

THE 1979



This is to certify that the
thesis entitled
A Comparison of Physician, Resident and Student
Performance in Two Medical Content Domains Using
Multiple-Choice and Simulated Clinical Encounter
Test Formats
presented by

Rivkah M. Lindenfeld

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Higher Education

Major professor

Date April 24, 1979



OVERDUE FINES ARE 25¢ PER DAY
PER ITEM

Return to book drop to remove
this checkout from your record.

--	--

A COMPARISON OF PHYSICIAN, RESIDENT, AND STUDENT PERFORMANCE
ON TWO MEDICAL CONTENT DOMAINS USING MULTIPLE-CHOICE AND
SIMULATED CLINICAL ENCOUNTER TEST FORMATS

By

Rivkah M. Lindenfeld

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Administration and Higher Education

1979

ABSTRACT

A COMPARISON OF PHYSICIAN, RESIDENT, AND STUDENT PERFORMANCE ON TWO MEDICAL CONTENT DOMAINS USING MULTIPLE-CHOICE AND SIMULATED CLINICAL ENCOUNTER TEST FORMATS

By

Rivkah M. Lindenfeld

The primary purpose of this study was to determine if the empirical relationship between scores on a multiple-choice test (MCQ) and performance ratings on a Simulated Clinical Encounter (SCE) hold across different levels of medical training and different medical content domains. Data was obtained from a field test of a newly developed certification examination for emergency medicine physicians. The sample (n=94) included three levels of medical training, namely physicians eligible for certification (n=36), second year residents in emergency medicine (n=36), and fourth-year medical students (n=22). The Multiple-Choice Test containing 199 items and the Simulated Clinical Encounters involving 4 different patients sampled two medical content areas, specifically, a Cardio-Pulmonary domain (CP) and a Skeletal-Trauma domain (ST).

The first hypothesis that the correlation between format scores would not be significantly different from zero was rejected. A correlation of .67 was significant at the .05 level using Grand Total scores on the MCQ test and the SCE test.

The second hypothesis posited that different levels of medical training would produce correlations between the two test formats that

would not significantly differ from each other. This hypothesis was accepted.

The third hypothesis assumed that a significant difference would not be found between the correlations that were obtained within each of the two different medical content domains sampled by the two tests. The results of the analysis showed that significantly different correlations between the two test formats were produced by the two content domains. However, when correlations within the three levels of medical training represented in the total sample are compared on the two medical content domains, significant differences are only observed for the residency population. The student population show similar differences in correlations on the two domains although these differences are not statistically significant at the .05 level. No differences in correlations on the two domains were observed for the physician population. Hypothesis 3, then, can be cautiously rejected with the caveat that it probably does not hold at the highest levels of training and practical experience.

The fourth hypothesis investigated whether SCE test format discriminated among the three levels of medical training better than did the MCQ test format regardless of medical content domains. The analysis of variance and Scheffé post hoc comparisons revealed that both test formats discriminate equally among the three levels of medical training.

As a result of this investigation the following conclusions were drawn:

1. This study has demonstrated that higher correlation can be obtained between MCQ and oral simulation, than were previously reported. The difference in the magnitude of the correlation were attributed to a) a greater variance of the population sampled in this study; b) reliability of the score produced by both test formats; and c) the item used in this study that were mainly designed to assess relevant clinical knowledge and specifically sampled two content domains of medical knowledge.
2. There seems to be some indication that multiple-choice tests and simulated clinical encounters may measure different competencies at the graduate and undergraduate level of medical training on some specific domains of knowledge.
3. This study demonstrated that different medical content domains may have an effect on the correlations that were observed for the graduate level of medical training and to some extent for the undergraduate level of medical training. No evidence was found to indicate that the different content domains had any effect on the observed correlations at the certification level of competence.
4. This study provides evidence for the validity of both the MCQ and SCE test formats. And that the concurrent

validity of the MCQ test format is suggested by the relatively high correlations with performance on SCE.

5. Both the MCQ and SCE test format demonstrated their ability to discriminate groups with different levels of clinical competence.

Therefore, it is suggested that both test formats be viewed as complementary measurement techniques and should be used, if possible, to obtain a more reliable and valid assessment of physician competence.

ACKNOWLEDGMENTS

I am indebted to many people who were helpful and without whose assistance and encouragement this study would have been impossible.

Dr. Jack L. Maatsch, director of the thesis, for being a major influence in the development of my educational and research skills; for his patience and support he provided opportunities for learning and is, in the truest sense, a teacher.

Dr. Richard Featherston, for assuming chairmanship of the committee, and for providing gentle but firm guidance.

Dr. Howard S. Teitelbaum, for his invaluable comments, critical review, friendly encouragements and for asking questions nobody else would.

Dr. Walter F. Johnson for his comments and warm encouragement.

Dr. Raywin Huang, a friend for his invaluable assistance in the analysis of the data.

Cathy Beegle for editing under unusual circumstances.

Marlene Dodge for typing the many drafts and friendly encouragements.

Judy Carley for typing the thesis.

These individuals have given freely of themselves and their time, that this study might be completed.

I would also like to thank the American College of Emergency Physicians and the American Board of Emergency Physicians for giving me the opportunity to use the data from the field test of the newly developed certification examination.

To all my colleagues and friends in the Office of Medical Education Research and Development, especially Dr. Ann Olmsted for her much needed friendship.

Last but not least, special thanks to my husband and son for their support and understanding.

I count myself fortunate, my sincere thanks to all.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
Chapter	
I. INTRODUCTION	1
Historical Background of Specialty Boards	2
Statement of the Problem	6
Purpose of the Study	9
Hypotheses	10
Definition of Terms	11
Summary and Overview	12
II. A REVIEW OF THE LITERATURE	13
Review of Reliability and Validity Studies of Oral Medical Specialty Certification Examinations	13
Simulations Used in Medical Specialty Certification Examination	18
Reliability and Validity of Simulations in Oral Medical Specialty Certification Examinations	24
Generalization from the Review of the Literature	31
III. DESIGN OF THE STUDY	33
Introduction	33
The Nature of the Population	34
Sampling Procedures	34
Organization and Procedures Used to Develop the ACEP Specialty Examination	36
Test Formats Developed for the Emergency Medicine Specialty Examination	37
Description of the Field Test	39
Examiner Orientation to the Field Test	40
Candidate Orientation to the Field Test	41
Description of the Treatment	42
Instructions to the Candidates Taking SCEs	45
Instructions to the Examiners	45
Administration of the Simulated Clinical Encounters	46

	Page
Rating Candidate Performance	46
Design of the Study	47
Analysis Methods	48
IV. ANALYSIS OF RESULTS	49
V. DISCUSSION AND SUMMARY	69
Conclusions	72
Recommendations	73
Summary	74
REFERENCES	77

LIST OF TABLES

Table	Page
1. Comparison of Means, Standard Deviation, Reliability and Correlation Coefficients Between Multiple-Choice Total Scores and Simulated Clinical Encounters Total Ratings Among Physicians, Residents and Students	51
2. Test of Statistical Significance of the Pairwise Differences Obtained in the Observed Correlations Among Physicians, Residents and Students	54
3. A Test of Statistical Significance Between Observed Correlations for the Total Population in the Cardio-Pulmonary and Skeletal-Trauma Content Domains	55
4. A Comparison of Means, Standard Deviation, Reliability and Correlation Coefficients Between Multiple-Choice Scores and Simulated Clinical Encounter Ratings Among Physicians, Residents and Students in the Cardio-Pulmonary Content Domain	57
5. A Comparison of Means, Standard Deviation, Reliability and Correlation Coefficients Between Multiple-Choice Scores and Simulated Clinical Encounter Ratings Among Physicians, Residents and Students in the Skeletal-Trauma Content Domain	58
6. Test of Statistical Significance of the Differences of the Observed Correlations Between Content Domains Among Physicians, Residents and Students	59
7. Analysis of Variance of the Multiple-Choice Grand Total Scores	61
8. Analysis of Variance of the Simulated Clinical Encounters Grand Total Ratings	63
9. Analysis of Variance of the Multiple-Choice Scores in the Cardio-Pulmonary Content Domain	64
10. Analysis of Variance of the Simulated Clinical Encounter Ratings in the Cardio-Pulmonary Content Domain	65

Table	Page
11. Analysis of Variance of the Multiple-Choice Scores in the Skeletal-Trauma Content Domain	66
12. Analysis of Variance of the Simulated Clinical Encounter Ratings in the Skeletal-Trauma Content Domain . . .	67

CHAPTER I

INTRODUCTION

In the United States, an individual wishing to practice general medicine must obtain a medical degree from an accredited medical school and successfully pass a state or national licensure examination. To practice as a specialist in a medical specialty, however, requires further graduate training in an approved residency training program. Upon completion of a residency program the physician may be certified by his or her peers in a particular medical specialty. This involves meeting certain qualifying requirements and sitting for a specialty certification examination to ascertain whether or not he or she has met acceptable standards of professional competence in the medical specialty.

The trend toward voluntary specialty certification after residency training has grown in recent years. At the same time, there have also been increasing doubts from both the public and the medical profession about the validity of medical specialty examination procedures. Do they indeed measure physician competence? Senior (1976) states:

"When, for example, medical audit reveals discrepancies (sic) between actual performance and the expectations of performance ability based on documents of certification doubt is cast on the certification procedure."

The most commonly used specialty certification tests contain two parts: an objective written examination and an oral examination. These two parts of the certification examination are purported to measure competency as defined by the National Board of Medical Examiners (1973) as:

"The ability and/or qualities for patient care, diagnosis, treatment and management as distinguished from theoretical or experimental knowledge. Clinical competence include

such elements as skill in obtaining information from patients, ability to detect and interpret symptoms and abnormal signs, acumen in arriving at a reasonable diagnosis and judgment in the management of patient."

Licensing and certification examinations are taken upon completion of an undergraduate program or a residency program. These examinations are developed by an agency outside the educational system, and are intended to evaluate the capability of a physician to perform health care generally or in his or her specialty.

HISTORICAL BACKGROUND OF THE SPECIALTY BOARDS

The standards of performance for the purpose of certification are established by the American Board of Medical Specialties in conjunction with each particular medical specialty. This board was previously known as the Advisory Board of Medical Specialties and was established in 1933. The Board was later organized into a loose federation of five medical specialties. It was again reorganized in 1970 under the current name of the American Board of Medical Specialties. It is now composed of representatives from twenty approved specialty boards. Holden (1969) described the objectives of the boards and states:

"The most important objective of the boards was to establish minimal requirements for the education of the specialist and to conduct examinations which when passed successfully by a candidate certified to his competence to practice the specialty."

For a candidate to be granted permission to take certification examinations in a medical specialty, approximately half of the twenty specialty boards require between one to two years of practice in the specialty in addition to the completion of an approved graduate program.

Examinations for the purpose of certification in a medical specialty were first introduced in 1917 by the American Board of Ophthalmology. Since that time, the medical specialty boards responsible for the certification of physicians in each medical specialty recognized the need to develop better methods for assessing physician competence. In many instances, these medical specialty boards were given assistance by the National Board of Medical Examiners.

Various test methods are used by the different medical specialty boards. The predominant methods are a combination of written and oral examinations. An outline of the essential characteristics of both written and oral examinations used in certification examinations will be made in the following paragraphs.

First, the evolution of written examinations has seen the virtual elimination of the use of the essay portion of certification examinations in favor of the multiple choice test format that was first introduced by the American Board of Internal Medicine in 1946. Today, all specialty boards use the objective multiple choice test format as part of the written examination, and are assisted in their development, scoring and analysis by the National Board of Medical Examiners. Three types of multiple choice questions (MCQ) are used: One Best Response, Matching, and Multiple True False. The One Best Response type of multiple-choice question is the predominant type. Examples of the three multiple-choice questions can be found in "Measuring Medical Education" by J.P. Hubbard (1971).

The continued efforts to develop examination formats and assessment techniques that would achieve better assessment of clinical competence led the National Board of Medical Examiners in 1961, to research,

develop and then use the Patient Management Problem (PMP), as an examination technique (Senior, 1976). Many medical specialty boards have incorporated this format into their certification process since that time.

