



RETURNING MATERIALS:  
Place in book drop to  
remove this checkout from  
your record. FINES will  
be charged if book is  
returned after the date  
stamped below.

~~18 R109~~  
~~APR 19 84~~

~~MAY 11 84~~  
~~R123~~  
~~MAY 16 84~~

~~11 84~~  
~~184~~

Big Mick, Good Lookin'

Ben Mitchell ↑

Sarah Jane Mitchell, Susa girl

6-23-97

Michael J. Mitchell

Jane A. Mitchell

SSC

Ben Mitchell

Josh Mitchell

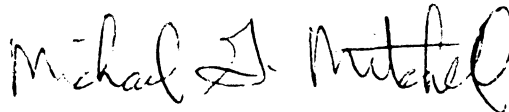
Michael Mitchell

The whole Mitchell Family  
We were all  
here 6-7-2004  
8:35 PM

EVALUATION THROUGH TELEVISION: A STUDY OF THE VALIDITY  
AND EFFECTIVENESS OF TELEVISION COMPARED TO LIVE  
OBSERVATION IN EVALUATING CANDIDATES'  
PERFORMANCE IN STRUCTURED ORAL  
CERTIFICATION EXAMINATIONS

By

Michael Glenn Mitchell

A handwritten signature in black ink, reading "Michael G. Mitchell". The signature is written in a cursive, flowing style with a large, prominent "M" and "G".

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Secondary Education  
and Curriculum

1979

## ABSTRACT

### EVALUATION THROUGH TELEVISION: A STUDY OF THE VALIDITY AND EFFECTIVENESS OF TELEVISION COMPARED TO LIVE OBSERVATION IN EVALUATING CANDIDATES' PERFORMANCE IN STRUCTURED ORAL CERTIFICATION EXAMINATIONS

By

Michael Glenn Mitchell

Many sectors of our American society have come under scrutiny in recent years. Demands for greater accountability have been placed on several of them. Medical specialty boards have felt the pressure to increase the validity and effectiveness of their certification examinations.

In an effort to more effectively predict physician competence and performance, medical specialties are turning to the use of structured oral examinations. The problem is, however, that oral examinations tend to be unreliable and difficult to administer.

Technology, in the form of television, offers potential to improve quality control of oral examinations. Traditionally, oral exams are evaluated by an examiner present in the room (live observation). Television offers an alternative form of observation that produces

a referable product--the videotape. The videotape presents many opportunities for quality control in the oral examination process. Television also offers possibilities for cost effectiveness.

This dissertation purposes to determine if there were meaningful differences in scores obtained from evaluating oral simulations through mediated versus live observation. This line of research was undertaken with the anticipation that evaluation through television might prove to be as valid an evaluation technique as live evaluation.

The experimental design was set up to compare the scores of examiners' rating from television versus a live situation. The data for this dissertation experiment were analyzed using power analysis. The results demonstrated that the alternative hypothesis could not be rejected. In other words, media and live observation cannot be considered the same. Given inconclusive results, one would ultimately need to replicate the study under more optimal conditions. The direction the follow-up study would take should be based on what the original data appeared to indicate. The data for this dissertation indicated that the estimated population size effects was very large, about 27 percent. Based upon these findings, another researcher conducted a

Michael Glenn Mitchell

follow-up study. The follow-up study compared examiners' rating alone to examiners rating in groups. Significant differences were found between the two treatments. The results of this follow-up study are useful in interpreting the findings from the dissertation experiment. It is reasonable to suspect that peer effects are confounded in the findings from the comparison of the examiner rating in the Field Test (live observation) to the group of examiners rating a candidate via television (post-Field Test Experiment). The computed size effects difference for the follow-up study was 29 percent. This finding suggests that the equivalence of the two treatments used in the dissertation experiment was much closer than what was found.

The implications from this dissertation are varied. First and foremost, if an investigator plans on using television in the same manner as was used in this dissertation, results obtained will probably be confounded with peer effects. If it is impossible to eliminate the peer effects, it is suggested that television be used in such a way that individual raters are isolated with the television set. An advantage of this arrangement is that it could allow for investigating the nature of the peer effect. Is it caused by proxemics, or other unknown factors? Television is one method

Michael Glenn Mitchell

that can be used in designing experiments to isolate the unknown nature of peer effects.

Apart from the above implications is the validity issue of the scores received from mediated versus live evaluation. The results of this dissertation experiment indicate that scores from media are more conservative than live evaluation. Which method is more valid? Laymen would probably argue for the media method. The medical profession would probably favor the live method. The answer is a judgment call and therefore is outside the intended parameters of this dissertation.

Two major conclusions arise from this dissertation: As contrasted in this study, (1) the two treatments, media and live, cannot be considered the same and (2) the differences between the two treatments are probably confounded with known effects.

## ACKNOWLEDGEMENTS

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

First and foremost my wife, Jane, who has loved and supported me throughout all my graduate studies.

Dr. William Jernigan, Dr. Carl Hamilton, and Oral Roberts University, for sponsoring my doctoral program.

Dr. Thomas F. Holmes, 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. Bruce Miles, for assuming chairmanship of the committee, and for providing gentle but firm guidance.

Dr. Kent Creswell, for his invaluable comments, critical review, and friendly encouragement.

Dr. Ben Bohnhorst, for his comments and warm encouragement.

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

*Your Beautiful  
Wife  
Jane  
2004*



My loving family, for their prayers and support.

These persons 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 Medicine 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, I am indeed grateful.

I count myself fortunate. My sincere thanks to all.

## TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	vi
LIST OF FIGURES . . . . .	vii
 Chapter	
I. THE PROBLEM . . . . .	1
Introduction . . . . .	1
The Problem . . . . .	1
Training and Certifying Physicians . . . . .	3
Historical Background of Specialty Boards . . . . .	5
Certification Examinations . . . . .	6
Oral Examinations . . . . .	7
The Emergency Medicine Specialty Certifi- cation Examination (EMSCE) . . . . .	9
Purpose of this Dissertation . . . . .	10
Hypotheses . . . . .	11
Definition of Terms . . . . .	13
Summary and Overview . . . . .	14
II. REVIEW OF THE LITERATURE . . . . .	16
Introduction . . . . .	16
Oral Assessment Techniques . . . . .	17
Use of Simulation in Medical Specialty Certification Examinations . . . . .	21
Observational Evaluation . . . . .	25
Observer Perception . . . . .	31
Summary . . . . .	34
III. DESIGN OF THE STUDY . . . . .	36
Introduction . . . . .	36
The Emergency Medicine Specialty Certification Examination Field Test . . . . .	37
The Post-Field Test Experiment . . . . .	38
Experimental Facilities . . . . .	39
Television Production . . . . .	40
Instrumentation . . . . .	43

Chapter	Page
Power Analysis . . . . .	44
Design of the Study . . . . .	49
Hypotheses Tested . . . . .	51
Summary . . . . .	52
IV. ANALYSIS OF DATA . . . . .	54
Introduction . . . . .	54
Hypotheses . . . . .	55
Design of Experiment . . . . .	55
Discussion . . . . .	57
Summary . . . . .	59
V. SUMMARY, DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS . . . . .	60
Introduction . . . . .	60
The Problem . . . . .	60
The Literature . . . . .	62
Design . . . . .	63
Findings . . . . .	64
Discussion . . . . .	65
Indications of the Data . . . . .	66
Implications . . . . .	68
Conclusions . . . . .	69
Recommendations . . . . .	69
LIST OF REFERENCES . . . . .	71

## LIST OF TABLES

Table	Page
3.1. Design of Post-Field Test Experiment . .	50
3.2. Design of Dissertation Experiment . . .	50
4.1. Design of Dissertation Experiment . . .	56
4.2. Design for Analysis of Dissertation Experiment with Accompanying Values . .	56
4.3. Myers (1979) Critical Non-Central F-Values; Effect size: $p^2 = .06$ ; <u>Beta</u> = .95 . . . . .	58

## LIST OF FIGURES

Figure	Page
3.1. Floor Plan of Ballroom . . . . .	40
3.2. Floor Plan for Individual Examination Booth . . . . .	41
3.3. Floor Plan of Post-Field Test Experiment .	41
3.4. The Relationship of Power I and Power II (Myers, 1979) . . . . .	46

## CHAPTER I

### THE PROBLEM

#### Introduction

Recent years have seen a movement in this country towards greater accountability in all sectors of our society. One sector that has been seriously impacted by this movement is the health care system. The costs associated with health care delivery in America have grown tremendously in comparison to such other factors as personal income. A part of this phenomenal growth in costs can be attributed to the rising incidence of malpractice suits. Insurance rates for doctors have increased tenfold in the last 15 years (Burton, 1977). This situation has placed tremendous pressures on the medical specialty certification agencies in the country. Now, more than ever before, valid methods for evaluating competence and predicting physician performance are needed.

#### The Problem

In an effort to more effectively predict physician competence and performance, medical specialties are expanding the testing procedures to include structured

oral examinations. McGuire (1966) has found that properly designed oral examinations can more rapidly assess the clinical skills of a physician than traditional test formats (e.g., multiple-choice questions). However, problems are associated with oral examinations. They tend to lack reliability and are difficult to administer (Hubbard, 1971).

Technology, in the form of television, offers potential to improve the quality control of oral examinations. Television has several inherent characteristics that can provide quality control to the oral examination process. First, and most important, television can provide a referable product--a high-fidelity recording of the oral examination/simulation (Salomon, 1974; Gibson, 1947). During a normal live oral examination, the rater has to interact with the candidate, at times administer the examinations, while at the same time evaluate the performance demonstrated. By using television, however, the rater could concentrate on administering and/or interacting with the test candidate. Afterwards, his attention can be focused on evaluating the simulation by viewing the videotape. The use of multiple raters (to increase reliability) can be provided because television provides a referable product. Thus television could allow valid post hoc evaluation opportunities. Television

also provides a basis for accountability if judicial contestment were to occur. Fortunately it is also an inexpensive medium that can be utilized by nontechnical laymen.

Another aspect on the use of television is that of cost effectiveness. Television provides a basis for more flexible scheduling and evaluation of oral examinations. This could mean more efficient use of physician time. Prior to examination administrations, videotapes of varying candidate performance could be presented to train raters for the oral simulations. Television could also be used by a person outside the testing agency to critique the examination itself.

In order to provide the reader with more background, a more detailed discussion of the present process of medical certification is appropriate.

#### Training and Certifying Physicians

The training of physicians in the United States has traditionally followed a long-established process: (1) three or four years of medical school, (2) a state licensing examination, (3) a 1-6-year residency program, and (4) a medical specialty examination. The use of the medical specialty examination, imposed by the medical profession to upgrade the quality of medical care, is a procedure where by a physician is examined by his peers



to determine if he has attained a prescribed level of professional competence.

The trend toward voluntary certification has grown in recent years. However, growing doubts have surfaced in both the public and the profession about the examination procedures. Do they indeed measure physician competence? Senior (1976) states:

When for example, medical audit reveals discrepancies between actual performance and the expectations of performance ability based on documents of certification, doubt is cast on the certification procedure (p. 18).

Certification examinations are supposed to measure competency as defined by the National Board of Medical Examiners (1974):

The ability and/or qualities for patient care, diagnosis, treatment and management as distinguished from theoretical or experimental knowledge. Clinical competence includes such elements as skill in obtaining information from patients, and the ability to detect and interpret symptoms and abnormal signs acumen in arriving at a reasonable diagnosis and judgment in the management of patients (p. 23).

