent onset of angina. Esam: P 130 reg., resp. 24, T 99po, BP 190/100. Lungs - diffuse inspiratory and expiratory wheezing. Cardiovascular exam - JVP greater than 10 cm., S3 gallop, and paradoxically split S2. Extremities - no edema or phlebitis. The most likely diagnosis

   1. acute asthma attack
   2. pneumococcal pneumonia
   3. acute pulmonary edema
   4. pulmonary emboli
   5. pneuthorax

   Which of the following is/are helpful in distinguishing pulmonary embolism from pulmonary infarction

   A. blood gases
   B. chest X-ray
   C. lung scan
   D. presence or absence of hemoptysis

2. Paracentesis abdominis is most likely to be of therapeutic benefit in a patient with peritoneal fluid due to

   1. tuberculous peritonitis
   2. nephrotic syndrome
   3. systemic lupus erythematosus
   4. hepatic cirrhosis
   5. congestive cardiac failure

   The therapeutic removal of large volumes of ascitic fluid by paracentesis may be complicated by

   A. ptosis of the abdominal viscera
   B. circulatory collapse
   C. acute gastric dilatation
   D. plasma albumin depletion

3. The aim of the initial therapy in acute pulmonary edema due to left ventricular failure is to

   1. slow the heart rate
   2. allay anxiety
   3. improve left ventricular contractility
   4. decrease pulmonary blood volume
   5. remove the excess fluid from dependent parts

APPENDIX B
(continued)

Myxedema is frequently associated with

A. increased cardiac size, as seen in the chest x-rays
B. hoarseness of voice
C. bradycardia
D. pretibial accumulations of subcutaneous myxedema

4. The most frequent mechanism of cardiac arrest in the hospitalized patient with an acute myocardial infarction is

1. ventricular fibrillation
2. asystole
3. electro-mechanical dissociation
4. cardiac rupture
5. atrial fibrillation

Morphine sulfate is often used to relieve pain in patients with acute myocardial infarction. Side effects that should be of concern in this situation are

A. bradycardia due to vagotonic action
B. hypotension primarily related to venous dilitation and pooling of blood
C. respiratory depression via direct action on the medulla
D. diarrhea

BIBLIOGRAPHY

BIBLIOGRAPHY

Albanese, M.A.; Kent, T.H.; Whitney, D.R. A comparison of the difficulty, reliability, and validity of complex multiple choice, multiple response, and multiple true-false items. Paper presented at the Annual Meeting of the American Association of Medical Colleges, Washington D.C., November 1977.

Board, C.; Whitney, D.R., The effect of selected poor item-writing practices on test difficulty, reliability and validity. Journal of Educational Measurement, 1972, 9, 225-233.

Boynton, M. "None of these" makes spelling items more difficult. Educational and Psychological Measurement, 1950, 10, 431-432.

Burmester, M.A.; Olson, L.A. Comparison of item statistics for items in multiple-choice and in alternative-response form. Science Education, 1966, 50, 467-470.

Campbell, D.T.; Stanley, J.C. Experimental and Quasi-Experimental Designs for Research. Chicago:Rand McNally College Publishing Company, 1963.

Cronbach, L.J. An experimental comparison of the multiple true-false and multiple multiple-choice tests. Journal of Educational Psychology, 1941, 32, 533-543.

_____. Note on the multiple true-false test exercise. Journal of Educational Psychology, 1939, 30, 628-631.

Dryden. R.E.; Frisbie, D.A. Comparative reliabilities and validities of multiple choice and complex multiple choice nursing education tests. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Washington D.C., April 1975.

Dudycha, A.L.; Carpenter, J.B. Effects of item format on item discrimination and difficulty. Journal of Applied Psychology, 1973, 58, 116-121.

Dunn, T.F.; Goldstein, L.G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 1959, 19, 171-179.

Ebel, R.L. Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1979.

_____. The ineffectiveness of multiple true-false test items. Educational and Psychological Measurement, 1978, 38, 37-44.

_____. Writing the test item. In E.F. Lindquist (Ed.) Educational Measurement, Washington, D.C.: American Council on Education, 1951, Pp. 185-249.

Eurich, A.C. Four types of examinations compared. Journal of Educational Psychology, 1939, 22, 268-278.

Flanagan, J.C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. Journal of Educational Psychology, 1939, 30, 674-680.

Frisbie, D.A. Multiple-choice versus true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

_____. The effect of item format on reliability and validity: A study of multiple-choice and true-false achievement tests. Educational and Psychological Measurement, 1974, 34, 885-892.

Harris, R.J. A Primer of Multivariate Statistics. New York: Academic Press, 1975.

Hughes, H.H.; Trimble, W.E. The use of complex alternatives in multiple-choice items. Educational and Psychological Measurement, 1965, 25, 117-126.

Magnusson, D. Test Theory. Reading, Massachusetts: Addison-Wesley Publishing Company, 1967.

McMorris, R.F.; Brown, J.A.; Snyder, G.W.; Pruzek, R.M. Effects of violating item construction principles. Journal of Educational Measurement, 1972, 9, 287-295.

Mehrens, W.A.; Lehmann, I.J. Measurement and Evaluation in Education and Psychology, Second Edition. New York: Holt, Rinehart and Winston, Inc., 1978.

Meuller, D.J. An assessment of the effectiveness of complex alternatives in multiple-choice achievement test items. Educational and Psychological Measurement, 1975, 35, 135-141.

Oosterhof, A.C.; Glasnapp, D.R. Comparative reliabilities of the multiple-choice and true-false formats. Paper presented ar the Annual Meeting of the American Education and Research Association, Chicago, Illinois, April 1972.

Pyrczak, F. Validity of the discrimination index as a measure of item quality. Journal of Educational Measurement, 1973, 10, 227-231.

Rimland, B. The effects of varying time limits and of using "right answer not given" in experimental forms of the U.S. Navy arithmetic test. Educational and Psychological Measurement, 1960, 20, 533-538.

Ruch, G.M.; Stoddard, G.D. The comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 1925, 16, 89-103.

Schmeiser, C.B.; Whitney, D.R. Effect of two selected item-writing practices on test difficulty, discrimination and reliability. Journal of Experimental Education, 1975, 43, 30-34.

Skakun, E.N.; Nanson, E.M.; Taylor, W.C.; Kling, S. An investigation of three types of multiple choice questions. Paper presented at the Annual Meeting of the American Association of Medical Colleges, Washington, D.C., November 1977.

Terranova, C. The effects of negative stems in multiple-choice test items. (Doctoral dissertation, State University of New York at Buffalo) Ann Arbor, Michigan: University Microfilms, 1969, 69-20, 512.

Wason, P. Response to affirmative and negative binary statements. British Journal of Psychology, 1961, 52, 133-144.

Wesman, A.G. Writing the test item. In R.L. Thorndike (Ed.) Educational Measurement. Second edition. Washington, D.C.: American Council on Education, 1971.

Wesman, A.G.; Bennett, G.K. The use of "None of these" as an option in test construction. Journal of Educational Psychology, 1946, 37, 541-549.

Williams, B.J.; Ebel, R.L. The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. Fourteenth Yearbook, National Council of Measurements Used in Education. Princeton, N.J., 1957, 63-65.

Williamson, M.L.; Hopkins, K.D. The use of "None-of-these" versus homogeneous alternatives on multiple-choice tests: experimental reliability and validity comparisons. Journal of Educational Measurement, 1967, 4, 53-58.

Zern, D. Effects of variations in question phrasing on true-false answers by grade school children. Psychological Reports, 1967, 20, 527-533.