The Patient Management Problem was developed to assess certain aspects of clinical competence not assessed by the MCQ test format, notably sequential problem solving and decision making. McGuire and Babott (1969) describe the PMP format as a method for the assessment of the candidate's ability to identify, solve, and manage patient problems. At the start of each problem, a short description of the patient's chief complaint is given. The candidate must then decide how he or she will solve the problem and record the decision by raising information through the use of a special pen. The information revealed directs the candidate to appropriate sections of this test under different headings (e.g. patient history, physical data, laboratory data, etc.). The PMP also named "written simulations" and "paper-and-pencil simulation" includes two types: linear and branching PMP. The linear type of PMP is solved sequentially. All candidates are forced to use the same sequence to solve the problems, as opposed to the branching type which is not. In the branching PMP candidates are directed to different sections of the problem depending upon which responses are made initially. However, because of difficulties involved in evaluating the many possible sequences of responses in the branching problems, the linear type of PMP is the most commonly used.

The oral examination component of medical specialty certification examinations, is considered by many specialty boards as the final and most valid step in the certification process.

Originally, in an attempt to assess physician clinical competence, the oral examination was conducted in the form of a bedside oral. This method included direct observation of a candidate's performance in a real clinical setting. The candidate was required to examine two or more real patients, while two or more examiners evaluated the candidate's clinical performance. The candidate was then asked by the examiners to discuss his or her findings pertaining to diagnosis, interpretation of laboratory and radiological data, and provide a plan of management for the patient's problems. The assessment of the candidate's performance in the clinical setting was discontinued in 1963 for a variety of reasons, among which were low reliability (Hubbard, 1971; Senior, 1976). Senior (1976) explains the reason for this low inter-rater reliability when he states:

"The scores awarded were influenced by three major variables: the competence of the candidate, the difficulty of the problem, and the level of expectation of the examiner. While only the competence of the candidate was at issue, the control of the other two variables posed substantial difficulties with respect to the reliability of the whole assessment."

In addition to the low reliability, logistical problems became more difficult. The growing number of candidates created problems in locating patients who would tolerate repeated examinations, and the financial burden for both the examiners and candidates increased.

Consequently, the bedside examination was replaced by the currently used oral examination in an interview format. And is very similar to that used in universities for doctoral candidates. This oral examination format has continued until recently as a dominant feature of the oral component of the certification examination. However, questions concerning the reliability and validity of the oral examination were also raised.

In 1962 extensive studies and revisions of the oral test format were initiated by the American Board of Orthopaedic Surgery and the Center for the Study of Medical Education of the University of Illinois. As a result, the oral role-playing simulation was introduced as part of the oral certification examination administered by this specialty. Oral role-playing simulations are described by Maatsch et. al., (1977) whereby:

"Personnel are trained to act like patients, simulation material are provided to aid personnel in role playing patients or others in health care setting."

STATEMENT OF THE PROBLEM

Currently, the most commonly used test formats in certification examinations are written objective tests, predominantly consisting of multiple-choice questions. PMPs, traditional oral examinations, and in some instances the oral role-playing simulations are used to augment the MCQ test.

Theoretically, performance demonstrated in examination settings should closely parallel the required behavior in real life. As stated by Tyler (1950):

"An established educational principle is that an evaluation should allow the student to duplicate the type of behavior being evaluated."

If we accept this principle, then, the multiple-choice test format which mainly assesses factual knowledge may not be adequate. Although proven highly reliable and objective, the multiple-choice test format has been questioned. Abrahamson (1976) clarifies:

"The problem should be clear by now. Our best examinations procedures are written and in the objective format, unfortunately, while these examinations have extremely high reliability and objectivity, there are serious

questions raised concerning their validity. Somehow or other, the use of objective type examinations has to be justified through demonstrating that the results of such examinations procedures are significantly and closely related to the competency to practice specialty medicine."

On the other hand, the traditional oral examinations have also been faulted for their lack of reliability and objectivity. Although they are useful in expert assessment of interaction and in observation of attitude and communication skills, Hubbard (1971) remarks:

"Reliance upon the widely accepted and time honored oral examination is, however, widely challenged for purposes of certification at the professional level. Examiners and examining boards appear to be increasingly aware that examinations are a form of measurement and, like other forms of measurement are subject to test accuracy. When reliability of the oral examination is studied, it almost invariably fails to equal the reliability that can be demonstrated for good multiple choice examinations."

In a recent study conducted by Williamson (1976) that examined the validity of certification procedures, it was found, that when certification scores and measures of actual clinical performance were compared that there seemed to be little or no relationship between grades on certification examinations and the quality of subsequent clinical performance. Williamson recommended:

"... Some immediate action aimed at trying to improve examination content validity. For immediate and long range planning the most needed action involves improving criterion validity or predictive validity of these tests."

The problem then becomes that, while performance on multiple-choice examinations has been highly reliable, that same performance may not be predictive of actual clinical performance.

A question must be raised as to whether multiple-choice examinations, although seemingly objective and efficient, can in fact reliably predict performance in the actual clinical setting. Simulations provide an

alternative to the actual clinical setting that, with proper design, offers new possibilities for the assessment of clinical skills that define competence in a physician. The added cost and time required in simulations is counterbalanced by their usefulness. Previous studies have shown that simulations seem to meet acceptable standards of content validity (Levine and McGuire, 1968; Lamont and Hennen, 1972). As Levine and McGuire (1968) state:

"The use of role-playing as an evaluation technique appears to provide insights into important dimensions of performance not sampled by more conventional methods of testing and gives promise of extending the now limited usefulness of oral examinations."

Studies which measured the relationship between MCQ tests and oral role-playing simulations were performed by Levine and McGuire (1968, 1970). In these studies the correlation coefficients that were obtained ranged from .19 to .35. These correlations were interpreted by the authors as an indication that these two test formats measured somewhat different aspects of physician competence. Lamont and Hennen (1972) also measured the relationship between MCQs and simulated office orals obtaining a correlation coefficient of .09. They also interpreted the results as an indication that the two test formats measured somewhat different aspects of physician competence.

On the other hand, Kelley et. al., (1971) obtained a correlation coefficient in the low .60's when MCQ test scores were correlated with scores from a structured oral examination. He suggested that these results could be interpreted in a different manner and stated that:

"In the judgment of the ABA examiners, the written examination and the oral examination measure somewhat different aspects of knowledge, skills and clinical competence of the candidate. The correlations between the written and the oral tend to confirm this judgment."

on the other hand, the fact that the correlations between the two sections of the oral is not much higher than the correlation between the oral and the written suggests another possibility. Perhaps each section of the oral is simply probing somewhat different areas of medical knowledge than those being tapped by the written examination rather than measuring areas of 'clinical competence' not covered by the written test."

In conclusion, studies cited in the preceding discussion indicate the need to improve the validity of certification examination procedures in medical specialties. To date, multiple-choice tests have not proven to be effective in predicting future performance in clinical settings.

However, the use of clinical simulations offers an opportunity to assess physician's clinical performance as a substitute for actual clinical performance. A study of the correlations of subjects performance on the two test formats will address whether or not multiple-choice tests measure different competencies, or simply, sample other areas of medical knowledge. This study may help resolve this issue, and lead to certification examinations that have greater predictive validity.

PURPOSE OF THE STUDY

The purpose of this study is to investigate the empirical relationship between a multiple-choice test format and an oral Simulated Clinical Encounter test format. The study uses data collected to experimentally evaluate a newly developed certification examination for the American College of Emergency Physicians (1977). Specifically, this study will seek to determine if a relationship exists between scores derived from a multiple-choice test and examiner ratings of performance on Simulated Clinical Encounters. It will further determine whether the observed relationship between these two test

formats holds when considering specific medical content domains and different levels of physician training.

The third objective will be to determine which of the two test formats best discriminate between: a) physicians eligible for certification; b) residents in emergency medicine; and c) medical students.

These objectives can be translated into the following hypotheses:

HYPOTHESES

1. Correlations between total scores derived from a multiple choice test, and ratings of performance in oral simulated clinical encounters will be positive and significantly different from zero.
2. Groups with different levels of training will produce correlations between total scores derived from a multiple choice test and ratings of performance on oral simulated clinical encounters that are significantly different from each other.
3. When knowledge of two different medical content domains are sampled, correlations between scores derived from a multiple choice test and ratings of performance on oral simulated clinical encounters will be significantly different from each other.
4. Ratings of performance on oral simulated clinical encounters will differentiate between the three levels of training significantly better than scores derived from a multiple choice test; and this relationship will hold regardless of medical content domains.

DEFINITION OF TERMS

The following definitions are given to clarify the important words and terms that are used in this study.

Multiple Choice Questions (MCQ): MCQ's have two parts: a) a stem consisting of a direct question or an incomplete statement; and b) two or more options consisting of answers to the question or completion of the statement. The examinees task is to choose the correct, or best answer option in terms of the question posed by the item. (Ebel, 19)

The multiple choice test format used in the ACEP Certification Examination will consist of a question and one-best answer option, and are classified in two categories. The description follows:

Level I MCQ's: Lower cognitive level: Recall recognition items which only require the examinee to remember or recognize specific facts, definitions, or standard procedures. (Downing, 1977)

Level II MCQ's: Higher cognitive level: Problem solving/ Application requiring the examinee to apply factual knowledge to unfamiliar situations, to reason, to make evaluative judgments to clinical problems. (Downing, 1977)

Simulated Clinical Encounter (SCE): A carefully planned simulation of a real patient(s) case which a physician might encounter. The SCE assesses how well a candidate can diagnose a patient's medical problems, how well the candidate uses appropriate cognitive, affective, and psychomotor skills, and how well the candidate manages the patient from initial contact to discharge. The examiner has specific history, physical, laboratory and case outline data for reference. There also may be a variety of materials (x-rays, lab reports, etc.) that are handed to the candidate for interpretation, if such data has been

ordered. The candidate plays the role of the physician, and the examiner plays the role of the patient and other health providers as the need arises, in addition to that of actual examiner. (Maatsch et al., 1977)

SUMMARY AND OVERVIEW

This chapter reviewed the requirements made of physicians if they wish to practice medicine as specialists in a medical specialty. The historical background of the Medical Specialty Boards was briefly discussed. And the evaluation of the various tests currently used in medical specialty certification examinations were reviewed. Also discussed were the concerns related to the validity of the various test methods currently used in certification examinations, and the relationship of these concerns to physicians future performance. The purpose for this study and four hypotheses concerning the relationship between multiple-choice tests and simulated clinical encounters are identified.

Chapter II will consist of a review of the literature pertaining to certification examination techniques. Particular emphasis will be on the oral examinations used in certification examinations. Chapter III will describe the design of the study. An analysis of the data will be described in Chapter IV. Chapter V will provide a discussion of the results, conclusions, recommendations and summary of this study.

CHAPTER II

REVIEW OF THE LITERATURE

The primary focus of this study is on the relationship between written objective multiple-choice tests and oral simulation tests in a medical specialty certification examination. Consequently, the review of the literature is, of necessity limited to studies published since 1966, when simulations were first introduced as an assessment technique in the oral component of medical specialty certification examination. Earlier studies which investigated the relationship between objective written multiple-choice examinations and oral examinations will be included only if they are of special significance to this study. The review of the literature has been organized under the following headings:

1. Review of Reliability and Validity studies of the oral examination in medical specialty certification examination.
2. Description and use of simulations used in medical specialty oral certification examinations.
3. Review of reliability and validity studies of oral simulations in medical specialty certification examinations.

REVIEW OF RELIABILITY AND VALIDITY STUDIES OF THE ORAL MEDICAL SPECIALTY CERTIFICATION EXAMINATIONS

In 1954, Dr. J. Cowles addressed the Advisory Board of Medical Specialties in which he expressed concern about the need to improve the oral component of the certification examinations. He emphasized that efforts to improve the oral examinations were equally important to the efforts to improve written examinations.

By the early 1960's, researchers in a number of medical specialty boards realized that the written component of the certification examination which used objective multiple-choice was well established and accepted as a reliable testing format. They began to focus their efforts on making the oral medical specialty certification examination format equally valid.