Certification examinations are defined as extramural examinations that are taken at the completion of an approved residency program. These examinations are developed by an agency outside the educational system, and evaluate the capability of a physician to perform at subsequent levels of training after licensure. Licensure examinations, in contrast, are intramural

examinations that assess achievement in educational institutions, such achievement being recognized by the bestowal of the medical degree and a license from the State granting permission to the candidate to practice medicine (NBME, 1973).

### Historical Background of Specialty Boards

The standards of performance for the purpose of certification are established by the American Board of Medical Specialties in conjunction with the particular specialty. This body was previously known as the Advisory Board of Medical Specialties and was established in 1933. The Board was reorganized in 1970 into a loose federation of five specialties with the responsibility for approving new specialties. The organization of the American Board of Medical Specialties now includes representatives from 20 approved specialty boards. Holden (1969) stated that 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 his competence to practice the specialty.

For a candidate to be granted permission to take certification examinations, approximately half of the boards require one to two years of practice in the

specialty in addition to completion of graduate work in an approved residency program.

### Certification Examinations

Hubbard (1971) declares that test formats used by specialty boards are as variable as the boards themselves. The predominant custom has been a combination of written and oral "practical" examinations. To date, the written test format is more developed and reliable than the oral form.

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" (p. 33). If we accept this principle, then the multiple-choice test format that 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 examination procedures are written and in the objective format, unfortunately, while these examinations have extremely high reliability and objectivity, there are serious questions concerning their validity. Somehow or other, the use of objective type examinations has to be justified through demonstrating that the

results of such examination procedures are significantly and closely related to the competency to practice specialty medicine (p. 11).

### Oral Examinations

In the past, a physician candidate was tested by oral questions as well as by live observation of his performance in history taking, physical examination, diagnosis, and management of a real patient. However, live observation of the candidate's performance with patients was discontinued in 1963 for a variety of reasons, among them the lack of reliability, cost, and problems with logistics (Hubbard, 1971).

In 1962, extensive studies and revisions of the oral testing format were initiated by the American Board of Orthopedic Surgery and the Center for the Study of Medical Education of the University of Illinois. The Oral Role Playing Simulation was introduced as part of the oral examination. The advantage that oral simulation has over the question-and-answer method and/or live observation in a natural clinical setting is that it allows for greater control over the presentation of stimulus materials. Therefore, it allows for greater precision in judging the particular aspects of clinical competence possessed by the examinee (Gordon, 1978).

On the other hand, traditional oral examinations have also been faulted for their lack of objectivity and

reliability. 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 the reliability of the oral examination is studied, it almost invariably fails to equal the reliability that can be demonstrated for good multiple-choice examinations (p. 46).

In a recent study that examined the validity of certification procedures, Williamson (1976) found that when certification scores were compared to actual performance measures there seemed to be little relationship between grades in certification examinations and the quality of subsequent performance. He 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 (p. 30).

Simulations, when properly designed, offer the possibility of assessing the clinical skills that a competent physician must have. As stated by McGuire (1966):

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 the promise of extending the now limited usefulness of oral examinations (p. 269).

Studies cited in the preceding discussion indicate the need to improve the certification examination procedures in medical specialties. To date, multiple-choice examinations have not proven to be effective in predicting future performance. Simulations, however, even though more costly and time consuming, are destined to play a much more significant role in certification examinations.

The Emergency Medicine Specialty  
Certification Examination  
(EMSCE)

In response to the need for more predictive certification examinations, the American College of Emergency Physicians (ACEP) entered into a contractual agreement in late 1974 with the Office of Medical Education Research and Development (OMERAD) at Michigan State University. OMERAD agreed to develop a multiformat criterion-referenced certification examination in Emergency Medicine for ACEP. The proposed examination would differ from conventional norm-referenced examinations (used by many other specialties) in its extensive use of structured oral simulations to assess the application of clinical

knowledge. From March 1975 to July 1977, an interdisciplinary team of OMERAD faculty worked closely with task forces of emergency physicians to develop the certification exam. The author of this dissertation was a graduate assistant in OMERAD, assigned to the ACEP project. His role was to aid in the development of some oral simulations as well as coordinate the media efforts during the developmental phase and finally the Field Test of the EMSCE.

In October 1977, the Field Test of the EMSCE took place in Lansing, Michigan. The purpose of the Field Test was to validate the entire test item library using 94 subjects from throughout the United States. Data for this dissertation were collected during both the Field Test and a post-Field Test experiment of the EMSCE. A description of the experiment can be found in Chapter III.

#### Purpose of this Dissertation

The reader will recall that in the statement of the problem earlier in this chapter, television was reported to hold potential for increasing the quality control and cost effectiveness of oral examinations. This is based on the assumption that television and live observation are equivalent forms. A large body of literature from instructional research supports this assumption (Chu and Schramm, 1968; Dubin and Hedley, 1969). However, whether or not these observational methods are equivalent

for purpose of certification remains to be empirically verified.

The purpose of this dissertation is to determine if there will be meaningful differences in scores obtained from evaluating structured oral simulations through mediated versus live observation. The Field Test of the Emergency Medicine Specialty Certification Examination (EMSCE) provided the basis for the experiment in this dissertation. An additional objective of this study is to explore the usefulness of using power analysis to test for meaningful differences. A discussion on power analysis and its application can be found in Chapter III. A final objective is to suggest procedures to implement the use of television technology in conducting and evaluating oral simulations.

### Hypotheses

Before the hypotheses are stated, a discussion of the conceptual framework for labeling and using the hypotheses in this dissertation will be presented. This will aid the reader as he progresses through the remainder of this study. Henkel (1976) states:

The term "null hypothesis" is often confusing, as it is used in at least three distinguishable, but overlapping, senses. In the sense that Fisher introduced it, "null hypothesis" meant a hypothesis complementary to (or the negation of) a research hypothesis that one believed to be true. One set up the null hypothesis



specifically to be rejected, or "nullified," so that its complement, the research hypothesis, could be considered to be true. A second sense in which the term null is used is to specify a parameter of zero. The third sense in which null is used is an outgrowth of the decision theory approach to significance testing which uses the term to specify the hypothesis which is being tested, that is, the hypothesis on which the sampling distribution is based. In the decision theory approach, one is deciding between two statistical hypotheses, one of which defines the sampling distribution and is specified as the null (p. 36).

Multiple meanings for the term null hypothesis could result in two contrasted hypotheses, both being labeled the null for different reasons. To avoid this confusing possibility in this dissertation the author limits the term null hypothesis to specify a parameter of 0, or equivalent to no difference. Conversely, the alternative hypothesis specifies a difference. (Chapter III will demonstrate that for power analysis, the magnitude of the difference must also be specified.)

The reader should note that the above definition of null is independent of the notion of rejection. This being so, either  $H_0$  or  $H_1$  could conceptually be rejected thus by indirect inference providing support for the remaining hypothesis. The hypothesis to be supported is considered and labeled the research hypothesis. This kind of labeling is consistent with Myers' (1979) system of labeling which is the technology used in this dissertation. Using this conceptual framework, the research

hypothesis for this dissertation will be the null hypothesis ( $H_0$ ). Therefore, the sampling distribution will be used with the alternative hypothesis ( $H_1$ ) to test it for possible rejection. Based upon this discussion, the following hypotheses will be studied:

Null Hypothesis ( $H_0$ )

The explained variance due to type of observation will be equivalent to no difference ( $.06p^2$  or less).

Alternative Hypothesis ( $H_1$ )

The explained variance due to type of observation will be greater than  $.06p^2$ .\*

Definition of Terms

The following definitions are provided to clarify the important words and terms used in this dissertation.

Live Observation: The actual physical presence of the rater/examiner in the same room as the candidate being evaluated.

Television Observation: The use of a television camera to record and play back the performance of a candidate in an oral simulation. The camera is placed in a position that will duplicate, as closely as possible, the

---

\*The value of  $.06p^2$  in the alternative hypothesis can be interpreted as: Do media distort the normal live estimate of the candidate's ability more than 6 percent? If it does not, the author is willing to accept media as equivalent.

view from the same perspective as that of the rater/examiner.

Simulated Clinical Encounter (SCE): A carefully planned simulation of a real patient(s) case that 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 have been ordered. The candidate plays the role of the patient and other health providers as the need arises, in addition to that of actual examiner (Maatsch et al., 1978).

#### Summary and Overview

This chapter reviewed the need for greater accountability in certifying physicians to practice medicine in this country. A brief discussion on certification examinations and the background of American specialty boards was presented. The problems associated with the use of multiple-choice and oral examinations were detailed. The development and Field Test of the

Emergency Medicine Specialty Certification Examination were discussed and the application of this dissertation to the EMSCE was described. Finally, the purpose of this dissertation, the hypotheses to be tested, and the definition of terms were presented.

Chapter II will present a review of the literature pertaining to oral examinations and observational evaluation techniques. Chapter III will describe the design of the experiment for this dissertation. An analysis of the data will be described in Chapter IV. Chapter V will provide a summary and a discussion of the results, conclusions, implications, and recommendations of this dissertation.

## CHAPTER II

### REVIEW OF THE LITERATURE

#### Introduction

The focus of this study is the comparison of two different methods of observation to evaluate the oral simulation format of a medical specialty certification examination.

Reviewed first in this chapter are oral assessment techniques used in medical specialty certification examinations including studies of various oral certification examinations. This review indicates that when oral examinations were properly structured and controlled, the results were much more reliable.

Second, literature on the subsequent use of Simulated Clinical Encounters (SCE's) in medical specialty certification examinations is reviewed. This literature indicates that SCE's represent the most recent evolution in the trend to assess aspects of competence not assessed by other written examination forms. This section also shows that use of SCE's offers the potential to increase both reliability and validity of oral examinations.

Finally, literature on the use of observational evaluation is reviewed. The focus here is on the use of television observation as an alternative to live observation.

Although the majority of the literature on television is from an instructional perspective, Salomon's analysis provides the link to use of this information for evaluative purposes (Salomon, 1974). He states the major role of television in an instructional or evaluative mode is that of transmitting information. Therefore, the findings from the literature can be applied to either modality. The review of this literature indicates that for evaluative purposes, no difference generally exists between the two forms of observation.

#### Oral Assessment Techniques

The oral examinations used in the medical certification examination process were always considered to be, and still are by many, the final and most valid step of the certification process (McGuire, 1966).

Originally, oral examinations were conducted in the bedside oral form. Candidates were requested to perform a bedside examination of one or two patients. The candidates were required to report their findings, interpret laboratory and radiological data, and provide

a management plan to solve the patient's problem. They were then interviewed by two or three observing examiners.

With time, two major problems became evident. First, the very low inter-rater reliability of the format (Hubbard, 1971; Senior, 1976). Senior explains the reason for the low reliability of this format when he states:

The score awarded was 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 (p. 21).

In addition to the low reliability, the growing number of candidates created logistical problems in finding patients to tolerate repeated examination, and resulted in a financial burden for both the examiners and the candidates. Thus, the bedside oral examination format was discontinued, and replaced by the traditional oral examination format similar to that used in universities for doctoral candidates. This format has continued until recently to be a dominant feature of the oral component of the certification examination despite its more obvious drawbacks, especially that of reliability.