Bull (1959) conducted correlation studies between multiple-choice tests and oral examinations to investigate the relationship between the two test formats. The analysis yielded a correlation coefficient which was not significant. The author accounted for the low correlations by noting that factual knowledge as assessed by the multiple-choice examinations play a small part in the scores allotted for performance on oral examinations. Attempts to specifically isolate a student's ability to elicit physical signs when assessing performance in an oral examination revealed that only sixteen percent (16%) of the score allotted to the oral examination dealt with this specific skill. A third score which had been obtained from interviews with students on a non-medical subject by a non-medical examiner correlated .45 with the final examination grade. These results provided more direct evidence on the effect of a student's personality in an oral examination according to Bull.

In his conclusions, Bull posed the possibility that the influence of a physician's personality on the score given in an oral interview examination should not be overlooked. He further states that:

"I do not feel that we are doing a great injustice to our students by continuing interview examinations. Probably in the last analysis a doctor's personality is as important as his knowledge."

Carter (1962) investigated the assumption that oral examinations were unreliable, using data made available to him by the directors of

the American Board of Anesthesiology. Analysis of the data produced an inter-rater reliability coefficient of .62 between two examiners who had rated the same candidate in a single session. The Spearman-Brown Prophecy Formula was applied, and indicated a reliability coefficient of .78 for the total oral examination. He also investigated the relationship between the oral examination and the multiple choice test of the certification examination. This study yielded a correlation coefficient of .45. Based on these results, Carter concluded that, when oral examinations are systematically and carefully conducted, they could be shown to be reliable. He also suggested that the positive but moderate correlations between the multiple choice examination and oral examinations could be interpreted as evidence that the oral examinations assess aspects of competence not adequately assessed by MCQs.

Observational studies of the oral examinations of the American Board of Orthopaedic Surgery were conducted by McGuire (1966). As a result of these observations McGuire concluded that the oral examinations: 1) predominantly assessed the ability to recall factual knowledge rapidly; 2) revealed that candidates rarely cited evidence to support their answers; and 3) showed the standards by which examiners conducted and judged candidate performance were not clear nor uniformly applied.

Pokorny and Frazier (1966) reported on their observations of oral examinations administered annually to psychiatry residents. The report was based on an investigation of examinations administered in 1965 to sixteen residents. This examination consisted of a practical examination, an oral interview examination, and a written examination. The primary external criteria used in this study was supervisors' ratings.

The purpose of this study was to evaluate the three methods of testing as they related to each other.

Correlation analysis between the three test formats yielded a correlation coefficient of .06 between the average grade on the practical examination and the oral examination, and between the written examination of .07. The correlation between the averaged grade on the oral examination and the written examination yielded a correlation coefficient of .73. Based on the results of their study, Pokorny and Frazier suggested that oral examinations could be considered a viable test method for the purpose of certification and licensure, and for screening incompetent physicians on a pass/fail basis. They should not, however, be considered when a precise ranking and grade are required.

Foster et. al. (1969) analyzed data from certification examinations for the Pediatric Board of Cardiology. Results showed differences among examiners when they scored candidates, however, the nature of these differences could not be identified. Examiners were found to consistently rate candidates either high or low on scales, and consistently asked specific types of questions.

Kelley et. al. (1971) conducted studies of the oral examination for the American Board of Anesthesiology, and obtained an inter-rater reliability coefficient that ranged from .69 to .80 averaging .75 based on examiner agreement in the rating of candidates in a single session. Correlation studies between the MCQ and oral examinations yielded a correlation coefficient in the low 60's after corrections for attenuation. Based on the results of the study the authors concluded that the oral examinations are more reliable than has previously been expected. Again, the positive but moderate correlation between the

MCQ and oral examination format suggests that each format assesses particular aspects of a physician's competence that are not assessed by the other test format. However, the authors suggested, that because the correlation between two oral examinations was not much higher than the correlation between the multiple-choice examination and oral examination another reasonable assumption might be that oral examinations evaluate different aspects of medical knowledge than those assessed by the MCQ examination rather than assessing different aspects of clinical competence. And that, the positive but moderate correlation enable the maximum usefulness of both test formats in certification examinations.

In summary, since the beginning of the 60's medical specialty boards began to concentrate their efforts on improving the reliability and validity of oral certification examinations. The studies undertaken by the various medical specialty boards of the oral examinations revealed some interesting results. First, if oral examinations were given in a structured and systematic manner, more acceptable inter-rater reliability could be achieved (Carter, 1962; Kelley et al., 1979). Second, studies investigating the relationship between multiple-choice examinations and oral examinations revealed positive but low to moderate correlations. These results were interpreted as an indication that the two test formats measure somewhat different aspects of knowledge, skills and clinical competence. However, Kelley (1971) suggests an alternative interpretation, and that is, that the two test formats measure different areas of medical knowledge, rather than different areas of clinical competence not assessed by the MCQ test.

The varied efforts and research conducted on the reliability and validity of oral tests of medical specialty certification examinations

eventually led to the development and use of simulations as an oral test method.

SIMULATIONS USED IN MEDICAL SPECIALTY CERTIFICATION EXAMINATIONS

The past decade has seen an increasing use of oral simulations in lieu of traditional oral examinations to assess aspects of physician competence not assessed by multiple-choice tests and Patient Management Problems in medical specialty certification examinations.

The use of simulations as an assessment technique provides the opportunity to assess aspects of competence recognized to be important in overall competence, for example physician interaction with patients and colleagues. The structured and standardized simulation format provide the necessary prerequisite to achieve better reliability and validity of the oral format in certification examinations. The advantages of simulation when compared to multiple-choice test is succinctly stated by Maatsch and Gordon (1978):

"Simulations stress the application of relevant knowledge and skills in a manner appropriate to the clinical problem or task presented. Multiple-choice examinations test the ability to select the best alternative offered. The latter abilities are not called upon frequently or directly in clinical reality, so the evaluator can only assume that the student's possession of factual knowledge demonstrated on multiple-choice test, will correlate highly with his ability to appropriately apply the knowledge and other skills appropriately in clinical situation."

Attempts to improve methods for assessment of aspects of physician clinical competence that were not adequately assessed by MCQs had led to the development and use of Patient Management Problems as a test format. However, the increase use of this test format in certification examinations has brought to light certain of its limitations and drawbacks.

Senior (1976) identifies the following problems: 1) the inability to prevent candidates from reading ahead and obtaining clues to the possible problem; 2) increased difficulties in the development of questions which avoid cueing; and 3) no method of recording candidate's choice of sequence in the resolution of the problem.

A study was conducted by McCarthy (1966) to determine whether differences in achievement could be identified using PMPs and oral role playing simulations. In both test formats, identical clinical problems were used. The study revealed that when visual cues were present as was the case in the PMP format, there was an increase in the total number of options chosen by the students and an increase in the extent to which these options were correct, particularly when compared with student's performance on oral role playing simulations.

The raw scores suggested that students who encountered difficulties in the non-cued (simulations) format were especially helped by the cued format (PMP). The mean total score for the oral role playing simulations was 109.3, and the corresponding mean for the PMP was 152.6. The correlation coefficient between the scores from the two test formats was -0.09, which McCarthy suggests may be an indication that the two test formats may assess different aspects of competence. In his concluding remarks concerning this study, he states:

"Cueing items seemed to aid particularly the students unable to recall relevant information. The suggestion is made that the cued erasure-type problems, and perhaps other formats involving selection from a list of responses may not provide optimal evaluation of certain aspects of physician performance."

To date, three medical specialties use some form of role playing for their oral certification examinations. These are the American Board of Orthopaedic Surgery (ABOS), the Canadian College of Family Physicians (CCFP), and the American College of Emergency Physicians (ACEP).

Levine and McGuire (1968) describe three types of role playing simulations developed for, and used by the ABOS. They are:

1. The Simulated Diagnostic Interview.
2. The Simulated Proposed Treatment Interview.
3. The Simulated Patient Management Conference.

These simulations form a structural problem examination where the candidate plays the role of the physician and the examiner provides appropriate information to the candidate as he or she requests it. Each of three simulations is designed to assess different aspects of physician competence.

The Simulated Diagnostic Interview is designed to assess the candidate's ability to obtain from a simulated patient a medical history, physical findings and laboratory findings. The objective of the Simulated Proposed Treatment Interview is to assess how well the candidate in the role of the physician can relate to the examiner in the role of the patient, explaining the nature of the patient's illness and gain patient cooperation for a proposed treatment. In the Simulated Patient Management Conference, five candidates are provided with basic information about two medical problems and are required to discuss the management and treatment of each case. The aim of this simulation is to assess minimum acceptable competence similar to cases discussed in a typical staff conference.

Lamont and Hennen (1972) and Van Wart (1974) describe the three types of oral simulations used in the Canadian College of Family Physician certification examinations. They are:

1. The Formal Oral
2. The Role-Playing Oral
3. The Simulated Office Oral

The formal oral examination is a tightly structured format, clinical content and protocols of the conduct of this examination are set in advance. The examiner provides the candidate with information on the preselected cases, and essentially assesses the candidate's problem-solving ability. In the Role-Playing Oral the examiner playing the role of the patient and the candidate the role of the physician assesses the candidate's ability to manage and treat the medical problem presented to him or her. The Simulated Office Oral assesses the candidate's interviewing skills to direct information from a patient, and establish rapport with a patient. For the purpose of this oral simulation examination "simulated patients" are used and are played by actors trained to simulated specific medical problems.

The American College of Emergency Physicians (ACEP) is the third medical specialty to incorporate the use of simulation in the oral specialty certification examination. The simulations used in this newly developed certification examination for ACEP were an outgrowth of the Patient Games that were developed at Michigan State University (M.S.U.) by the Office of Medical Education Research and Development (OMERAD) under the guidance of J.L. Maatsch, Ph.D. and in cooperation with faculty from the College of Human Medicine at M.S.U. These Patient Games were developed in an effort to find new methods of teaching medical students decision making strategies and other clinical aspects of patient care. The Patient Games are an instructional and/or evaluation technique that is used by clinical faculty, and described by Maatsch (1974) as:

" ... providing clinical information in realistic scenarios consistent with classroom constraints, learning principles and instructional objectives."

As a result of this successful method of clinical instruction and evaluation, the Office of Medical Education Research and Development at M.S.U. and in cooperation with the American College of Emergency Physicians, developed a multi-format criterion based certification examination (Maatsch, et. al., 1977) for the Emergency Medicine Specialty.

The Simulated Clinical Encounters (SCE) are used as the oral component of the certification examinations, and include two types of oral role playing simulations. Generally, the SCEs assess appropriate cognitive knowledge and selected psychomotor skills required of emergency physicians. Because this certification examination is criterion-referenced the following guidelines were adhered to in the selection of the content for, and the development of, the Simulated Clinical Encounters:

1. Frequently encountered medical cases and problems in emergency medicine.
2. Necessary knowledge (clinical or basic) that must be remembered at all times during the practice of emergency medicine.
3. Problems and knowledge related to time-sensitive emergency medicine decision and treatment.
4. Problems and knowledge related to life-threatening situations.
5. Problems or knowledge that is common to all emergency department settings.

The Simulated Clinical Encounter oral examination includes two types of oral role-playing simulations. They are:

1. The Simulated Patient Encounter (SPE).
2. The Simulated Situation Encounter (SSE).

The Simulated Patient Encounter is designed to measure candidate ability to assess, diagnose and manage a single patient. The candidate functions in the role of an emergency physician. The examiner, in addition to rating the candidate's performance, plays the role of the patient, colleague, nurse and other health personnel when required.

The Simulated Situation Encounter is more complex than the SPE, and is designed to assess the candidate's ability to effectively utilize knowledge and skills, and provide leadership in an emergency situation. In the SSE the candidate is required to treat three patients in an emergency situation and assessed in his own ability to rapidly diagnose, triage, and organize available health care while performing several life saving procedures.

In summary, the continuous efforts to develop more reliable and valid oral testing methods to assess aspects of physician clinical competence not assessed by multiple-choice tests and/or Patient Management Problems, and to provide an alternative to the oral test methods as has led to the introduction and use of oral simulations as an assessment technique in medical specialty certification examinations.

Of the various simulation formats currently in use the Simulated Clinical Encounters represent the most recent addition to this method of evaluation. Through the use of simulation, two important features were introduced in oral examinations to make this test format more valid than the traditional oral examination. These are: the standardization of the oral examination, and the unique design that allows for the assessment of specific aspects of competency not assessed by other test methods, but which are considered necessary for the evaluation of the overall competence of a physician in a clinical specialty.

The effectiveness of simulations used in certification examinations prior to the development of SCEs as demonstrated by studies of reliability and validity is reviewed in the following section.

RELIABILITY AND VALIDITY OF SIMULATIONS IN ORAL MEDICAL SPECIALTY CERTIFICATION EXAMINATIONS

Studies available in the literature pertaining to the reliability and validity of oral certification examinations that use simulations as a test method are limited to investigations that were based on the data collected from certification examinations that were administered by two medical specialty boards: The American Board of Orthopaedic Surgery and the Canadian College of Emergency Physicians.

Levine and McGuire (1968) report on the study based on scores obtained from the American Board of Orthopaedic Surgery Certification examination administered to 383 candidates in 1966. Three types of oral role playing simulations were used for the oral examination. These were:

1. The Simulated Diagnostic Interview
2. The Simulated Proposed Treatment Interview
3. The Simulated Patient Management Conference

To assess the reliability of these assessment techniques, inter-rater reliability studies were conducted. The inter-rater reliability coefficient obtained through the correlation of scores of two examiners rating candidate performance on the same case at the same time was .58 for the Simulated Diagnostic Interview, .72 for the Simulated Proposed Treatment Interview, and .14 for the Simulated Patient Management Conference. The authors summarized their observations upon completion of the study by stating:

"Rater reliability as measured by the correlation between two independent ratings is sufficiently high for the Patient Interview to indicate that these techniques should prove to be of value in a battery of tests designed to evaluate individual candidates."

However, they also report that the results of a one way analysis of variance showed that some examiner teams applied more stringent standards when scoring than did other examiner teams. And suggest, that the following should be considered in future administration of the oral simulations: 1) the importance of training the examiners; 2) that judgment about individual candidates be made on the basis of a sufficiently large number of observations so as to eliminate examiner bias; and 3) the results obtained for the Simulated Patient Management Conference indicated the need to revise the rating techniques that were used in this type of oral simulations.

The content validity and construct validity of the oral examinations were also evaluated. Content validity studies based on observation revealed that most observers felt this method had assessed certain behavioral aspects of competence not assessed by the other test formats or by the traditional oral examination.

Construct validity studies were also performed correlating the scores obtained on role-playing simulations and other test formats (MCQ and PMP) used in the certification examination. The correlations that were obtained were generally low, however, they felt that the results were compatible with their hypotheses, that is, that simulation as an evaluative technique measured some components of professional competence not assessed by the traditional test formats. In their conclusions Levine and McGuire stated that:

"The use of role playing as an evaluative technique appears to provide insight into important dimensions of performance not sampled by more conventional methods of testing and gives more promise of extending the now limited usefulness of oral examinations."

Another study conducted by Levine and McGuire (1970) with the purpose of investigating the validity and reliability of oral examination. This study was based on data collected from the 1969 certification examination administered by the American Board of Orthopaedic Surgery.

Three problem solving oral simulations were used to assess candidate performance on diagnostic decision making, emergency treatment, and management of complications. The candidates were rated on the following aspects of competence:

1. Recall of Factual Information.
2. Analysis and Interpretation of Clinical Data.
3. Problem Solving Abilities -- Clinical Judgment.
4. Relates Effectively: shows desirable attitudes.

Reliability of this problem solving oral examination was obtained by pooling the scores for the entire group of oral examinations. The authors contended that the pooling of the scores was legitimate, since all oral scores were pooled for the purpose of certification. To obtain an estimate of the reliability test format, the Spearman-Brown correlation formula was used. This yielded a reliability estimate of .47 for the four combined tests, with an average reliability of .18.

The authors found these results to be consistent with those obtained in a previous study, in which an analysis of variance on the four groups of candidates where each group was rated by the same team of examiners. The reliability estimates that were obtained ranged from .40 to .63 with an average of .53. The authors concluded that:

"Although pooled data from four oral examinations may be extremely useful in generalizing about groups, they are not sufficiently reliable to use alone to certify individuals; to be of value for this purpose, they must be used in combination with other test data."

Content validity of the problem-solving simulation oral examination was investigated by using two methods: systematic observation and a questionnaire. The authors considered content validity of this test attained if this examination did, in fact, assess a candidate's ability to solve problems. Results of the observation and questionnaire data reflected opinions of both the examiners and the candidates, and showed that this test format had high content validity as reflected by the behavior of the candidates and examiners.

Concurrent validity studies were performed by correlating supervisor's rating of candidates on ten traits (skills, knowledge, attitude, etc.) with the sub-scorers of each of the oral tests, and other test formats. Correlation coefficients obtained between supervisor ratings and multiple choice test scores ranged from .10 to .33. And, between supervisor ratings and the sum of the scores for the problem solving oral from .17 to .30. Summarizing their study Levine and McGuire observed:

"The results obtained in the Orthopaedic Training Study suggests that the methods described above for structuring the examination, standardizing case materials, training the examiners and pooling their ratings can be employed to minimize many validity and reliability problems that plague traditional oral examinations and can, thereby, increase substantially the arsenal of techniques available for the assessment of clinical competence."

Studies have also been conducted by Levine and McGuire (1970) to investigate the use of role-playing simulations to assess affective skills in medicine, based on scores obtained from the Orthopaedic Surgery Certification Examination.

For the purpose of this investigation, role-playing simulations were specifically designed to assess affective skills such as physician-patient and physician-colleague relationships, and were part of an extensive battery of tests including multiple-choice tests, and Patient Management Problems test, and oral problem-solving simulation tests.

The reliability of this test method was investigated by means of inter-rater reliability studies correlating the scores given by two examiners rating the same candidate at the same time. Results obtained yielded an inter-rater correlation coefficient of .73 and an overall reliability coefficient of .84.

To investigate the construct validity of the test, it was administered to residents at different levels of training. The results of this investigation revealed that 29 first-year residents achieved a mean of 6.2 (on a 12 point scale) and a standard deviation of 2.6, and 79 fourth-year residents achieved a mean of 7.4 and standard deviation of 2.8. Although the authors (Levine and McGuire) did not consider these results to be statistically significant, they considered the results to be understandable in view of the data obtained from questionnaires in which residency program supervisors had commented that very little supervision and observation is done by residents as they relate to patients and colleagues. In addition, forty percent (40%) of the examiners who administered this oral simulation examination agreed that, in general, candidates were not adequately trained to partake in this type of examination.

The relationship between the oral simulations and the multiple-choice test format was also investigated. Results of this investigation show that the correlation between the multiple-choice test and the problem solving simulation ranged between .29 and .35, and between the

multiple-choice test and the affective oral simulation correlation coefficient of .19 was obtained.

A study by Lamont and Hennen (1972) investigated the "simulated office" oral examination based on scores collected from certification examinations administered by the Canadian College of Family Physicians. The authors report that the content validity of the simulated office oral examination was not performed in a systematic manner. However, comments made by both candidates and examiners revealed that both groups considered this type of testing method to be the closest thing to real-life ever experienced in an examination.

In addition to the simulated office oral examination, two different types of oral examinations were administered: the formal oral and the role-playing oral examinations. The candidates participating in the certification examination were asked to rank the three oral test formats. Data showed that forty-four percent (44%) of the candidates felt that the simulated office oral examination was the most valid, as an evaluation technique to assess physician's competence. Thirty three percent (33%) considered the formal oral examination to be the most valid, and sixteen percent (16%) felt that the role-playing oral simulation had the greatest validity. The relationship between the simulated office oral and MCQ test format was also studied. The results obtained from this investigation show a correlation coefficient of .09 between the multiple-choice test and the office oral. Correlation studies between the office oral and the other two oral examinations that were administered yielded a correlation coefficient of .19 between the office oral and the role-playing oral, and a correlation of .35 between the office oral and the formal oral. Although these correlations were not statistically

significant, the authors did suggest that the results point to some discriminative validity, in view of the fact that each of the oral examinations were designed a priori to assess specific aspects of competence not assessed by the other formats. And, concluded that the simulated office oral examination was a valid technique to assess a physician's ability to interact with patients, and was also representative of the type of patients encountered by family physicians in their daily practice.

Van Wart (1974) conducted an investigation of the formal oral examination which was administered by the Canadian College of Family Physicians in 1973. He reported that no objective study had been performed to assess the validity of this test format. However, examiner comments, on an informal and formal level gained by the use of questionnaires, made it clear that they considered the formal oral a valid test for assessing candidate ability to practice medicine. Candidates responding to a questionnaire concerning the validity of this test format revealed that ninety percent (90%) of the respondents thought the formal oral to be valid.

Van Wart also reported reliability studies that were conducted in 1975 by Dr. J. Corley on the formal oral examination. Dr. Corley analyzed inter-rater reliability by correlating scores given by examiners and observers of the same candidate at the same time. In addition, comparison was made of the amount of agreement between two examiners and the average scores given to all candidates who were examined by the same examiners.

Results showed that three examiner teams were extremely reliable, while five teams were reliable, three were marginally reliable, and

four teams were found to be unreliable. Peer validation of those same ten teams replicated the results as obtained through statistical analysis.

In summary, studies investigating the reliability and validity of simulations used in the oral component of medical specialty certification examinations were reviewed. The reliability of this assessment technique based on inter-rater agreement of the performance of a candidate, was found to be adequate.

Content validity of simulations met acceptable standards. Studies investigating the relationship between multiple choice examinations and the various oral simulation technique generally yielded low to moderate correlation coefficient. These correlations were interpreted as an indication that the two test formats probably assessed somewhat different aspects of a physician's competence.

GENERALIZATION FROM THE REVIEW OF THE LITERATURE

The following generalizations have emerged from a review of the literature as they relate to the use of simulations as an oral test method in medical specialty certification examinations:

1. Simulations were introduced as an oral test method as a possible solution for the lack of reliability and validity of the traditional oral examination.
2. The characteristics of simulations allow for the assessment of specific aspects of content and skills not assessed by other test methods.
3. Studies of reliability and validity of simulation show that acceptable levels of reliability and

content validity can be achieved in oral examinations through the use of structured simulations and proper examiner training.

4. Correlation studies between multiple-choice tests and various types of simulations reveal a low to moderate degree of relationship. A greater correlation is found between multiple-choice tests and oral simulations of cognitive content than between multiple-choice tests and simulations assessing affective content.
5. Correlations between MCQ and simulation orals have been interpreted both as evidence for the concurrent validity of MCQs as well as evidence that MCQs measure different competencies than measured by oral simulations.

CHAPTER III

DESIGN OF THE STUDY

INTRODUCTION

The Emergency Medicine Specialty Examination was developed by the Office of Medical Education Research and Development (OMERAD) at Michigan State University in close collaboration with and under contract to the College of Emergency Physicians (ACEP). This study is based on data that was collected when this newly developed specialty examination was Field Tested under a grant from HEW (1977).

This chapter outlines the research procedures used to determine:

- 1) if a relationship exists between multiple choice test format and simulated clinical encounter test format; and 2) whether this relationship holds for different levels of physician training and within medical content domains.

The following topics will be described in this chapter:

1. Nature of the population and sampling procedures.
2. A brief description of the procedure used to develop the ACEP specialty examinations.
3. A brief description of the test formats developed for the ACEP specialty examination.
4. A description of the Field Test.
5. Description of the treatment.
6. Design of the study.

THE NATURE OF THE POPULATION

The population from which the sample subjects were selected consisted of three different groups at various medical training levels.

These groups were:

- A. Physicians Eligible for Certification: This was a population of physicians (M.D.) who were eligible to participate in the Field Test by virtue of the length of time they have practiced emergency medicine, or physicians who were eligible by virtue of having completed a residency in emergency medicine.
- B. Residents: This population represented subjects who were in their second year of residency in emergency medicine.
- C. Students: This population consisted of subjects who were in their fourth year in medical school. These students were from the Colleges of Human Medicine and Osteopathic Medicine at Michigan State University.