By the early 1960's concerted efforts were made by researchers in a number of specialty boards who realized that the written component (i.e., multiple-choice

questions and patient management problems) of the certification process was well established as a reliable testing format and focused their efforts on the oral examination format.

Carter (1962) investigated the assumption that oral examinations were unreliable, and used data made available to him by the Directors of the American Board of Anesthesiology. The best method of computing reliabilities is to use the interclass correlation coefficient. Applying this process, Carter found the data produced correlation coefficients of .62 between two examiners who scored the candidate in the same room, with a reliability coefficient of .89 for six examiners working in three teams. Based on the analysis of the data, Carter concluded that when oral examinations are systematically and carefully conducted, they can be more reliable.

McGuire (1966) conducted observational studies of oral examinations for a major specialty board. Results of this study showed that the traditional oral interview examination, (a) predominantly assessed the ability to recall factual knowledge rapidly, (b) that candidates rarely cited evidence to support their answers, and (c) the standards under which examiners conducted and judged candidates' performance were not clear, nor uniformly applied.



Foster, et al., (1969) analyzed data from a major specialty certification examination. Results showed differences among examiners when they scored candidates; however, the nature of these differences could not be identified. Also, no significant differences were found with regard to the level of difficulty among the different case studies used in the oral examination. Examiners consistently scored candidates either high or low, and consistently asked specific types of questions.

Kelly, et al., (1971) conducted an analysis of the oral examination for the American Board of Anesthesiology. The data yielded interclass correlation coefficients that ranged from .69 to .80 based on examiner agreement in scoring candidates in a single session. Correlation studies were performed between two oral sections, and yielded coefficients in the low 60s. Based on the results, the authors of this study concluded that structured oral examinations were more reliable than traditional oral examinations.

In summary, medical specialty boards, since the establishment of the first specialty board in 1917, have continuously striven to develop more reliable, valid, and practical testing formats designed to assess the competence of physicians in medical specialty certification examinations. However, the oral examinations in

the interview format continued to create concern because of lack of reliability. By 1960, medical specialty boards started to concentrate their efforts on the oral component of the certification process.

The studies that were undertaken by the various boards of the oral examinations produced two important findings. First, if the oral examinations were given in a structured and systematic manner, more acceptable inter-rater reliability coefficients could be achieved than in the traditional interview format. Second, the oral examinations in their present format mainly assessed factual recall.

Use of Simulation in Medical  
Specialty Certification  
Examinations

In the last decade, simulation technology has been increasingly used to develop and assess aspects of clinical competency not assessed by written formats. Maatsch and Gordon (1978) state the following advantages of simulations when compared to multiple-choice questions:

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 recognize factual information or 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 a multiple-choice test, will correlate highly

with his ability to apply the knowledge and other skills appropriately in a clinical situation (p. 173).

To date, at least two medical specialty boards use some form of Simulated Clinical Encounter for their oral certification examinations: The American Board of Orthopedic Surgery, and The Canadian College of Family Physicians.

Levine and McGuire (1966, 1970) describe three types of Role Playing Simulation (SCE's) developed for, and used by the American Board of Orthopedic Surgery.

1. The Simulated Diagnostic Interview:

In this simulation format the candidate plays the role of the physician, and the examiner plays the role of the patient while a second examiner rates the candidate's performance. The purpose of this simulation is to assess the candidate's ability to obtain from the "patient" a history, physical, and laboratory data. The candidate is required to explain and support his diagnostic procedure.

2. The Simulated Proposed Treatment Interview:

In this simulation format the candidate plays the role of the physician. The objective is for the candidate to gain patient compliance for the proposed treatment plan. The treatment plan is based on the medical problems presented to the candidate by a specific

patient. The role of the patient is played by an examiner, while the candidate is rated on his performance by a second examiner.

3. The Simulated Patient Management Conference:

This format is based on the leaderless group discussion. Its purpose is to simulate a staff conference. Five candidates are provided with basic information relating to two clinical problems. In an allotted period of time the candidates are required to discuss the management of each clinical problem, and arrive at an agreed course of action. The purpose of this format is to assess the minimal acceptable level of competence, rather than leadership qualities.

In all three formats described, candidates' performance is judged on the following competencies:

- A. Recall of factual knowledge.
- B. Analysis and interpretation of data.
- C. Their problem-solving abilities: clinical judgment.
- D. Do they relate effectively: show desirable attitudes?

These different aspects of competency are weighed differently. Maximum weight is allotted to that aspect of competency for which the specific simulation was developed.

The Canadian College of Family Physicians is another medical specialty that has introduced the use of simulation for the oral component of the certification process. Lamont and Hennen (1972) and Van Wart (1974) described the two types of simulation formats used in the oral certification examination component by the Canadian College of Family Physicians.

1. The Simulated Office Oral

In this format the patients are simulated by actors, who were programmed to simulate a specific problem. Three "patients" are interviewed by each candidate, one of whom requires a partial physical examination. The candidate's performance is rated on the following competencies:

- A. Desirable attitudinal skills.
- B. Skill in problem solving.
- C. Skill in allaying patient anxiety.

2. The Formal Oral

In a structured-problem oral examination, the candidate plays the role of the physician, and the examiner provides appropriate information to the candidate as he requests it. The information provided is from a preselected case. The candidate is expected to gather pertinent data and make acceptable patient management decisions. The candidate is scored on his ability to

relate to patients and overall coordination of clinical skills.

In summary, the continuous efforts to provide more reliability and face validity in oral testing methods has led to the development and use of different types of simulations as evaluative techniques. SCE's represent the most recent evolution in this trend.

Through the use of simulations, two important features were introduced in oral examinations that would make them more reliable and valid. They are:

1. The standardization of the oral examination (reliability).
2. The assessment of specific aspects of competency not assessed by other testing methods, but which are necessary for the assessment of the overall competence of a physician in a medical specialty (face validity).

### Observational Evaluation

In this dissertation, television observation was compared to live observation for evaluating a candidate's performance on a Simulated Clinical Encounter (SCE). Television was selected because it could simulate as closely as possible the actual activity of evaluation presently used by examiners in the oral simulation format of the Emergency Medical Specialty Certification Examination (EMSCE). The questions then arise, how equivalent are live and television observation, and

what can be expected when using television in place of live observation? The following review is directed at these questions.

In the world around us today are many examples of the use of television in place of live observation. Such activities as space shots, where men in underground control rooms watch the rocket via television, security surveillance in banks, stores, federal buildings, and even drive-up tellers are all examples of television used as an alternative to live observation.

Television can also be used in training and evaluation. For instance, the Mutual of Omaha Insurance Company uses videotape replays of simulated client interviews to train and evaluate new salesmen. Extensive use of videotape is also made in the physical education field, especially to evaluate the performance of individuals.

In all of these applications, as well as use in this dissertation, television is being used in what Salomon (1974) refers to as the "transmitting function of the medium."

The first kind of media usage and the most common one associated with the mass media (such as television) emphasizes their transmission quality. It is obvious that certain media bridge distances over space by their distribution or dispersion qualities and, over time, by their record-keeping qualities. Very often these functions of media are their prime

justification for use in instruction or evaluation. The "front-row view" approach assigns media only a transmitting function. The medium involved is taken to be but an envelope and as such it is not expected to have an influence on the transmitted messages (p. 399).

The majority of the literature on the use of television is from an instructional perspective. However, the major role television operates in when used in either an instructional or an evaluative mode is that of transmitting information. Therefore, much of the literature on the use of television can be applied to its use in an evaluative mode. Gordon Liefer (1976) states that on the average, television and film will impart information as well as the average live teacher does. As shown below, innumerable studies support this assertion.

In 1968, Chu and Schramm reviewed more than 400 comparisons of live and television teaching at all grade levels. In 1969, Dubin and Hedley reviewed nearly 400 comparisons of live and television teaching at the college level, including many of those reviewed by Chu and Schramm. The evidence clearly supports the conclusion that television and film are as effective at imparting all kinds of information to students as are live teachers (Chu and Schramm, 1968; Dubin and Hedley, 1979).

In a major government study sponsored by the U.S. Office of Education, Travers (1967) found that the



quality of information transmitted by television was as reliable and valid a method as using a live person. Based on this finding, he recommends the use of television due to the inherent characteristics of quality control and availability. Holmes (1959) analyzed television research and found that in almost 90 percent of the comparisons there were no substantial differences in achievement or information gain over conventional instruction. In studies conducted at Michigan State University, television was found to be as adequate in presenting information to students as the live teacher (Davis and Johnson, 1966). Likewise, in a study comparing data collected through both live and videotape observation of classroom interaction, no significant differences were found (Long, 1971).

The use of videotape technology to present the testimony of witnesses unavailable to testify at time of trial is one methodology that is gaining increasing acceptance. In fact, during the past three years Presidents Nixon, Ford, and Carter have all given testimony via videotape. In a study of the use of videotape in the courtroom, Miller and Fontes (1977) found there were no significant differences between information retention of jurors when television was used to present testimonies instead of live observation.

Evidence did seem to indicate, however, that the credibility of some testimonies was enhanced when presented by television. This was explained by the fact that jurors felt anyone who presented testimony via television was probably important enough that he could not be present for the trial.

Videotape has proven to be an important simulation technique for self-assessment and evaluation in medical education. Assessing interviewing and counseling skills has been done effectively with videotape (Betts, 1974; Simpson, 1974; and Suess, 1970). DeMers, Lawrence, and Callen (1976) describe the role of videotape in evaluation of students in a decentralized medical education program. In addition to written examinations, students are assessed on their responses to videotaped, simulated clinical problems. The students themselves are also videotaped while interviewing programmed actors or patients with known diagnoses. Trained observers rate the histories, physical examinations, quality of analysis of patient problems, and management plans.

In 1965, while searching for better ways to evaluate clinical competence, the National Board of Medical Examiners experimented with the application of TV/film to its evaluation process. They found that TV/film offered a more manageable (as compared to live

observation) method that was an objective and reliable means of evaluating the ability of the intern to observe a patient accurately, to judge the competence and completeness of a physical examination, and to arrive at appropriate conclusions based upon what he has seen (Hubbard, Levit, Schumacher, and Schnabel, 1965).

The results of "no differences" between television and live observation are consistent across the literature in this area, but how important are these findings? What can be inferred from them? Salomon (1974) commenting on these types of comparative experiments states:

Unfortunately, the typical experiment in which the effectiveness of one medium is compared with that of another is in actuality a study of different technologies. When live teaching is compared with televised instruction, the real difference between the two is one of transmission rather than anything inherent in any particular symbol system. Thus, it is not surprising that the vast majority of such studies reveal no significant differences (p. 385).

The manner in which the previously mentioned experiments were conducted is referred to by Salomon and Clark (1977) as "research with media" as compared to "research on media." A major difference exists between the two. Whereas research with media employs them only as modes of stimulus presentation, but studies nothing inherently connected with them, research on media treats them as its major focus of investigation. The design of the comparative experiments cited in the

literature indicates they all tend to be simple one-way analyses of variance with Alpha levels set at .05. The problem with this is that the experiment was designed to minimize Type I errors, the decision to accept differences when in fact there were none (Cohen, 1976). The fact that no differences were found at the .05 Alpha level cannot lead to the inference that therefore the two media are equivalent. The critical omission in these studies is the power of their experiments. This refers to the Type II error or the ability of the experiment to detect true differences if indeed they do exist (Cohen, 1976). If no differences are expected in a study, then designing the experiment for a high power (.80 or above) (Cohen, 1976) will enable the experimenter to state his confidence in retaining the decision of no differences.

#### Observer Perception

Samph (1976) found that in studying observer effects, one is constantly faced with the problem of the observers' and observees' perceptions of the purpose of an observation. Ultimately, these perceptions should be standardized across each group.

Two other aspects of observational evaluation, both dealing with the issue of fidelity of mediated methods, relate to this dissertation. The first is what Miller (1975) referred to as selected exposure or

the manner in which camera angles are selected. In the study of jurors' reactions to videotaped courtroom proceedings, it was found that a split-screen effect, where the attorney and the witness appeared in the top half of the screen with a panoramic view of the entire courtroom in the bottom half, was considered a disadvantage compared to a single shot of a familiar setting. The split-screen effect was criticized because of its lack of realism. It did, however, produce a more visible though less natural presentation. A noted motion-picture researcher, James Gibson (1947), comments on this issue of realism:

Although the objects and events presented in motion picture form are not fully "real" they approximate reality more nearly than do ordinary photographs and pictures or verbal descriptions. By moving, they become animated, i.e., alive. This tendency for scenes on the screen to appear real should not be overlooked in designing certain types of tests. To a great extent, the observer loses himself in the scene, i.e., locates himself in the environment and in the situation being portrayed. This attitude of "being there and seeing it happen" is compelling; it can only be overcome by deliberately attending to the frame of the screen image or to objects in the projection room. This tendency to adopt the attitude of reality is much more striking in motion pictures than in any other form of pictorial or photographic representation. Participation by the observer in the situation being portrayed can be enhanced by a number of camera techniques. The location of the camera in the scene photographed can be such as to make the observer see what a participant sees . . . .

This use of the camera as a participant is in contrast with its more frequent use as a story-telling agent in entertainment films (p. 285).

Gordon (1978), in contrasting the use of media for instruction versus evaluation, states that when a one-way flow of information medium (such as television) is used for instruction, the content presented is forced to be in a linear fashion not allowing for interchange. Therefore, its effectiveness is limited. On the other hand, when a one-way flow of information medium is used for evaluation, the fact that interchange is not possible does not affect its use. What is important is that the critical variables necessary for effective evaluation are present in the medium. That is, the variables that are used in a live evaluation must be present in the mediated representation. McKeachie (1975) supports the use of television for examining performance for evaluation purposes, not for learning.

The results of current research suggest that simulation techniques are especially useful in evaluating: (a) skill in gathering and interpreting clinical and laboratory data; (b) judgment in patient management; and (c) the professional skills and attitudes required for effective interaction with patients and colleagues (McGuire and Wezeman, 1974). Clearly, greater reliability can be placed on any judgment about an evaluator's

interpretative skills if he is asked to interpret data presented in the form in which he usually encounters them rather than some other method. For this reason, the use of motion pictures and videotape is as valuable as live observation in assessing important aspects of clinical competence.

This aspect of simulation (its correspondence to real-life situations) adds face validity to the process of evaluation. Such perceived relevance is at the very least of psychological and motivational benefit (McGuire, 1973).

The other aspect of fidelity is that of technical quality. It is important to implement the medium used for observation in such a way that both the picture and sound are clear and precise. The medium should be used as unobtrusively as possible but not to the extent that the resultant quality will impair the evaluation procedure.

### Summary

This chapter reviewed literature related to two general topics. The first concerned the issue of oral assessment techniques and the use of Simulated Clinical Encounters (SCE's) in medical specialty certification examinations. The review indicated that in order to improve the reliability and face validity of oral

examinations, simulations were introduced. The characteristics of simulations allow for the assessment of specific aspects of content and skills not assessed by other test methods.

The review concluded that studies of reliability and validity of simulations indicate that acceptable levels of reliability and face validity can be achieved in oral examinations through the use of structured simulations and proper examiner training.

The second part of the chapter reviewed observational evaluation literature. The review indicated that there is generally no difference between live observation and observation via television. Many studies have been conducted comparing the two forms of observation/instruction but none have accounted for the power of their experiment to detect true differences if indeed they did exist. The studies are deceptive to the extent they imply that the two forms of observation are equivalent when no differences were found. Finally, it was noted that the perceptions of the participants in an evaluation should be standardized and the fidelity of the alternative to live observation should be high both in its relation to the real world and to its technical quality.



## CHAPTER III

### DESIGN OF THE STUDY

#### Introduction

The principal concern of this study is to determine if there are meaningful differences in scores received from evaluating oral simulations through mediated versus live observation. The data for this dissertation were collected as a subset of a larger research program. Elements of this program relevant to this dissertation are discussed below. This chapter begins with a description of the structure of the Emergency Medicine Specialty Certification Examination (EMSCE) Field Test and the corresponding relationship this dissertation has to it. Indicated within this description are the design, population, and sample used for the dissertation experiment. Next, the experimental facilities and television production used in the dissertation experiment are described, followed by a description of the instrument used to evaluate the Simulated Clinical Encounters (SCE's). The statistical analysis, including a discussion of power analysis, along with a listing of the hypotheses is then presented. Finally, a summary is provided.

The Emergency Medicine Specialty  
Certification Examination  
Field Test

Data for this dissertation were collected during the Field Test and a post-Field Test experiment of the EMSCE. The Field Test was conducted October 22-24, 1977, at the Hilton Motel in Lansing, Michigan. During those three days, all of the testing formats, (multiple-choice questions, patient-management problems, pictorial multiple-choice questions, and the Simulated Clinical Encounters) were presented to 94 candidates.

Twenty-four emergency physicians administered the problem and evaluated candidates' performance in the SCE's. These physicians were selected by the American College of Emergency Physicians, for whom the EMSCE was developed. Two days prior to the EMSCE Field Test, all 24 physician/examiners went through an extensive workshop on how to administer the SCE's. The purpose of the training workshop was to standardize rater performance.

During the three-day testing period, the author was responsible for videotaping a total of 24 15-minute SCE's. The videotaping procedures are described later in this chapter. The data from these SCE's were used to divide the 24 videotaped cases into two categories. The first category represented four levels of performance; poor, fair, good, and excellent. The second category

represented perceived age group of the videotaped candidate (e.g., medical student, resident, physician). Four cases, each representing a different mix of the combined categories (level of performance and perceived age) were randomly selected to be presented, via videotape, in a post-Field Test experiment.

#### The Post-Field Test Experiment

The post-Field Test experiment was designed to assess the post-test reliability of examiner ratings and the effects of candidate visual cues on scores (Holmes, 1979). Although there were 24 examiners in the Field Test, only 21 of them were available for the post-Field Test experiment. In this experiment, examiners were randomly assigned to two rating teams and teams were randomly assigned to the treatment groups. The two treatments were audio-video and audio only. One group (N = 12) viewed the four SCE's on two 23-inch television monitors (see Figure 3.3) while the other group (N = 9), in a different room, were exposed to only the audio portion of the four videotaped cases. The videotaped cases were presented sequentially, pausing approximately three minutes between SCE's to collect the rating forms and distribute another blank one. The 21 examiners represented in this post-Field Test experiment used the

same rating form that was used by the examiner during the Field Test to evaluate the candidate's performance. Later analysis revealed no significant differences in scores received from the two media methods ( $F = .09$ ,  $d.f. = 1, 16$ ,  $p = .84$ ). In order to analyze this decision, three raters had to be randomly omitted from the video rating team to provide a nested  $N$  for each group of 9. For the purpose of the dissertation experiment, the two media groups were pooled, giving a sample size ( $N$ ) of 21. The inter-rater reliability, reduced to one rater, for the combined media treatment group of this post-Field Test experiment was  $r = .75$ .

It is the author's intention in this dissertation to compare the rating scores received by the candidates from the combined media treatment group of the post-Field Test experiment, (evaluation through use of media) to the rating scores the same four candidates received during the Field Test (evaluation through live observation).

### Experimental Facilities

The SCE's were conducted in a large ballroom at the Hilton Motel. Eight curtained booths were set up in the ballroom, each with a 3 x 6-foot table with chairs on each side. Figure 3.1 is a floor plan of the ballroom in which the SCE's were administered. Figure 3.2 represents the floor plan of a single examination booth.

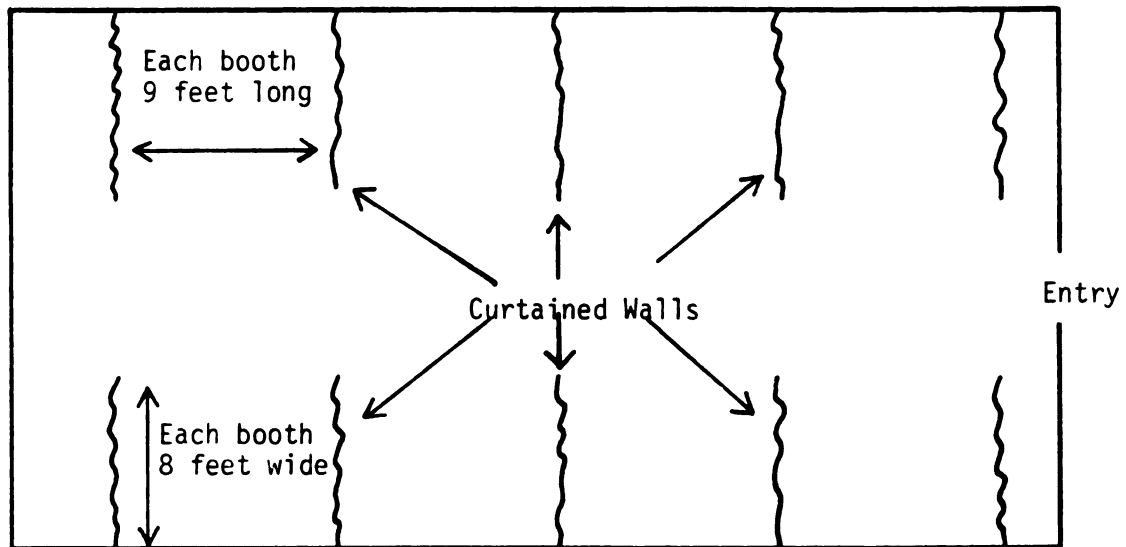


Figure 3.1.--Floor Plan of Ballroom

#### Television Production

Figure 3.2 shows the placement of the television camera in the individual examination booth. The camera lens was placed through a slit in the curtained wall and was the only indication that the examination was being videotaped. The camera itself, the tripod with wheels it was mounted on, and the camera operator, were not visible to the candidates. The camera was placed in a line as near to the head of the examiner as possible and framed on the candidate in such a way as to, as nearly as possible, duplicate the same field of view and from the same perspective as the examiner's. The

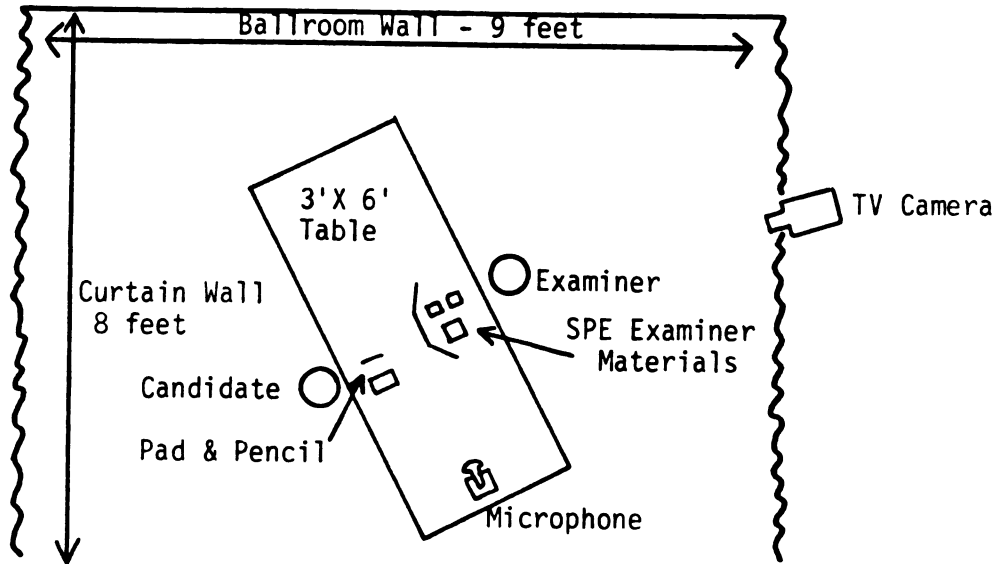


Figure 3.2.--Floor Plan for Individual Examination Booth.