SAMPLING PROCEDURES

The subjects chosen to participate in the Field Test of the Emergency Specialty Examination from the Physician Eligible Group were selected by peer nomination. To facilitate this process the American College of Emergency Physicians (ACEP) contacted the major ACEP organizations in the United States and in Canada. These organizations were requested to help in the peer nomination of candidates to be chosen to participate in the Field Test.

The criterion used for peer physician nomination was that a candidate possess high quality medical skills, and not necessarily be

articulate or popular. In addition, physicians had to meet the eligibility requirements established by the American Board of Emergency Medicine. Initially, this group was to consist of an equal number of physicians made eligible to participate by virtue of the length of time they had practiced emergency medicine, and of physicians eligible to participate by virtue of having completed a residency in emergency medicine. In spite of this effort to equalize the two subject groups, the actual sample population consisted of fewer physicians who were eligible due to the length of time they had practiced emergency medicine than those who were eligible because they completed a residency in emergency medicine.

After random samples were drawn by computer, fourteen (14) practice eligible physicians and twenty two (22) residency eligible physicians were able to attend the Field Test.

Subjects for the Resident Group were chosen in a process by which ACEP contacted Directors of emergency medicine residency programs. A request was made for the director's help in compiling a list of the program's second year residents. To arrive at a sample representative of the resident population, the directors of the residency programs were asked to rank the residents according to their overall clinical competence. An effort was made to select residents who were not necessarily rated as the best or the worst residents, but rather, rated somewhere in the middle. Also, at least one resident was chosen from each of the different programs. Finally a sample of thirty-six (36) Resident Group subjects were drawn from this list to form the study group.

Subjects from the Student Group were chosen from a pool of fourth year medical students. Clinical Community Coordinators were contacted by the Director of OMERAD informing them about the project. The importance of the project was agreed upon, and their cooperation was assured. All fourth year medical students were then personally contacted, informed about the project, and encouraged to participate in the Field Test. No probability sampling was done for this subgroup population. The student subjects participating in the Field Test consisted of twenty-two (22) fourth year medical students who were paid volunteers.

ORGANIZATION AND PROCEDURES USED TO DEVELOP THE ACEP SPECIALTY EXAMINATION

The organization and procedures used to develop the test items for the Emergency Medicine Specialty Examination are briefly described below.

Two major units were organized for the purpose of examination development. One unit consisted of ACEP members and the other unit was from OMERAD--faculty and staff. The ACEP unit was directed by Dr. Ronald Krome (M.D.), while the unit from OMERAD was directed by Dr. Jack L. Maatsch (Ph.D.), Principle Investigator.

An ACEP medical task force was then organized consisting of twenty-five (25) members divided into five committees. They were assigned the responsibility of identifying medical content domains which are essential and necessary to the practice of emergency medicine. Furthermore, additional task forces from ACEP were formed and given the task of developing items in the different test formats. A medical audit committee was created to then audit all medical content items generated.

The development unit from OMERAD was composed of individuals representing various scholarly disciplines (e.g. Psychology, Educational Psychology, Instructional Development, Nursing, etc.). These individuals were responsible for generating test development strategies and specific procedures from the items and scenarios that were submitted to them by the different medical task forces. The members of this OMERAD development team were assigned to work on specific test formats. Throughout the development of the examination OMERAD faculty worked in close collaboration with members of the ACEP team, who were assigned to the same test formats.

Finally, a Planning, Monitoring and Documentation (PMD) Group was established to oversee and evaluate the research and development activities. Throughout the period during which test items were developed, the assignments given to ACEP task force members were facilitated by a series of workshops, meetings and personal communication from various individuals of the OMERAD development unit.

TEST FORMATS DEVELOPED FOR THE EMERGENCY MEDICINE SPECIALTY EXAMINATION

The Emergency Medicine Specialty Examination is a criterion-referenced examination, that was developed to evaluate candidates on essential and frequently used knowledge in predetermined and specific content domains in emergency medicine. This evaluation is a multi-format examination that includes various testing procedures. Twenty-two medical content domains were identified by the medical task forces, with each then being represented in a specified percentage of the total examination. General guidelines were established for the

development of the test items, with specific guidelines established for each of the test formats used in the examination (Maatsch et. al., 1976).

Three test formats were developed for the Emergency Medicine Specialty Examination. These formats were:

1. A written objective test which included:

a) Multiple-choice test consisting of approximately 400 items developed by ACEP task force members. They were edited by members of OMERAD assigned to this format, and by the Lee Natress Corporation.

b) Pictorial multiple-choice items consisted of 140 items. Development of items for this format followed the same procedures as regular MCQ items. However, visual aids were incorporated within this format of MCQs. Final editing of these items was done by members of OMERAD assigned to this format.

2. Patient Management Problems (PMP): The ACEP medical task force assigned to this test format provided OMERAD members assigned to this format with scenarios, from which ten (10) PMPs were developed according to specific guidelines. The scenarios selected to this format had to be reviewed and approved by the American Board of Emergency Medicine. This format is not involved in the present study.

3. Simulated Clinical Encounters (SCE): Two types of clinical simulations representing the oral examination component were included within this test format. These clinical simulations were:

a) Simulated Patient Encounters (SPE): This format evaluated candidate performance while managing a single patient in the emergency department of a simulated hospital.

b) Simulated Situation Encounters (SSE): This format evaluates the simultaneous management of three patients in the emergency department of a simulated hospital.

Medical scenarios for both SCE formats were provided by the particular medical task force, and from which eight (8) SPEs and four (4) SSEs were developed. All the cases used in the SCE were reviewed and approved by the American Board of Emergency Medicine.

DESCRIPTION OF THE FIELD TEST

All the test items developed for the Emergency Medicine Specialty Examination were field tested over a period of three days. The purpose of the Field Test was to gather the data necessary to establish the validity and reliability of the test items.

The Field Test examination was administered to the ninety-four (94) candidates under the direction of J.L. Maatsch, Ph.D., Project Director. He was assisted by members of the development team from OMERAD, in close collaboration with ACEP members and staff and Chief Examiner, C. Roussi, M.D.

The candidates who participated in the three-day Field Test were divided into four groups of equal size. The groups were rotated every two hours to a different test session. The candidates were evaluated individually for a total period of twenty-four (24) hours on all items

included in the test library consisting of approximately 400 Multiple-Choice Questions, 140 Pictorial Multiple-Choice Questions, 10 Patient Management Problems, 8 Simulated Patient Encounters, and 4 Simulated Clinical Encounters.

EXAMINER ORIENTATION TO THE FIELD TEST

Prior to the Field Test, OMERAD organized a two-day orientation workshop for the 24 examiners administering the Simulated Clinical Encounters. The orientation session was deemed necessary because the use of SCEs as an oral examination format is relatively new and much more complex to administer than traditional oral examinations. The purpose of the workshop was threefold:

1. To familiarize all examiners with specific cases they had been assigned to with regard to content and visual aids.
2. To familiarize the examiners with the process of administering a Simulated Clinical Encounter.
3. To achieve standardization of examiner rating of candidate performance.

Three different sessions were convened to orient the examiners. The first session of the workshop examiners familiarized themselves to their specific cases with the help of OMERAD members assigned to the various teams. The second session of the workshop consisted of a brief training film, showing how to correctly administer Simulated Clinical Encounters. After viewing the film, examiners were paired and asked to administer this case to each other. During the third session of the workshop, the examiners practiced rating candidate

performance on Simulated Clinical Encounters using videotape examples. After the first rating, examiners were given verbal instructions and explanations of the rating form, and asked to rate the videotape examples a second time. The two sets of ratings were then compared for consistency and reliability.

Two studies were conducted during and after the Field Test to assess the reliability of examiner ratings of candidates performance on the Simulated Clinical Encounters. One study used verifiers during the examination, and the other study was performed after the administration of the Field Test and investigated the effect of a candidate's appearance on examiner ratings. The effect of candidate appearance on examiner ratings was tested by dividing the examiner group into two groups, one of which rated the performance of videotaped candidates with varying appearances, and another of which rated the same candidates on the basis of their voices and audible cues, but were not allowed to see the candidates. Results of the reliability studies are reported by Maatsch et. al., in a report to HEW. (1979 Progress Report on Grant No. HS 02038).

CANDIDATE ORIENTATION TO THE FIELD TEST

Prior to the administration of the Field Test Examination Dr. Maatsch and members of OMERAD development team presented a brief review of the test item development to the candidates participating in the examination. The presentation was assisted by G. Roussi, M.D. and other representatives from ACEP. Rules and procedures pertaining to the expected conduct of participants during the Field Test were also explained. Each different test format was reviewed. Special emphasis

was given to the Simulated Clinical Encounters. Since the SCEs were a new oral test format for most of the candidates, a film modeling the administration of example cases and instructions to candidates were presented. Administration of the two test formats were explained and discussed.

All candidates were debriefed in detail at the end of the Field Test and were asked to give their reactions and suggestions as to the effectiveness, efficiency and acceptability of each test format. Comments regarding the Field Test were obtained in a systematic discussion and from a detailed evaluation questionnaire.

To summarize, the Field Test of the Emergency Medicine Specialty Examination ran effectively and according to schedule for the duration of the Field Test. Despite complex and difficult time constraints every candidate completed the examination. All items in the test item library were field tested and subsequent data was collected for analysis. Initial results of the debriefing sessions indicated that candidates considered the use of the varied test formats to be an excellent method by which to demonstrate their knowledge and clinical competence of emergency medicine.

DESCRIPTION OF THE TREATMENT

This study used data that was collected during the administration of the Field Test of the Emergency Medicine Specialty Examination. The data was obtained from the written objective multiple-choice questions and the pictorial multiple choice questions, and from the Simulated Patient Encounters. Following is a brief description of the test formats and their administration:

Objective Written Examination consisted of two parts:

1. Multiple-Choice Question. This format used on most medical specialty examinations was administered to the candidates in four (4) two-hour sessions. At the start of each session, the candidates were given instructions pertaining to this test format, and sample questions reviewed. Candidates were requested to mark their answers on machine scorable answer sheets. The candidates were asked to review the questions at the completion of each session, and make written comments relating to unclear or questionable items. These comments were used as one of the means of identification of poor test items.
2. Pictorial Multiple-Choice Questions. This format was administered in two (2) two-hour sessions. It is a variation of the MCQ test format differing from the MCQ by adding a new dimension through the use of visuals as part of the questions, (e.g. x-ray, EKG, and pictures). The test booklet used in this format was organized in a manner so that the visual was on the left page while the MCQ item relating to the visual was placed on the opposite page, or on the right. Candidates were requested to examine and interpret a given visual, and to answer a series of MCQ's relating to the visual on machine scorable answer sheets.

Simulated Patient Encounters: This type of oral role-playing simulation involved the management of one patient. The candidate

verbally interacted with the examiner administering the SPE. The candidate in the role of an emergency physician, and the examiner in the role of the patient and other health providers as the need arose. In terms of design the SPE consisted of the following material that aided in the administration of the cases:

1. Data Panels for Each Case:
 - a) An outline of the Patient's Medical History.
 - b) The Patient's Physical Examination Panel.
 - c) The Patient Laboratory Data Panel.
2. Flow Panel for Each Case. For each patient, a flow panel was constructed that outlined the development of the case from the patient was admitted to the emergency department of a simulated hospital, to the point when the patient was discharged. The Flow Panel had two columns:
 - a) The Event Column: providing a list of actions taken by the candidate, and/or patient reactions during the course of the case.
 - b) The Information Column: listed all additional information pertaining to each specific event. Every information box related to a specific event was divided into two parts: an upper part providing relevant information when the candidate took an appropriate action, and a lower part providing information and contingencies whenever the candidate's actions were inappropriate.
3. Stimulus Material. All stimulus material required by the candidate to appropriately manage a case, e.g.

x-ray, laboratory test results, etc., were numbered in a sequence.