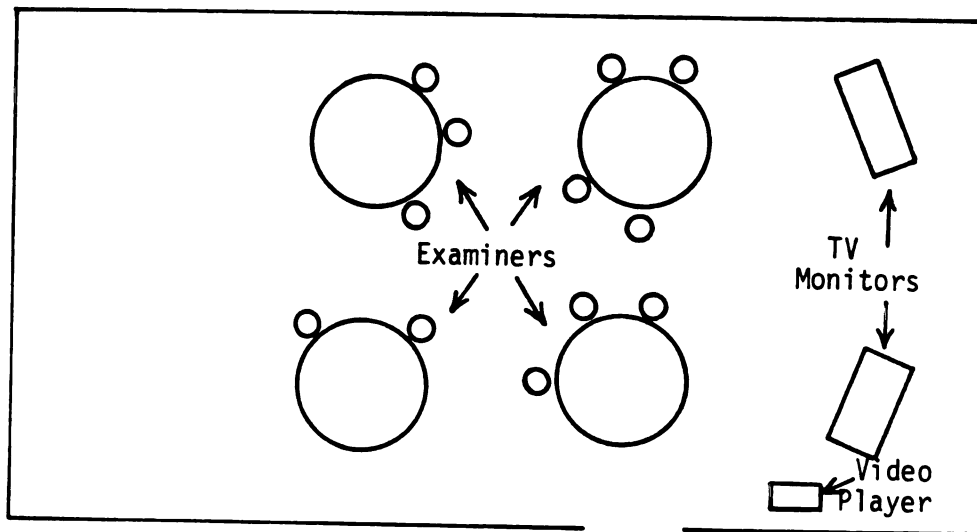


Figure 3.3.--Floor Plan of Post-Field Test Experiment.

objective was to communicate with as much fidelity as practical the information the examiner was receiving.

To record sound, an Altec, Model 945-S microphone mounted onto a desk stand was placed on the end of the table, well away from the examiner, the candidate, and the various materials used during the simulation (see Figure 3.2).

The recording and monitoring equipment used to tape the SCE's consisted of one SONY model VO-1800 video-cassette recorder and one SONY model TV 960 black-and-white video monitor. These two pieces of equipment were placed on a roll-around cart and positioned near the center of the ballroom. The roll-around cart was seldom moved because both the camera and microphone had a cable run of 25 feet back to the video-cassette recorder. Therefore, they could be moved into position for any of the eight booths in the room without moving the position of the video-cassette recorder itself.

All SCE's were taped in black-and-white using a SONY AV-3250 black-and-white camera. The decision was made to tape the SCE's in black-and-white owing to the fact that available lighting in the ballroom was only approximately 70 foot-candles. Therefore, to be able to videotape in color would require additional lighting which would greatly impair the testing environment and

violate the goal to conduct the experiment as unobtrusively as possible. An additional concern was the extra cost of color equipment and the required expertise to properly operate it.

During the Field Test of the EMSCE, the author, with an aide, was responsible for setting up the equipment and videotaping all of the SCE's.

### Instrumentation

The evaluation instrument designed to evaluate the SCE's was developed by OMERAD test development personnel along with various task forces made up of emergency physicians who were cooperating in developing specific case problems to be presented in the SCE form. All of the rating forms used to rate candidate performance were based on an eight-point semantic differential scale. Seven separate ratings of a candidate's performance were on specific and general clinical skills deemed necessary for a physician specializing in Emergency Medicine.

In addition, the rating form included a behavior checklist section that listed the minimum and essential critical actions required of a 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 by checking the appropriate yes or no column.



2. To aid in standardizing any subsequent subjective ratings.

The design of the instrument was influenced by experience gained from the Canadian College of Family Physicians and through development of various instruction and evaluation projects within OMERAD. One important aspect of the instrument is its close nature to the simulation itself; they were developed together. The instrument was approved for use in the Field Test of the EMSCE by the American Board of Emergency Medicine, the certification-granting agency for emergency medicine. The appropriateness of the instrument was tested during the 2-year simulation development phase of the EMSCE. The mean inter-rater reliability for 12 problems is .79.

In an effort to use a rationale for supporting the null hypothesis in this dissertation, a discussion of power analysis is appropriate.

### Power Analysis

In performing a statistical test of a null hypothesis, one is normally concerned with supporting  $H_1$ . As a consequence the probability of a Type I error is a concern, that is, rejecting  $H_0$  when it is actually true. The probability of the error is specified by Alpha, the significance criterion. Since a small value of Alpha is traditionally used (.05) one can be confident, when rejecting  $H_0$ , that the decision is very likely to be

correct. Hypothesis testing, however, also poses the risk that even if  $H_0$  is false, it may fail to be rejected, that is, a Type II error is made. The probability of this error is Beta. Therefore, Alpha is the probability of rejecting the null hypothesis when it is actually true (Type I error) and Beta is the probability of retaining the null hypothesis when it is false (Type II error). The complement of the Type II error,  $(1-\beta)$ , is the power of the statistical test. This could be defined as the probability of being correct in supporting the null hypothesis as the true state of nature.

It is almost universally accepted by educational researchers that the power (probability of rejecting a false  $H_0$ ) of a statistical test is important and should be substantial (Brewer, 1972; Cohen, 1969; Hays, 1974).

What needs to be emphasized at this point is that the type of power just described is the traditional sense of power. That is, power relating to the traditional method of hypothesis testing where one desires to reject a false null hypothesis and thus support the alternative hypothesis. Myers (1979) describes this as Power I. The corollary to this is Power II, or the probability of retaining a true null hypothesis and thus rejecting a false alternative hypothesis (see Figure 3.4).

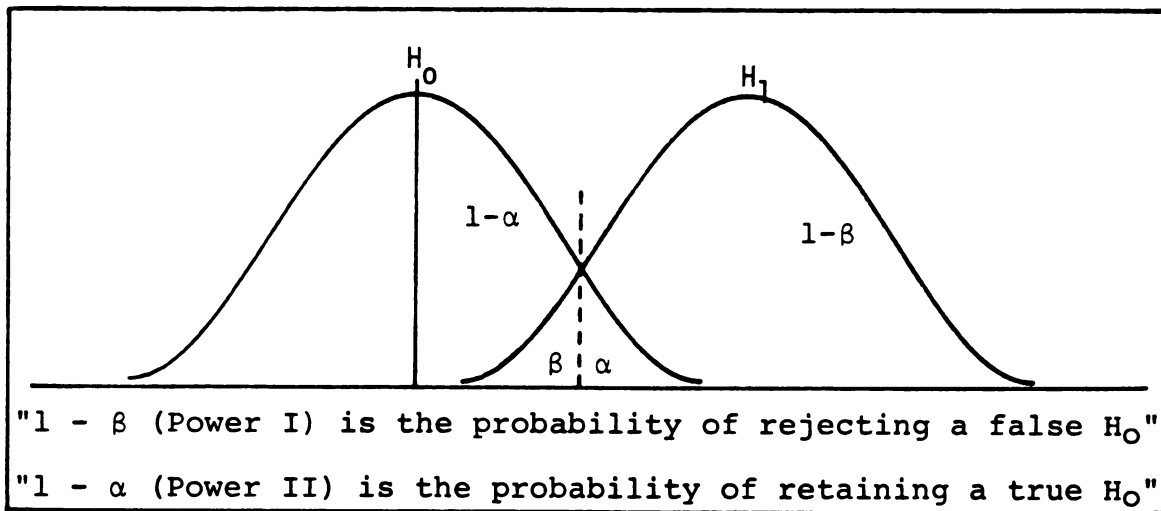


Figure 3.4.--The Relationship of Power I and Power II (Myers, 1979).

Myers (1979) has extended the technology of hypothesis testing to facilitate research in controlling for the probability of a Type II error (retaining a false  $H_0$ ). In this method, the desired outcome is to reject the alternative hypothesis. A noncentral sampling distribution must be defined so that the critical region may be located on it (the alternative hypothesis,  $H_1$ ) rather than on the usual central distribution. This noncentral sampling distribution of the test statistic is defined by the parameters of effect size and the appropriate degrees of freedom. The probability of an incorrect decision is chosen a priori (e.g., Beta = .05). An effect size is chosen that is the largest difference the researcher is willing to accept as having no practical

significance. There are two methods for doing this. Most desirable would be a theoretical rationale for choosing either a small, medium, or large effect size difference based upon what the literature indicates. If the literature does not indicate what the size effect may be, Cohen (1976) has suggested using a convention to define small, medium, and large effect sizes. He defines these three levels in terms of standard scores. The three standard score values he suggests are .20, .50, and .80.

Myers (1979) converts these standard score effect size units to the population squared multiple correlation -  $p^2$ . The  $p^2$  parameter is interpreted as the proportion of variance accounted for in a dependent variable by a linear model. Cohen's (1976) medium effect size of .50 standard score converts to a  $p^2$  of .0588. Next, the sample statistic,  $S_g$ , is derived. The alternative hypothesis will be rejected if  $S_g \leq$  the criterion ( $S_c$ ). The relationship of Power I and Power II can be found in Figure 3.4.

Some researchers support the Fisher (1949) philosophy of hypothesis testing. This philosophy holds that the only statistical decision that can be made is to reject  $H_0$  when the test statistic falls within the critical region or reserve judgment until further

evidence has been gathered. Neyman-Pearson (1933) use the phrase "reject or accept" the null hypothesis as two decision choices that can be made. Myers (1979) points out that if one were to read the original articles of Neyman-Pearson and others, as to the nature of the null hypothesis, it would become clear that one cannot prove or disprove a hypothesis but can only support or fail to support it.