INSTRUCTIONS TO THE CANDIDATES TAKING SCES

Each candidate was provided with specific instructions pertaining to the Simulated Clinical Encounters prior to taking the oral examination. The candidates were required to perform the role of an emergency physician managing a patient in an emergency department of a simulated hospital.

At the start of each case, the candidate was provided with basic written information about the patient such as: clinical setting, initial patient-data and circumstances for admission. Also, included was any history and physical findings obtained prior to the time when the candidate assumed management of the patient.

From this point on, the candidate was required to take charge of the case as he or she would in a real-life setting. He or she was asked to keep in mind that the staff would not initiate any action unless specifically requested or directed by the candidate.

INSTRUCTIONS TO THE EXAMINERS

Examiners responsible for the administration of the Simulated Clinical Encounters were required to play three roles throughout the interaction. The first role was that of the patient, and was acted using the first person tense. The second role was as the administrator of the case who provided information or stimuli when appropriate and was requested by the candidate. Examiners were asked, thirdly, to role-play as other health care providers (colleague, nurse, etc.). Most of the interactions within the Simulated Patient Encounter were initiated by the candidate, except at the beginning of the case, or

when major events occurred during the development of the case. Throughout the interaction, examiners were requested to not provide the candidate with any verbal or motion cues.

ADMINISTRATION OF THE SIMULATED CLINICAL ENCOUNTERS

Candidates were administered the SPEs as part of the oral examination. The SCEs were administered to the candidates in a large ballroom with eight (8) different test stations. Each SPE lasted for fifteen minutes. At the end of each SPE, the candidate rotated to the next station.

RATING CANDIDATE PERFORMANCE

All of the rating forms used to rate candidate performance were based on an eight-point scale. Seven separate ratings of candidate's performance were on essential specific and general clinical skills, deemed essential for a physician specializing in Emergency Medicine.

In addition, the rating form included a behavior checklist section which listed the minimum and essential critical actions required of a candidate in order to satisfactorily manage a patient in each clinical case. This section of the rating form had two major objectives:

1. To aid the examiners in keeping track of the critical actions made by the candidates checking the appropriate Yes or No column.
2. To aid in standardizing any subsequent subjective ratings.

DESIGN OF THE STUDY

This study attempts to determine whether a relationship exists between two test formats within three levels of medical training and within two specific content domains. In order to investigate this relationship three groups representing three separate populations at different medical training levels were used:

1. Physician (P) $n = 36$
2. Residents (R) $n = 36$
3. Students (S) $n = 22$

The items used in this study represented two medical content domains which were designated by the medical task force. The two medical content domains are:

Cardio-Pulmonary (CP)

Skeletal-Trauma (ST)

The Multiple-Choice items used in this study, the combined items from the Multiple Choice Questions and Pictorial Multiple Choice Questions, representing the two medical content domains after item analysis was performed. The combined items of MCQ and PMCQ will be designated as MCQ.

Multiple Choice Questions

CP = $n = 105$

ST = $n = 94$

Total MCQ (CP + ST) = 199

For the Simulated Patient Encounter test format the following number of cases were used in the specified content domains for this study.

CP = 2 Cases

ST = 2 Cases

Total SPE = 4 Cases

ANALYSIS METHODS

To test Hypothesis 1 the Pearson Moment Correlation Coefficient will be used to determine whether a general relationship between the two test formats exist.

To test Hypotheses 2 and 3, correlations between the levels of medical training and between medical content domains will be compared and tested for statistical significance.

To test Hypothesis 4, an analysis of variance will be performed to compare the ability of the two test formats to discriminate the different levels of competence represented in the three groups of candidates.

CHAPTER IV

ANALYSIS OF RESULTS

This chapter presents an analysis of test data collected from a sample of ninety-four (94) subjects representing three levels of medical training. Each test format (MCQ and SCE) contained two medical content domains; a) Cardio-Pulmonary (CP); and b) Skeletal-Trauma (ST). Items for the MCQ format (n = 199) and the SCE format (n = 4) were assigned to one of the above medical content domains by a special task force of emergency physicians.

The following research hypotheses were tested:

1. Correlations between total scores derived from a multiple-choice test, ratings of performance in oral Simulated Clinical Encounters will be positive and significantly different from zero.
2. Groups with different levels of training will produce correlations between total scores derived from a multiple-choice test, and ratings of performance on oral Simulated Clinical Encounters that are significantly different from each other.
3. When Knowledge of two different medical content domains are sampled, correlations between scores derived from a multiple-choice test and ratings of performance on oral Simulated Clinical Encounters will be significantly different from each other.
4. Ratings of performance on oral Simulated Clinical Encounters will differentiate between the three levels of training significantly better than scores derived from a multiple-choice test; and this relationship will hold regardless of medical content domains.

The testing of these research hypotheses proceed as follows:

1. The translation of the research question into an appropriate set of statistical hypotheses.

2. Selection of a statistical test appropriate to the statistical hypothesis.
3. Preparation and analysis of data.
4. A decision relative to the null statistical hypothesis.

For example, the first research hypothesis states that correlations between total scores derived from an objective multiple-choice test, and ratings of performance on oral simulated clinical encounters will be positive and significantly different from zero. Linton and Gallo state:

"A correlation coefficient indicates the strength of a relationship. It does not, however, indicate whether the relationship obtained differs significantly from zero. For that reason, each correlation must be followed by a test of significance."¹

Therefore, two statistical hypotheses can be formulated as follows:

$$H_0: \rho_{MCQ, SCE} = 0$$

$$H_1: \rho_{MCQ, SCE} \neq 0$$

If the null hypothesis can be rejected through a statistical test of significance of the observed correlation, then evidence is provided by the sample for the research hypothesis (H_1). In other words, the sample observations would have produced a positive correlation that is not, in all probability due to chance combinations of scores on the two test formats.

The correlations between total scores on MCQ's and ratings from SCE's are shown in Table 1. A correlation of .67 (Grand Total) is

¹Linton, M. and Gallo, P.S., Jr.: The Practical Statistician: Simplified Handbook of Statistics, pp. 342. Brooks/Cole Publishing Company, 1975.

TABLE 1

A COMPARISON OF MEANS, STANDARD DEVIATION, RELIABILITY AND
CORRELATION COEFFICIENTS BETWEEN MULTIPLE-CHOICE TOTAL
SCORES AND SIMULATED CLINICAL ENCOUNTER TOTAL RATINGS
AMONG PHYSICIANS, RESIDENTS AND STUDENTS

POPULATION	N	$\bar{X}$		S.D.		RELIABILITY		CORRELA- TION
		MCQ	SCE	MCQ	SCE	MCQ	SCE	
Physicians	36	108.8	69.8	10.3	12.7	0.82	NA**	.25
Residents	36	102.2	66.8	8.3	11.5	0.67	NA**	.30
Students	22	81.9	41.2	12.6	12.4	0.82	NA**	.45*
GRAND TOTAL	94	103.8	67.69	14.5	15.86	0.89	.80***	.67*

* $p < .05$

** Not Available

*** A weighted average of inter-rater reliabilities of each of the four SCEs used in forming a total performance score (Maatsch et. al., 1979)

observed for the total MCQ and SCE scores of the total study. Because the total sample is 94 ($n=94$), a correlation greater than .205 is needed in order to determine statistical significance at the .05 level.² In other words if H_0 is correct, a correlation of .205 or larger could be expected to occur by chance less than five percent (5%) of the time in a sample the size employed. Since the observed correlation coefficient is greater than the required value, the null hypothesis is rejected. Therefore, evidence is provided for the research hypothesis that the correlation would be positive and significantly different from zero.

Correlations between scores on MCQ's and ratings on SCE's for the three levels of physician training are also shown in Table 1. A correlation of .25 is observed between the scores for the physician group, .30 for the resident group, and .45 is observed for the student group. Since a correlation of .325 is required to be significant at the .05 level when $n=36$ the correlations produced by the physician and residents are not considered significantly different from zero. However, the .05 level for the student group is .404 ($n=22$) and the observed correlation is .45. Thus, the observed correlation is significantly different from zero.

These within group correlations are generally lower (.25, .45) than the observed correlation when all three groups are combined (.67). This is due in part to the restricted spread of scores within groups having a common level of training. In Table 1, it can be seen that the standard deviations of scores within each group are somewhat smaller

²Adapted from J.T. Spence et. al.,: Elementary Statistics, 2nd. Ed. Appleton-Century-Crofts, Pub. 1968, pp. 236.

than the standard deviation of the total scores for all three groups combined. The effect of this reduced variance of scores on both test formats is to reduce the size of the correlation between the two sets of scores. Hypothesis 2, however, concerns the differences between observed correlations for these three groups treated separately.

Table 2 presents an analysis of statistical significance of the differences between the observed correlations for the three levels of training. Following methods specified in Glass and Stanley (1970), differences in correlations as large as those observed, could have occurred by chance in the sample, fifty two percent (52%) of the time for the physicians and residents (see values in Table 2). The other contrasts also do not reach the .05 level required to reject the null hypotheses. In other words, despite the observed differences in the sample there is no evidence that different levels of training produce significantly different correlations and the null research hypothesis 2 is not rejected.

The third hypothesis concerns the performance of the total population on two different domains of medical knowledge. Because these two domains, Cardio-Pulmonary (CP) and Skeletal-Trauma (ST) require quite different medical knowledge and skills, it was anticipated that differences in correlations between MCQs and SCEs might be observed.

Table 3 compares total group performance on the Cardio-Pulmonary (CP) and the Skeletal-Trauma (ST) domain subtests. Because the comparison involves the same population, a test of significance developed by Hunter (undated) is used.

It should be noted in Table 3 that the ST domain items produce less score variance on the multiple-choice test than the CP domain

TABLE 2

TEST OF STATISTICAL SIGNIFICANCE OF THE PAIRWISE
DIFFERENCES OBTAINED IN THE OBSERVED CORRELATIONS
AMONG PHYSICIANS, RESIDENTS, AND STUDENTS

POPULATION	OBSERVED DIFFERENCES	Z	P
Physician-Resident	.05	.2198	.52
Physician-Student	.20	.7962	.22
Residents-Student	.15	.6083	.27

TABLE 3

A TEST OF STATISTICAL SIGNIFICANCE BETWEEN OBSERVED
CORRELATIONS FOR THE TOTAL POPULATION IN THE
CARDIO-PULMONARY AND SKELETAL-TRAUMA CONTENT DOMAINS

DOMAIN	N	$\bar{X}$		S.D.		RELIABILITY		CORRELATION	Z
		MCQ	SCE	MCQ	SCE	MCQ	SCE		
CP	94	74.5	32.5	11.0	8.7	.86	.84 ³	.66	2.00*
ST	94	25.8	29.5	4.5	9.2	.72	.76 ³	.42	

* $p < .05$

³As in Table 1, weighted averages (Fischer, 1921), for the inter-rater reliabilities of SCEs are reported for each domain. More specifically, Maatsch et. al., (1979) using the same subjects report inter-rater reliabilities for the two CP problems as .83 and .84 and for the two ST problems as .79 and .72. Using Fischer $r \rightarrow z$ transformation, averages of .84 and .76 respectively are shown in Table 3. Because only two problems were used to obtain an average score for each domain, it was felt that inter-rater reliability of problem scores was the most stable estimate of the reliability for purposes of comparison of the two domains. However, these reliabilities over-estimate the reliability of a total score made up of two or more problems. This is due to the fact that there is a significant amount of case to case variability of scores, i.e., case specific variance, observed in candidate performance (Maatsch et. al., 1979). This case specific variance would lower reliabilities of the total score for several cases even though the inter-rater reliability of each individual problem score might be quite high.

items (S.D. 4.5 vs. 11.0) and lower reliabilities for both test formats. The relative lack of score variability and lower score reliability could partially explain the lower correlations observed for the ST domain test.

The question of whether the observed differences in correlations between test formats for the two different domains of knowledge holds across the three levels of training was explored. The correlations between scores on MCQs and ratings from SCEs in the Cardio-Pulmonary (CP) medical content domain are shown in Table 4.