To summarize Myers's approach to hypothesis testing, when the desired outcome is to support the null hypothesis, the researcher:

1. Chooses a priori a high level of significance (e.g.,  $\beta = .05$ ) to protect against Type II error.
2. Chooses an effect size that represents the maximum difference the investigator is willing to accept as having no practical meaning. If estimates for differences fall within this size effects interval, they are interpreted as having no practical meaning.
3. Uses the effect size, appropriate degrees of freedom, and Beta to define a noncentral sampling distribution of the critical statistic ( $S_C$ ) upon which is located the critical region of rejection.
4. Determines the critical statistics ( $S_C$ ) value which is computed or read from a table and compared to the sample statistic ( $S_S$ ).
5. Decides;  $H_1$  is rejected if  $S_S \leq S_C$  and  $H_0$  is said to be supported.

Following Myers, it is clear that decisions to support the null can be made with as much confidence as

is traditionally done for supporting the alternative hypothesis.

### Design of the Study

This study attempts to determine the extent to which mediated and live evaluation can be considered meaningfully equivalent methods. Table 3.1 represents the post-Field Test experiment described earlier in this chapter.

One of the problems associated with this design is the concern whether or not the 21 raters in the media group are actually measuring the same thing as the one rater in the live group. In order to control for this concern, the media group was reduced to  $N = 1$ , representing the same rater from the Field Test. Thus the same rater is in the media and live cell for each candidate. Table 3.2 represents the relationship among candidates, treatments, and raters. There is concern that by reducing the media group to an  $N$  of 1, the investigator is throwing out information. Initially this may seem correct, but when the researcher attempts to determine the power of his statistical test, he is faced with a low effective  $N$  for the first design, Table 3.1.

Ordinarily, in research where variables are to be manipulated, it is possible to arrange the experiment so that the total available pool of subjects is divided

TABLE 3.1.--Design of Post-Field Test Experiment.

---

	Media	Live
C <sub>1</sub>	N = 21	N = 1
C <sub>2</sub>	N = 21	N = 1
C <sub>3</sub>	N = 21	N = 1
C <sub>4</sub>	N = 21	N = 1

---

TABLE 3.2.--Design of Dissertation Experiment.

---

	Media	Live
C <sub>1</sub>	N = 1	N = 1
C <sub>2</sub>	N = 1	N = 1
C <sub>3</sub>	N = 1	N = 1
C <sub>4</sub>	N = 1	N = 1

---

equally among the groups. This division is optimal. But it is not always possible or desirable to have equal sample sizes. Therefore, to determine power when  $N_1 \neq N_2$ , the harmonic mean is computed (Cohen, 1976).

$$\text{Harmonic Mean} = \frac{2N_1N_2}{N_1+N_2}$$

For example, if one study had a treatment with an  $N = 21$  and the other treatment  $N = 1$ , the harmonic mean would be calculated:

$$\frac{2 \frac{(21)(1)}{(21)+(1)}}{2} = 1.909$$

It is evident that when there are large differences between two treatment groups, the effective  $N$  is largely determined by the smaller group. Because of the fact that subjects are now the same for each treatment, and assuming that the data are interval and distributed as a  $t$ , a dependent  $t$ -test was used for analysis.

### Hypotheses Tested

The major concern in this dissertation is that in order to support the use of media in evaluating oral simulations, the scores cannot deviate more than  $.06p^2$  (a meaningless difference) than those from live evaluation. If that is found, media will be accepted as equally valid.



Due to the fact that the sample size in this disseration experiment was so small, it would be hard to determine small effects between the treatments. Therefore, the author selected a medium size effect ( $.50_z$ ) which will enable a more reasonable test to be used on the data. In fact, Lumsdaine (1963) suggests that no experiment should be conducted that will not detect a difference of one-half of a standard deviation,  $.50_z$ .

Therefore the following hypotheses were tested:

$$H_0: p^2 \leq .06$$

$$H_1: p^2 > .06$$

The  $.06 p^2$  value in the statement of hypotheses relates to the size effect or how much different these two methods can be and still be acceptable to the author as equivalent methods. This Myers (1979) value is very close to Cohen's (1976) medium size effect of .50 standard score ( $.50_z = .0588 p^2$ ). The results of using the Myers methodology to determine whether or not  $H_0$  can be supported are reported in Chapter IV.

### Summary

This chapter began with a description of the structure of the Emergency Medicine Specialty Certification Examination (EMSCE) Field Test and the corresponding

relationship this dissertation had to it. The design, population, and sample of the EMSCE post-Field Test experiment were presented along with a description of the experimental facilities and television production procedures. Next, the instrument designed to evaluate the Simulated Clinical Encounters (SCE's) was described, and a discussion of power analysis was presented. Finally, the design for this dissertation was presented along with a description of the tested hypotheses.

## CHAPTER IV

### ANALYSIS OF DATA

#### Introduction

The purpose of this study was to determine if there would be meaningful differences in scores obtained from evaluating oral simulations through mediated versus live observation.

As a part of the Emergency Medicine Specialty Certification Examination (EMSCE) Field Test, a post-test experiment was conducted to assess the reliability of examiner ratings and the effects of candidate visual cues on scores. During this post-Field Test experiment the raters of Field Test Simulated Clinical Encounters (SCE's) were asked to rate candidates' SCE performance on videotape.

For purposes of this dissertation, the scores obtained from the post-test media experiment were compared to the scores the same four videotaped candidates obtained in the Field Test. This chapter reports the results of this comparison.

### Hypotheses

The following hypotheses were tested:

$$H_0: p^2 \leq .06$$

$$H_1: p^2 > .06$$

### Design of Experiment

The experimental design for this dissertation described in Chapter III can be found in Table 4.1. Across the top of the table are the two observational methods that were compared. The four candidates that were in each treatment are represented on the left. The sample size for both treatments is placed in each cell.

Table 4.2 displays basic descriptive statistics. Values have been placed in each cell with the accompanying means and standard deviations for each treatment.

Because subjects are the same for each treatment and assuming the data are interval and distributed as a  $t$ , a dependent  $t$ -test was used to analyze the data. The resulting  $t$ -value was  $-1.58$ . To use the Myers (1979) methodology described in Chapter III, the  $t$ -value was converted to an  $F$ -value.

$$t^2 = F \qquad -1.53^2 = 2.4964$$

This  $F$ -value of 2.4964 represents the sample statistic ( $S_s$ ). In order to determine the critical statistic ( $S_c$ ),

TABLE 4.1.--Design of Dissertation Experiment.

---

	Media	Live
$C_1$	N = 1	N = 1
$C_2$	N = 1	N = 1
$C_3$	N = 1	N = 1
$C_4$	N = 1	N = 1

---

TABLE 4.2.--Design for Analysis of Dissertation Experiment with Accompanying Values

---

	Media	Live
$C_1$	1.2857	3.4286
$C_2$	4.1429	3.5714
$C_3$	4.5714	6.2857
$C_4$	6.4286	7.00
$\bar{x}$	4.1071	5.0714
s	2.217	1.839

---

Myers (1979) provides a table of critical F-values for a noncentral distribution (see Table 4.3). The decision rule is if  $S_s \leq S_c$ , the alternative hypothesis can be rejected at a Beta level of .05 and a power of .95.

Using the degrees of freedom from an experiment, one can find the corresponding critical statistic ( $S_c$ ) for the level of Beta and effects size desired. For this dissertation, the sample size (N) was 4. Therefore, the degrees of freedom are one for the numerator, three for the denominator. To test the hypothesis ( $H_1$ ), the author selected a Beta of .05 and an effect size of  $p^2 = .06$ . Going to the Myers F-table, first column third row, the corresponding critical statistic ( $S_c$ ) is found as .006. Applying the decision rule,  $S_s \leq S_c$ , 2.4964 is much larger than .006. Therefore, it is evident the alternative hypothesis cannot be rejected for these data.

### Discussion

Based on the results, there can be no statistical statements made other than that the investigator failed to find support for the equivalence of the two observational methods. There are, however, indications that the differences between the the two methods are substantial-- $F = 2.49$ . This indication does not coincide with what has been found in the literature. In a situation with

Table 4.3 - Myers (1979) Critical Non Central F-Values; Effect Size:  $p^2 = .06$ ;  $\text{Beta} = .95$

	1	2	3	4	5	6	7	8	9	10	12	16	20	30	60
1	0.008	0.061	0.110	0.143	0.165	0.182	0.194	0.203	0.211	0.217	0.226	0.239	0.246	0.256	0.266
2	0.006	0.062	0.119	0.161	0.191	0.214	0.231	0.244	0.255	0.264	0.278	0.296	0.308	0.323	0.339
3	0.006	0.063	0.125	0.172	0.207	0.233	0.254	0.270	0.283	0.294	0.311	0.334	0.348	0.367	0.397
4	0.006	0.065	0.130	0.180	0.218	0.247	0.270	0.288	0.303	0.316	0.335	0.361	0.377	0.400	0.424
5	0.007	0.067	0.134	0.187	0.227	0.258	0.283	0.303	0.319	0.332	0.353	0.382	0.400	0.425	0.452
6	0.007	0.069	0.138	0.193	0.235	0.267	0.293	0.314	0.331	0.345	0.368	0.399	0.418	0.446	0.475
7	0.008	0.071	0.142	0.198	0.241	0.275	0.302	0.324	0.342	0.357	0.381	0.413	0.434	0.463	0.495
8	0.008	0.073	0.146	0.203	0.247	0.282	0.309	0.332	0.351	0.366	0.391	0.425	0.447	0.478	0.512
9	0.008	0.076	0.149	0.207	0.253	0.288	0.316	0.339	0.359	0.375	0.401	0.436	0.459	0.491	0.527
10	0.009	0.078	0.153	0.212	0.258	0.294	0.323	0.346	0.366	0.382	0.409	0.445	0.469	0.503	0.540
12	0.010	0.083	0.160	0.221	0.267	0.304	0.334	0.358	0.379	0.396	0.423	0.461	0.486	0.522	0.562
16	0.013	0.094	0.174	0.237	0.285	0.323	0.354	0.379	0.400	0.418	0.447	0.487	0.513	0.552	0.596
20	0.016	0.106	0.190	0.253	0.302	0.340	0.371	0.396	0.418	0.436	0.465	0.507	0.534	0.575	0.622
30	0.030	0.143	0.233	0.296	0.344	0.382	0.412	0.436	0.457	0.475	0.504	0.545	0.573	0.615	0.666
60	0.174	0.325	0.400	0.448	0.483	0.510	0.531	0.549	0.565	0.578	0.600	0.631	0.654	0.690	0.737
100	0.828	0.735	0.712	0.704	0.703	0.703	0.706	0.709	0.712	0.715	0.721	0.733	0.743	0.762	0.793
200	3.762	2.269	1.773	1.527	1.378	1.281	1.214	1.162	1.122	1.091	1.046	0.990	0.960	0.922	0.895
500	16.048	8.464	5.936	4.674	3.917	3.412	3.048	2.779	2.570	2.404	2.152	1.837	1.650	1.402	1.160
1000	40.155	20.540	14.024	10.750	8.786	7.477	6.543	5.842	5.297	4.861	4.208	3.392	2.901	2.249	1.599

indeterminate results, the best procedure is to conduct another study. An additional study could make this dissertation more meaningful.

Based upon the indications from the results of this dissertation, a follow-up study using EMSCE data (Maatsch et al., 1979) was conducted. This follow-up study aids in interpreting the results obtained in this dissertation. A description of the follow-up study can be found in Chapter V.

### Summary

This chapter has reviewed the experimental design of this dissertation. The scores received from two observational methods, mediated and live, were compared. The decision rule was if the sample statistic ( $S_s$ ) is equal to or less than the critical statistic ( $S_c$ ), the alternative hypothesis can be rejected. The  $S_s$  for the dissertation experiment was computed as 2.4964. The critical statistic from the Myers (197) table of noncentral distribution F-values was .006. Therefore, the alternative hypothesis could not be rejected. Based on these results, Chapter V will present a summary and discussion, and the conclusions and recommendations of this dissertation.



## CHAPTER V

### SUMMARY, DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

#### Introduction

This chapter contains general summaries of the problem and purpose of this study, the relevant literature, the experimental design, and the findings. Based upon these summaries, there are discussion, conclusions, indications, implications, and a number of recommendations drawn for further research and evaluative practice.

#### The Problem

Many areas of our country have come under public scrutiny in recent years. Demands for greater accountability have been placed on several sectors of American society. One area that has been seriously impacted by this movement is the health care system. Medical specialty boards have felt the pressure to increase the validity and effectiveness of their certification examinations. They must be better able to assess the competence of, and predict performance for, physicians.

In an effort to more effectively predict physician competence and performance, medical specialties are

turning to the use of structured oral examinations (McGuire, 1966). Hubbard (1971) points out that the problems associated with oral examinations are that they tend to be unreliable and difficult to administer.

Technology, in the form of television, offers potential to improve the quality control of oral examinations. Traditionally, oral examinations are evaluated by an examiner present in the room (live observation). Television offers an alternative form of observation that produces a referable product--the videotape. The videotape presents many opportunities for quality control in the oral examination process. Television also offers possibilities for cost effectiveness.

The purpose of this dissertation was to determine if there were meaningful differences in scores obtained from evaluating oral simulations through media versus live observation. This line of research was undertaken with the anticipation that evaluation through television might prove to be as valid an evaluative technique as live evaluation. If this dissertation found that television is a valid alternative evaluative technique, several assertions could be made to support the theory that television can be more useful than live evaluation.

Television could provide the basis for increasing reliability of examiner ratings by allowing several

raters to view the videotaped simulation and evaluate a candidate's performance. Television could allow for further post-hoc evaluation opportunities. It could also provide the basis for more flexible scheduling and evaluation of examinations. Television can be used to train raters for oral examinations as well as provide the basis for an outsider to critique the examination.

### The Literature

Reviewed in the second chapter of this dissertation were two general bodies of literature. The first concerned the problems associated with oral assessment techniques used in medical specialty certification examinations and subsequent use of Simulated Clinical Encounters (SCE's). The review indicates that when oral examinations were properly structured and controlled, the results were more reliable. Simulated Clinical Encounters represent the most recent evolution in the trend to assess aspects of competence not assessed by other written examination forms. They also offer potential to increase both reliability and validity of oral examinations.

The second part of Chapter II reviewed the literature on the use of observational evaluation. Due to the nature of this study, the use of television

observation as an alternative to live observation is emphasized.

Although the majority of the literature on television is from an instructional perspective, Salomon's (1974) research provides the link to use of this information for evaluative purposes. He states that the major role of television in an instructional or evaluative mode is that of transmitting information. Therefore, the findings from the literature can be applied to either modality. The review of this literature indicates that for evaluative purposes, generally no difference exists between the two forms of observation, provided there is obvious quality control. The majority of the studies cited in the literature lacked any determination of the power of the comparative experiment. The philosophy of hypothesis testing does not allow the researcher to support  $H_0$  when he simply fails to reject it. Therefore, one needs to know the power his test had to find differences if indeed they did exist.

### Design

The basis for this dissertation study was provided through the Field Test of a new Emergency Medicine Specialty Certification Examination (EMSCE). As part of the EMSCE, a test was conducted to determine if visual cueing had effect on the scores a candidate

receives in the oral simulations. In order to do this, several Simulated Clinical Encounters were videotaped during the Field Test. In a post-test session, 21 physician examiners from the Field Test rated candidate SCE performance on videotape. For purposes of this dissertation, candidate scores received in this media treatment were compared to the scores the same candidates received in the Field Test.

### Findings

The hypotheses for the dissertation were:

$$H_0: p^2 \leq .06$$

$$H_1: p^2 > .06$$

Due to the fact that subjects were the same for each treatment, a dependent t-test was used to analyze the data. The resulting t-value was -1.58.

The Myers (1979) perspective of using power analysis was then applied to the data. First the sample statistic ( $S_s$ ) was calculated by converting the observed t-value to an F-value ( $t^2 = F$ ). This  $S_s$  was then compared to the critical tabled F-value ( $\beta = .05$ ) ( $S_c$ ) in order to determine if  $H_1$  could be rejected. If  $S_s \leq S_c$  then  $H_1$  can be rejected. Since 2.4964, the sample statistic ( $S_s$ ) is greater than .006, the critical statistic ( $S_c$ ),  $H_1$  could not be rejected.

### Discussion

Two approaches can be used when a researcher wants to derive meaning from a set of data. The first, and most powerful, are the traditional methods of hypothesis testing. In this approach, tests are run in order to find support for the hypothesis of interest: null or alternative. When a hypothesis is rejected, the other is then supported by indirect inference. If neither  $H_0$  or  $H_1$  can be rejected, the researcher is faced with no statistical meanings other than that he was unable to find significant differences. If one has complete experimental control, this situation can be avoided by predetermining the needed sample size. If the investigator does not have this control, this approach is still an appropriate first step in what could be considered exploratory experimental research.

Given inconclusive results, one would ultimately need to replicate the study under more optimal conditions. The direction the follow-up study would take should be based on what the original data appeared to indicate. Many times a researcher is placed in a position where there are indeterminate results but a decision still has to be made. Mosteller and Tukey (1977) point out the primary value of data lies in what they indicate, what they appear to show. Sometimes indication is as far

as we need carry an analysis. Indication can be numerical or qualitative. Thus indication can be descriptive statistics (mean, standard deviation, etc.) or it can include any hints and suggestions obtained from data by an understandable process, suggestions that might prove informative to a reasonable mean.

### Indications of the Data

A look at what the data appear to indicate in this dissertation was insightful to the author. The fact that differences between the media and live treatment groups were not significant cannot discount the fact that differences were suggested in the mean score from each treatment. One of the indications of the data was the computed size effects of the difference using the Myers's (1979) formula:

$$\hat{p}^2 = \frac{df_1 (F-1)}{df_2 + (df_1) (F)}$$

Using this formula, the estimated population size effects was very large, about 27 percent. In light of this indication, a follow-up study (Holmes, 1979) was conducted by the EMSCE test development group. The purpose of this study was to determine if there is a difference between an examiner rating alone (Field Test) versus in groups (verifiers in Field Test). In other

words, does the presence of a peer affect an examiner's rating?

During the Field Test, approximately one-third of the oral simulation testing encounters had a verifier examiner in addition to the administrator examiner. The data for these SCE's were analyzed by a dependent t-test analysis with examiners ( $N = 24$ ) as the unit of analysis. Significant differences were found ( $t = 3.28$ ,  $d.f. = 23$ ,  $p = .002$ ). It appears that examiners are slightly more liberal in their scoring when they are alone with candidates. This was a totally unanticipated outcome. In future studies, raters need to be trained to eliminate this peer effect. If this is impossible, scores could be adjusted statistically to remove this effect.

It is reasonable to suspect that these Peer Study effects are confounded in the experiment used in this dissertation. This evidence tends to indicate why the estimated population size effects were so large. One might hypothesize that if peer effect were eliminated, the two treatments would be much closer to each other. Applying the Myers (1979) formula, the estimated population size effects of the Peer Effects Study was 29 percent. This finding suggests that the equivalence of the two treatments used in this dissertation was much closer than what was found.



### Implications

Many implications can be drawn as a result of this dissertation and the follow-up Peer Effects Study. First and foremost, if an investigator plans to use television in the same manner as was used in this dissertation, results obtained will probably be confounded with peer effects. If the peer effects can be eliminated through rater training or statistical methods, then the methods used in this dissertation could be implemented. A caveat here would be that no effects are associated with media. This needs to be confirmed empirically. If it is impossible to eliminate the peer effects, it is suggested that television be used in such a way that individual raters are isolated with the television set. An advantage of this arrangement is that it could allow for investigating the nature of the peer effect. Is it caused by proxemics, or that the presence of another rater causes more objectivity, or other unknown factors? Television is one method that can be used in designing experiments to isolate the unknown nature of peer effects.

Separate from the above implications is the validity issue of the scores received from mediated versus live evaluation. The results of the dissertation experiment indicate that scores from media are more

conservative than live evaluation. Which method is more valid? The public would probably argue for the mediated method. The medical profession would probably favor the live method. The answer is a judgment call and therefore is outside the intended parameters of this dissertation.

### Conclusions

Two major conclusions emerge from this dissertation:

1. As contrasted in this study, the two treatments, media and live, cannot be considered the same.
2. The differences between the two treatments are confounded with known effects.

### Recommendations

Based on the findings of this exploratory study, several recommendations can now be made.

1. Additional research should be conducted to determine if there is a media effect independent of peer effects.
2. Researchers should take into account possible peer effect in the use of media as an evaluative technique.
3. Statistical power should not be considered a monolithic concept. To date the literature has dealt

with power only as it relates to the probability of being able to reject a false  $H_0$ . Myers (1979) has extended the logic of hypothesis testing to include a second type of power, the probability of retaining a true  $H_0$ . Researchers like Cohen (1976) and Brewer (1972) have implied and suggested that for a fixed sample size (N), the researcher can sacrifice Alpha in order to build up the power of a test (what Myers refers to as Power I,  $1-\beta$ ). The problem with this approach is that in increasing Power I, the probability of rejecting a false  $H_0$ , the researcher decreases Power II ( $1-\alpha$ ), the probability of retaining a true  $H_0$ , for a given sample size. As a result, the researcher can be confronted with interpreting inconclusive results. It is clear now that other things being equal, Power I and Power II are inversely related. For a given effects size, the only way to achieve both powers concurrently is to increase the sample size (N).

It is hoped that the present study has made two contributions: (1) in general, to the use of technology in the evaluative process of oral certification examinations; and (2) in particular, to the understanding and use of power analysis.

## 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, 1976, pp. 15-17.
- Allen, William. "Instructional Media Research: Past, Present, and Future." Audiovisual Communicative Review 19 (1971): 5-18.
- Anderson, Scarvin B.; Ball, Samuel; Murphy, Richard T.; and Associates. Encyclopedia of Educational Evaluation. London: Jossey-Bass Publishers, 1975.
- Armsey, James W., and Dahl, Norman C. An Inquiry Into the Uses of Instructional Technology. New York: 1973.
- Barbatsis, G. S. "The Nature of Inquiry and Analysis of Theoretical Progress in Instructional Television from 1950-1970." Review of Education Research 48 (Summer 1978): 399-414.
- Brewer, J. K. "On the Power of Statistical Tests in the American Educational Research Journal." American Educational Research Journal 9 (Summer 1972): 391-401.
- Bronowski, J. A Sense of the Future. Cambridge, Mass.: The MIT Press, 1977.
- Bull, G. M. "Examinations." Journal of Medical Education 34 (December 1959): 1154-1158.
- Burton, J. R. "The Issues in Medical Education Today." Journal of Medical Education 52 (1977): 8-12.
- Caffarella, Edward P., Jr. "The Cost-Effectiveness of Instructional Technology: A Propositional Inventory of the Literature." Unpublished Ph.D. dissertation, Michigan State University, 1973.