The magnitude and the pattern of correlations for this domain are very similar to the correlations reported in Table 1 when scores of both domains are combined. The correlation between scores on MCQ's and ratings from SCE's in the Skeletal-Trauma (ST) medical content domain are shown in Table 5.

For this domain, a much different magnitude of correlations are observed. Only the Grand Total score correlation is statistically significant. Correlations within groups are virtually zero. Score variability and reliability of the multiple-choice battery are also much lower than observed for the CP domain (see Table 4).

The third test was performed to determine whether the observed correlations within each level of training in the two medical content domains were significantly different. Hunter's test of significance of two correlations involving the same group of subjects was again used. The results are shown in Table 6.

Results of this analysis show that only the resident group produced a statistically significant difference between the correlations of the two test formats observed in the two medical content

TABLE 4

A COMPARISON OF MEANS, STANDARD DEVIATION, RELIABILITY
AND CORRELATION COEFFICIENTS BETWEEN MULTIPLE-CHOICE SCORES
AND SIMULATED CLINICAL ENCOUNTER RATINGS AMONG PHYSICIANS,
RESIDENTS AND STUDENTS IN THE CARDIO-PULMONARY CONTENT DOMAIN

POPULATION	N	$\bar{X}$		S.D.		RELIABILITY		CORRELATION
		MCQ	SCE	MCQ	SCE	MCQ	SCE	
Physicians	36	80.2	35.8	8.4	9.4	.79	NA**	.26
Residents	36	76.3	36.0	7.0	4.9	.65	NA**	.41*
Students	22	61.5	21.0	11.4	8.4	.84	NA**	.54*
GRAND TOTAL	94	74.51	32.58	11.0	8.7	.86	.84	.66*

* $p < .05$

** Not Available

TABLE 5

A COMPARISON OF MEANS, STANDARD DEVIATIONS, RELIABILITY AND
CORRELATION COEFFICIENTS BETWEEN MULTIPLE-CHOICE SCORES
AND SIMULATED CLINICAL ENCOUNTER RATINGS AMONG PHYSICIANS,
RESIDENTS AND STUDENTS IN THE SKELETAL-TRAUMA CONTENT DOMAIN

POPULATION	N	$\bar{X}$		S.D.		RELIABILITY		CORRELATION
		MCQ	SCE	MCQ	SCE	MCQ	SCE	
Physicians	36	28.5	33.9	2.8	6.6	.47	NA**	.22
Residents	36	25.9	30.7	3.3	9.4	.50	NA**	-.13
Students	22	20.3	20.1	4.1	7.4	.56	NA**	.09
GRAND TOTAL	94	25.8	29.50	4.5	9.2	.72	.76	.42*

* $p < .05$

** Not Available

TABLE 6

TEST OF STATISTICAL SIGNIFICANCE OF THE DIFFERENCES OF THE
OBSERVED CORRELATIONS BETWEEN CONTENT DOMAINS AMONG
PHYSICIANS, RESIDENTS AND STUDENTS

POPULATION	CORRELATIONS		DIFF.	Z	P
	CP	ST			
Physicians	.26	.22	.04	.1935	.53
Residents	.41	-.13	.54	2.3066	.01
Students	.54	.09	.43	1.5299	.07

domains. The students approach but do not attain the criterion for statistical significance. The observed correlations for physicians are nearly identical.

In conclusion, the evidence related to hypothesis 3 can be summarized as follows:

- a) The two content domains produce significantly different correlations between the two test formats when the three levels of training are combined.
- b) When these correlations are broken down into different levels of training, however, the differences are due to differences in the resident group and to a lesser extent, the student group. Although, the student difference is not statistically significant at the .05 level, it is at the .07 level, which can be considered a possible trend.
- c) The physician group shows virtually no differences in performance on the two medical content domains.

The fourth hypothesis concerns which of the two test formats can best discriminate between the three groups representing different levels of medical training. Several analysis of variance were performed to investigate the above hypothesis. The analysis of variance for the Grand Total scores on MCQs is shown in Table 7. Results of the Scheffé post-hoc analysis show that the Grand Total scores on MCQs differentiate between all three groups at the .05 level.

TABLE 7

ANALYSIS OF VARIANCE OF MULTIPLE-CHOICE GRAND TOTAL SCORES

SOURCE	DF	SS	MS	F
Between Groups	2	10142.14	5071.07	48.32*
Within Groups	91	9549.81	104.94	

* $p < .05$

The significant F test statistic in Table 7 indicates that further investigation between the groups is warranted. Such an investigation was done using the Scheffé method of post-hoc comparisons. This procedure isolated which group differences contributed to the overall significant F test statistic. The results of this procedure produced the following results:

1. Significant difference at .05 level between the student and resident group.
2. Significant difference at .05 level between the student and physician group.
3. Significant difference at .05 level between the resident and physician group.

Analysis of variance was also performed to investigate whether the Grand Total scores on SCE differentiate between the three levels of training. The results are shown in Table 8. The Scheffé technique shows a discrimination between two of the three groups.

A comparison of the two post hoc analyses shows that the MCQ test differentiates between residents and physicians, but the SCE scores do not. The total score in this analysis combine scores from both content domains. As shown in Table 3, these domains produce somewhat different means and standard deviations, as well as correlations between the two test formats. It was found that the cardio-pulmonary domain produced correlations closely approximating those observed using the Grand Total scores but that the skeletal-trauma domain did not. Because of these differences in performance on the two content domains, separate analyses of variance on each set of domain scores were performed.

Table 9 and Table 10 show the results of an analysis of variance for MCQ and SCE in the cardio-pulmonary content domain.

Results of the analysis of the cardio-pulmonary domain scores show that both test formats fail to differentiate between residents and physicians but differentiate student from these two groups.

An analysis of variance was performed on the skeletal-trauma content domain for MCQ and SCE scores and is presented in Table 11 and Table 12.

The analysis of the skeletal-trauma domain scores produces the same finding as the analysis of the cardio-pulmonary domain scores. Only the student group scores are significantly different from the other two groups.

TABLE 8
ANALYSIS OF VARIANCE OF
SIMULATED CLINICAL ENCOUNTERS GRAND TOTAL RATINGS

SOURCE	DF	SS	MS	F
Between Groups	2	12549.18	6274.59	41.73*
Within Groups	91	13681.80	150.34	

* $p < .05$

The significant F test statistic in Table 8 indicates that further investigation between the groups is warranted. Such an investigation was done using the Scheffé method of post-hoc comparisons. This procedure isolates which group differences contribute to the overall significant F test statistic. The results of this procedure produced the following results:

1. Significant difference at .05 level between the student and resident group.
2. Significant difference at .05 level between the student and physician group.

TABLE 9
ANALYSIS OF VARIANCE OF MULTIPLE-CHOICE SCORES
IN THE CARDIO-PULMONARY CONTENT DOMAIN

SOURCE	DF	SS	MS	F
Between Groups	2	4991.79	2495.89	32.67*
Within Groups	91	6950.84	76.38	

* $p < .05$

The significant F test statistic in Table 9 indicates that further investigation between the groups is warranted. Such an investigation was done using the Scheffé method of post-hoc comparisons. This procedure isolates which group differences contribute to the overall significant F test statistic. The results of this procedure produced the following results:

1. Significant difference at .05 level between the student and resident groups.
2. Significant difference at .05 level between the student and physician groups.

TABLE 10

ANALYSIS OF VARIANCE OF SIMULATED CLINICAL ENCOUNTER RATINGS
IN THE CARDIO-PULMONARY CONTENT DOMAIN

SOURCE	DF	SS	MS	F
Between Groups	2	3732.22	1865.11	31.23*
Within Groups	91	5437.26	59.75	

* $p < .05$

The significant F test statistic in Table 10 indicates that further investigation between the groups is warranted. Such an investigation was done using the Scheffé method of post-hoc comparisons. This procedure isolates which group differences contribute to the overall significant F test statistic. The results of this procedure produced the following results:

1. Significant difference at .05 level between the student and resident groups.
2. Significant difference at .05 level between the student and physician groups.

TABLE 11
ANALYSIS OF VARIANCE OF MULTIPLE-CHOICE SCORES
IN THE SKELETAL-TRAUMA CONTENT DOMAIN

SOURCE	DF	SS	MS	F
Between Groups	2	915.39	457.69	40.32*
Within Groups	91	1032.81	11.34	

* $p < .05$

The significant F test statistic in Table 11 indicates that further investigation between the groups is warranted. Such an investigation was done using the Scheffé method of post hoc comparisons. This procedure isolates which group differences contribute to the overall significant F test statistic. The results of this procedure produced the following results:

1. Significant difference at .05 level between the student and resident group.
2. Significant difference at .05 level between the student and physician group.

TABLE 12
ANALYSIS OF VARIANCE OF SIMULATED CLINICAL ENCOUNTER
RATINGS IN THE SKELETAL-TRAUMA CONTENT DOMAIN

SOURCE	DF	SS	MS	F
Between Groups	2	2705.81	1352.59	19.86*
Within Groups	91	6196.31	68.09	

* $p < .05$

The significant F test statistic in Table 12 indicates that further investigation between the groups is warranted. Such an investigation was done using the Scheffé method of post-hoc comparisons. This procedure isolates which group differences contribute to the overall significant F test statistic. The results of this procedure produced the following results:

1. Significant difference at .05 level between the student and resident group.
2. Significant difference at .05 level between the student and physician group.

In conclusion, the results of the several analyses of variance and Scheffe post-hoc analyses, show that the MCQ format differentiate the three levels of training in the hypothesized direction, i.e., physicians highest, students lowest. SCEs can clearly differentiate the student group from the other two groups at the .05 level of significance. In specific content domains the student group performs significantly lower than the physician and resident groups on both formats.

Since our research hypothesis was that the SCE format will differentiate between the three levels of training better than the MCQ format, and that relationship will hold within medical content domains, we must accept the null hypothesis and conclude the one format seems to discriminate groups about as well as the other.

In summary, the first and third null hypotheses were rejected and the second and fourth null hypotheses are accepted. A significant positive correlation was observed between performance on the two test formats when total scores are utilized for the total sample. Groups with different levels of medical training did not produce significantly different correlations using Total Scores. However, when format correlations are compared between the two content domains subtests, their magnitude are significantly different. Further investigation shows that this difference is due primarily to the difference in the magnitude of correlations produced by the residents group and to a lesser extent students group. Test score correlations of physicians on the two content domains were virtually the same. Finally, it was observed that the Simulated Clinical Encounter format did not discriminate the three levels of training of the population sampled better than the multiple-choice format. In fact, they performed about equally well.

CHAPTER V

DISCUSSION AND SUMMARY

Throughout this study questions were raised relative to the concurrent validity of the different test formats used in medical specialty certification examinations to measure physician competence. In the review of the literature we cited studies by Levine and McGuire (1968, 1970) which measured the relationship between MCQ tests and oral simulation tests based on data that the authors obtained from certification examinations that were administered by the American Board of Orthopaedic Surgery. Their studies yielded correlation coefficients that ranged from .19 to .35. These correlations were interpreted by the authors as being in accord with the hypothesis that the MCQ format and oral simulation formats measured somewhat different physician competencies.

Lamont and Hennen (1972) also performed studies that investigated the relationship between measures of performance on the "office oral" which is a form of simulation and MCQ test scores. Their study produced a correlation of .09. They also interpreted this result to be in the expected direction and magnitude because the various test formats that were used in the Canadian College of Family Physicians certification examination were presumably designed to assess different physician competencies. Furthermore, they stated that when the "office oral" ratings were correlated with the two other types of oral simulations namely, a role-playing oral and a formal oral, correlation coefficients of .19 and .35 respectively, were observed. They argue that these low correlations further support their assumption that the different test formats indeed assessed different competencies.

Correlations obtained in this study confirm the results that were obtained in previous studies cited only in part. The low to moderate correlations are obtained only when MCQ scores and SCE ratings are correlated within a specific level of medical training and especially at the highest level of training, i.e., the group of physicians who were eligible for certification. This group corresponds in their medical training and practice experience to the population on which Levine and McGuire and Lamont and Hennen based their studies. However, it should be noted that the correlation coefficients that were obtained in the present study when the three levels of medical training are combined into one group are much higher than correlations reported by both Levine and McGuire. For example, we found a correlation of .67 for the grand total scores as compared to .09 for Lamont and Hennen and .19 to .35 for Levine and McGuire.

The differences in the magnitude between the correlations reported in the two previous studies cited and the correlations that were obtained in this study could be attributed partially to the different statistical factors that influence correlations. Specifically, the magnitude of correlations obtained are influenced by the variability of scores and reliability of the scores on each test format.

Maatsch et. al., (1978) argues that the two different test formats sample different aspects of the same general competency. Maatsch points out that when correlations are corrected for attenuation to account for lack of reliability on otherwise highly reliable measures, the estimate of the underlying true correlation is in the .90's. These results are based on a more complete set of data collected from the same Field Test of the ACEP Certification Examination. He argues that

this study would seem to demonstrate that the same general competency is being measured by the two test formats, and that the low correlations obtained by others are probably due to other statistical and sampling factors.

The results of this study seem to indicate that content domains may have an effect on the correlation coefficients that are obtained at the undergraduate and graduate levels of medical training. However, when correlations are compared between the two different content domains using the physician group who are at specialty certification level of medical training and experience, they are the same magnitude. The observed correlations in the CP content domain are very similar to the correlations that were obtained when the scores for the two content domains were combined (see Table 1). In contrast, the observed correlations in the ST content domain are very different in their magnitude from the correlations using the CP content domain across all levels of training. In fact, they are virtually zero.

Two plausible explanations can be given for these differences. First, results of the analysis show that there was more variability and higher reliability of the scores in the CP content domain than in the ST content domain. Second, the nature of the content domains that were used in this study could also account for some of the differences that were observed. The CP content domain is a well defined medical content domain, whereas, the ST content domain is much broader and harder to define. It encompasses most all medical content domains whenever complications occur as a result of trauma. A very broad knowledge of medicine as well as specific surgical procedures and skills are required for management in this domain. To reliably sample

knowledge and performance in such a broadly defined domain is much more difficult and in turn could produce lower correlations.

CONCLUSIONS

Given the limitations of this study, we conclude that the results that were obtained only partially support previous studies which measured the relationship between MCQ tests and oral simulation tests. The higher correlations that were obtained in this study could be attributed for the most part to the increased variance within the population sampled. It represented a much broader range of medical competence than represented in previous studies. Another factor which may have contributed to the higher correlations was the fact that both test formats attempted to sample the same two domains of knowledge. The previous studies sampled broadly across a large number of specific domains of knowledge within their respective medical specialties.

Care should be taken when interpreting the results of correlation studies. As we have indicated in the discussion above, the results of this study seem to indicate that MCQ and SCE test formats may measure different competencies at the graduate and undergraduate level of medical training on some specific domains of knowledge. For example, knowledge of the ST content domain as measured by MCQ test does not correlate highly with clinical performance as measured by the SCE test format. In other words, knowledge of and the ability to perform may constitute different competencies at these levels of medical training for specific domains. However, low to moderate correlations that were observed could have been as a result of a variety of reasons both statistical and substantive and not necessarily the fact that test

formats are measuring different competencies as other authors have proposed.

The purpose of certification examinations in a medical specialty is to assess not only the candidate's knowledge, but also to assess the ability of the candidate to apply that knowledge in a variety of clinical situations. Attempts to reliably and validly measure physician competence either during or at the completion of their medical training has been an objective pursued for many years by the various medical professional associations. To achieve this goal of finding better ways to assess physician performance, assessment techniques are needed that will assess or sample the different aspects of knowledge and performance required for physicians to provide health care. This study has demonstrated that both the MCQ test format and the SCE formats are valid techniques to assess physician knowledge and competence and both discriminate the different levels of competence represented in the three groups about equally. Therefore, these two test formats can be considered as complimentary to each other as measurement techniques to more validly assess physician's knowledge and performance. Where possible both formats should be used to more adequately measure physician competence.

RECOMMENDATIONS

From the preceeding discussion, a number of recommendations for future research in the assessment of physician competence are suggested. They are:

1. To initiate comparative studies between multiple-choice tests and Simulated Clinical Encounters within different medical content domains in order

to determine whether specific medical content domains show similar or different relationships than the relationships that were obtained in this study. The results of such a study may assist in arriving at a more generalizable theory concerning the use of rationally defined medical content domains in specialty certification examinations.

2. To assess the effectiveness of Simulated Clinical Encounters as a unique measure of the growth of clinical competence at the different levels of training in medical schools.
3. To investigate the predictive validity of both the multiple-choice and the Simulated Clinical Encounters tests by comparing scores from certification examinations on both these test formats with the performance of physicians in clinical settings in order to ascertain which of these two test formats best predict future performance of the physicians in real clinical settings.

SUMMARY

The aim of this study was to establish whether an empirical relationship between multiple-choice tests and Simulated Clinical Encounters holds across three levels of medical training and two medical content domains. The review of the literature demonstrated the continuous efforts of the medical specialty boards to develop better and more valid test formats to assess physician competence.

Through these efforts the Multiple-Choice test and the Patient Management Problem test format were introduced in the specialty certification process. However, the oral examination which was widely used early in the development of certification examinations remains a source of concern as to its reliability when compared to the more reliable MCQ test format. More recently there have also been questions raised as to the validity of both the oral and MCQ test formats. As an alternative to the traditional oral examination, a more standardized form of simulations were introduced in an effort to obtain a more reliable and valid performance examination.

The review of the literature related to studies which measured the relationship between the MCQ test format and oral simulation test format is very limited. These studies were based on the data made available to researchers from certification examinations that were administered by two specialty boards which have incorporated oral simulations in their certification examinations, namely, the American Board of Orthopaedic Surgery and The Canadian College of Family Physicians.

The investigation reported in this study was based upon data which was collected during the Field Test of a newly developed certification examination for the American College of Emergency Physicians. The results of this investigation revealed that correlations obtained in this study were higher than those previously reported. These differences in the magnitude of the correlations were explained as due in part to a variety of statistical factors which could influence the results of correlations obtained. Three such factors were identified:

- 1) greater variance of the population which was sampled in this study,
- 2) the reliability of the scores produced by the two test formats,

and 3) the items used in the present study were by and large designed to assess relevant clinical knowledge and specifically sampled two domains of medical knowledge.

This study demonstrates that different content domains may have an effect on the correlations that were observed for the residency levels of medical training and to some extent on the undergraduate level of medical training. However, there is no evidence that different domains effect the correlations observed at the certification level of competence.

In any event, this study provides evidence for the concurrent validity of both test formats. Concurrent validity of the MCQs is suggested by the relatively high correlations with performance on SCEs, but more importantly both test formats demonstrate their ability to discriminate with different levels of clinical competence. Both test formats are viewed as complimentary measurement techniques. It is suggested that both should be used, if possible, to obtain a more reliable and valid assessment of physician competence.

LIST OF REFERENCES

LIST OF REFERENCES

- Abrahamson, S. "Validation in Medical Education." Proceedings from a Conference on Extending the Validity of Certification, American Board of Medical Specialties, Pub. 1976, pp 15-17.
- Bull, G.M. "Examinations." Journal of Medical Education, Vol. 34, Dec. 1959, pp 1154-1158.
- Carter, H.D. "How Reliable are Good Oral Examinations." California Journal of Educational Research, Vol. XIII, No. 4, Sept. 1962, pp 147-153.
- Cowles, J.T. "Current Trends in Examination Procedures." Journal of American Medical Association, 155, 1954, pp 1383-1387.
- Downing, S.M. Multiple-Choice Item Writing Handbook. Office of Medical Education Research and Development, Michigan State University, 1977.
- Ebel, R.L. Essentials of Educational Measurement. Prentice-Hall, 1972.
- Foster, J.T., Abrahamson, S., Lass, S., Girard, R., and Garriss, R. "An Analysis of an Oral Examination used in Specialty Board Certification." Journal of Medical Education, Vol. 44, 1969, pp 951-954.
- Holden, W.D. "The Evolutionary Functions of American Medical Specialty Boards." Journal of Medical Education, Vol. 44, 1969, pp 819-829.
- Hubbard, J.P. Measuring Medical Education. Lea and Febiger, Pub. 1971.
- Hunter, J.E. "A Statistical Test for the Stability of a Correlation Coefficient." Department of Psychology, Michigan State University. Undated.
- Kelley, P.R., Mathews, J.H., and Schumacher, C.F. "Analysis of the Oral Examination of the American Board of Anesthesiology." Journal of Medical Education, Vol. 46, 1971, pp 982-988.
- Lamont, C.T. and Hennen, K.E. "The Use of Simulated Patients in a Certification Examination in Family Medicine." Journal of Medical Education, Vol. 47, 1972, pp 789-795.
- Levine, H.G. and McGuire, C.H. "Role-Playing as an Evaluative Technique." Journal of Educational Measurement, Vol. 5, No. 1, Spring, 1968, pp 1-8.

- Levine, H.G. and McGuire, C.H. "The Use of Role-Playing to Evaluate Affective Skills in Medicine." Journal of Medical Education, Vol. 45, 1970, pp 700-705.
- Levine, H.G. and McGuire, C.H. "The Validity and Reliability of Oral Examinations in Assessing Cognitive Skills in Medicine." Journal of Educational Measurement, Vol. 7, No. 2, Summer, 1970, pp 63-74.
- Linton, M. and Gallo, P. Jr. The Practical Statistician: Simplified Handbook on Statistics. Brooke/Cole Publishing Co., 1975.
- Maatsch, J.L. An Introduction to Patient Games: Some Fundamentals of Clinical Instruction. Biomedical Communication Center, Michigan State University, 1974.
- Maatsch, J.L. (Principle Investigator) Model for Criterion-Reference Medical Specialty Test. HEW Grant No. I R18 HS 02038, National Center Health Service Research, July, 1977.
- Maatsch, J.L. and Gordon, M.J. "Assessment Through Simulations." Chapter 10 in Evaluating Clinical Competence in the Health Professions. Edit. Morgan, M.K. and Irby, D.M. C.V. Mosby Co., 1978.
- Maatsch, J.L., Hoban, J.D., Sprafka, S.A., Hendershot, N.A. and Messick, J.R. A Study of Simulation Technology in Medical Education. Final Report to the National Library of Medicine, 1977.
- Maatsch, J.L., Holmes, T., Downing, S.M., Sprafka, S.A. "Towards a Testable Theory of Physician Competence: An Experimental Analysis of a Criterion-Referenced Specialty Certification Test Library." Symposium Presentation, RIME Conference, 1978.
- Maatsch, J.L., Krome, R.L, Sprafka, S.A. and Maclean, C.B. "The Emergency Medicine Specialty Certification Examination (EMSCE)." Journal of College of Emergency Physicians, July, 1976. Special Contribution.
- McCarthy, W.H. "An Assessment of the Influence of Cueing Items in Objective Examinations." Journal of Medical Education, Vol. 41, 1966, pp 263-266.
- McGuire, C.H. "The Oral Examination as a Measure of Professional Competence." Journal of Medical Education, Vol. 41, 1966, pp 267-274.
- McGuire, C.H., Solomon, L.M., and Bashook, P.L. Construction and Use of Written Simulations. The Psychological Corp., 1976.
- Pokorny, A.D. and Frazier, S.H. "An Evaluation of Oral Examinations." Journal of Medical Education, Vol. 41, 1966, pp 28-40.

Senior, J.R. Toward the Measurement of Competence in Medicine. Carnegie Corporation of New York and the Commonwealth Fund, 1976.

Tyler, R.W. Basic Principles of Curriculum and Instruction. University of Chicago Press, 1950.

Van Wort, A.D. "A Problem-Solving Oral Examination for Family Medicine." Journal of Medical Education, Vol. 49, 1974, pp 673-680.

Williamson, J.W. "Validation by Performance Measures." Proceedings from a Conference on Extending the Validity of Certification, American Board of Medical Specialties, 1976.

Evaluation in the Continuum of Medical Education. Report to the Committee on Goals and Priorities of the National Board of Medical Examiners, 1973.

Extending the Validity of Certification. Conference Proceedings, American Board of Medical Specialties, 1976.