- Carter, H. D. "How Reliable are Good Oral Examinations?" California Journal of Educational Research 13 (September 1962): 147-153.
- Chu, G. C., and Schramm, W. Learning From Television: What the Research Says. Washington, D.C.: National Association of Broadcasters 1968.
- Clark, R. E. "Doctoral Research Training in Educational Technology." Educational Communications and Technology Journal 26 (Summer 1978): 165-173.
- Cohen, J. Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press, 1969.
- Cohen, J. Statistical Power Analysis for the Behavioral Sciences. Rev. ed. New York: Academic Press, 1977.
- Cohen, J.; Welkowitz, J.; and Ewen, R. B. Introductory Statistics for the Behavioral Sciences. New York: Academic Press, 1976.
- Cowles, J. T. "Current Trends in Examination Procedures." Journal of American Medical Association 155 (1954): 1383-1387.
- Davis, Robert H., and Johnson, F. Craig. Final Report: Evaluation of Regular Classroom Lectures Distributed by CCTV to Campus and Dormitory Classrooms. East Lansing: Michigan State University, 1966.
- Dayton, C. M. The Design of Educational Experiments. New York: McGraw-Hill Book Company, 1970.
- DeMers, J. L.; Lawrence, D.; and Callen, W. B. Educating New Health Practitioners: The MEDEX Northwest Approach. Seattle: University of Washington, 1976.
- DeMers, J. L.; Lawrence, D.; and Callen, W. B. Educating New Health Practitioners: The MEDEX Northwest Approach. Seattle: University of Washington, 1976.
- Dubin, R., and Hedley, R. The Medium May be Related to the Message: College Instruction by TV. Eugene: University of Oregon Press, 1969.

- Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.
- Ebel, R. L.; Noll, V. H.; and Bauer, R. M. Encyclopedia of Educational Research. 4th ed. Toronto: The Macmillan Company, 1969.
- 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.
- Fisher, R. A. The Design of Experiments. 5th ed. Edinburgh: Oliver and Boyd, 1949.
- Foster, J. T.; Abrahamson, S.; Lass, S.; Girard, R.; and Garris, R. "An Analysis of an Oral Examination used in Specialty Board Certification." Journal of Medical Education 44 (1969): 951-954.
- Frick, T., and Semmel, M. I. "Observer Agreement and Reliabilities of Classroom Observational Measures." Review of Educational Research 48 (Winter 1978): 157-184.
- Gage, Nathaniel Lees. Handbook of Research on Teaching. American Education Research Association. Chicago: Rand McNally, 1963.
- Gibson, J. J. Motion Picture Testing and Research. Washington, D.C.: U.S. Government Printing Office, 1947.
- Glass, Gene V., and Stanley, Julian C. Statistical Methods in Education and Psychology. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1970.
- Grant, D. "Testing the Null Hypothesis and the Strategy and Tactics of Investigating Theoretical Models." Psychological Review 69 (1962): 34-61.
- Hempel, C. G. Philosophy of Natural Science. Englewood Cliff, N.J.: Prentice-Hall, Inc., 1966.

- Henkel, R. E. Tests of Significance. Beverly Hills, Calif.: Sage Publications, 1976.
- Hitchens, Howard B. "The Benefits of Instructional Technology." In To Improve Learning: An Evaluation of Instructional Technology, Vol. 1. Edited by Sidney Trickton. New York: R. R. Bowker and Company, 1972.
- Hoban, Charles F., and van Ormer, E. B. Instructional Film Research 1918-1950 (Rapid Mass Learning). SCD Human Engineering Project 20-E-4, The Pennsylvania State College, Department of the Army and Department of the Navy, 1951.
- Holden, W. D. "The Evolutionary Functions of American Medical Specialty Boards." Journal of Medical Education 44 (1969): 819-829.
- Holmes, Presely D., Jr. Television Research in the Teaching-Learning Process. Detroit: Wayne State University, Division of Broadcasting, 1959.
- Holmes, T. Section on "Results-Part III." Model for a Criterion-Referenced Medical Specialty Test--A Progress Report on Grant Number HS 02038. East Lansing: OMERAD, Michigan State University, 1979.
- Hubbard, J. P.; Levit, E. J.; Schumacher, C. F.; and Schabel, T. G. "An Objective Evaluation of Clinical Competence." New England Journal of Medicine 272 (June 24, 1965): 1321-1328.
- Hubbard, J. P. Measuring Medical Education. New York: Lea and Febiger, 1971.
- Jamison, Dean; Suppes, Patrick; and Wells, Stuart. "The Effectiveness of Alternative Instructional Media: A Survey." Review of Educational Research 44 (Winter 1974): 1-67.
- Jason, Hilliard. "Educational Uses of Simulations; Attributes, Assumptions and Applications." Keynote Speech of Symposium on Simulation in Medicine. East Lansing, Mich.: Office of Medical Education Research and Development, 1973.



- 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 47 (1972): 789-795.
- Leifer, G. A. The Psychology of Teaching Methods. Chicago: N.S.S.E., 1976.
- Levine, H. G., and McGuire, C. H. "Role-Playing as an Evaluative Technique." Journal of Educational Measurement 5 (Spring 1968).
- Levine, H. G., and McGuire, C. H. "The Use of Role-Playing to Evaluate Affective Skills in Medicine." Journal of Medical Education 45 (1970): 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 7 (Summer 1970): 63-74.
- Lumsdaine, Arthur H. "Instruments and Media of Instruction." In Handbook of Research on Teaching. Edited by N. L. Gage. Chicago: Rand McNally, 1963.
- Maatsch, J. L., and Gordon, M. J. "Assessment Through Simulations." In Evaluating Clinical Competence in the Health Profession. Edited by M. K. Morgan and D. M. Irby. New York: C. V. Mosby Co., 1978.
- Maatsch, J.; Elstein, A.; Holmes, T.; Sprafka, S.; and Downings, S. Model for a Criterion-Referenced Medical Specialty Test--A Progress Report on Grant Number HS 02038. East Lansing: OMERAD, Michigan State University, 1979.
- 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.
- McGuire, C. H.; Solomon, L. M.; and Bashook, P. L. Construction and Use of Written Simulations. The Psychological Corporation, 1976.
- McGuire, C. H. "The Oral Examination as a Measure of Professional Competence." Journal of Medical Education 49 (1974): 18-34.



- McKeachie, W. J. Review of Educational Research 3. Itasca, Illinois: F. E. Peacock Publishers, Inc., 1975.
- Miller, G. R.; Bender, D. C.; Boster, F. J.; Florence, B. T.; Fontes, N. E.; Hocking, J. E.; and Nicholson, H. E. "The Effects of Videotaped Testimony in Jury Trials." Brigham Young University Law Review (1975): 331-373.
- Miller, G. R., and Fontes, N. E. Video Technology and the Legal Process. Beverly Hills, Calif.: Sage Publishers, 1978.
- Miller, G. R. "Jurors' Responses to Videotaped Trial Materials: Some Recent Findings." Personality and Social Psychology Bulletin 1 (1975): 561-569.
- Miller, G. R.; Bender, D. C.; Florence, B. T.; and Nicholson, H. E. "Real Versus Reel: What's the Verdict?" Journal of Communication 24 (1974): 99-111.
- Mosteller, F., and Tukey, J. W. Data Analysis and Regression. Reading, Mass.: Addison-Wesley Publishing, 1977.
- Myers, B. "Investigating Procedures Employing Effect Size and Level of Significance to Test Statistical Hypotheses." Unpublished Ph.D. dissertation, Southern Illinois University, 1978.
- Myers, B. E. "The Null Hypothesis as the Research Hypothesis." Paper presented at the American Educational Research Association, Division D, San Francisco, 1979.
- Neyman, J., and Pearson, K. "The Testing of Statistical Hypotheses in Relation to Probabilities a priori." Proceedings of the Cambridge Philosophical Society 24 (1933): 492-510.
- Pokorny, A. D., and Frazier, S. H. "An Evaluation of Oral Examinations." Journal of Medical Education 41 (1966): 28-40.
- Popper, Karl. The Logic of Scientific Discovery. New York: Harper and Row, 1959.
- Rozeboom, W. "The Fallacy of the Null Hypothesis Test." Psychological Bulletin 57 (1960): 416-428.

- Salomon, G., and Clark, R. E. "Reexamining the Methodology of Research on Media and Technology in Education." Review of Educational Research 47 (Winter 1977): 99-120.
- Salomon, Gabriel. "What is Learned and How it is Taught: The Interaction Between Media, Message, Task and Learners." In Media and Symbols: The Forms of Expression, Communication, and Education. Edited by David R. Olson. Chicago: University of Chicago Press, 1974.
- Samph, Thomas. "Observer Effects on Teacher Verbal Classroom Behavior." Journal of Educational Psychology 68 (1976): 763-741.
- Scanlon, Robert G. "Improving Educational Productivity Through the Use of Technology." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 18, 1974.
- Senoir, J. R. Toward the Measurement of Competence in Medicine. Carnegie Corporation of New York and the Commonwealth Fund, 1976.
- Simpson, Michael A. "Judging Clinical Competence and Videotape Techniques." (Letter). Lancet. (September 28, 1974), 778.
- Struening, E. L., and Guttentag, M., eds. Handbook of Evaluation Research. Beverly Hills, Calif: Sage Publications, 1975.
- Subkoviak, M. J., and Levin, J. R. "Fallibility of Measurement and the Power of a Statistical Test." Journal of Educational Measurement 14 (Spring 1977): 47-52.
- The Carnegie Commission on Higher Education. The Fourth Revolution: Instructional Technology in Higher Education. New York: McGraw-Hill Book Co., 1972.
- Tickton, Sidney G. To Improve Learning: An Evaluation of Instructional Technology, Vol. 1. New York: R. R. Bowker Co., 1972.
- Tiku, M. "Tables of Power of the F Test." Journal of the American Statistical Association 62 (1967): 525-536.

- Travers, Robert M. W. Second Handbook of Research on Teaching. Chicago: Rand McNally and Co., 1973.
- Travers, Robert M. W. Research and Theory Related to Audio-visual Information Transmission. Rev. ed. Western Michigan University, OE Contract No. 3-20-003, U.S. Department of Health, Education and Welfare, 1967.
- Tyler, R. W. Basic Principles of Curriculum and Instruction. Chicago: University of Chicago Press, 1950.
- Van Wort, A. D. "A Problem-Solving Oral Examination for Family Medicine." Journal of Medical Education 49 (1974): 673-680.
- Wilkinson, Gene L. "Cost Evaluation of Instructional Strategies." A-V Communication Review 21 (1973): 11-30.
- Williamson, J. W. "Validation by Performance Measures." Proceedings from a Conference on Extending the Validity of Certification, American Board of Medical Specialties, 1976.
- Winer, E. Statistical Principles in Experimental Design. 2nd ed. New York: McGraw-Hill, 1971.

MICHIGAN STATE UNIV. LIBRARIES



31293104377803