This is to certify that the

thesis entitled

An Analysis of the Effects of Different Multiple-Choice Item Selection Strategies on the Reliability and Validity of Measures of Physician Competence in Specialty Certification

presented by

Steven M. Downing

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

Robert L. Ebel

Major professor

Date May 24, 1979

AN ANALYSIS OF THE EFFECTS OF DIFFERENT MULTIPLE-CHOICE ITEM SELECTION STRATEGIES ON THE RELIABILITY AND VALIDITY OF MEASURES OF PHYSICIAN COMPETENCE IN SPECIALTY CERTIFICATION

by

Steven M. Downing

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling,
Personnel Services,
and Educational Psychology

1979

ABSTRACT

AN ANALYSIS OF THE EFFECTS OF DIFFERENT
MULTIPLE-CHOICE ITEM SELECTION STRATEGIES
ON THE RELIABILITY AND VALIDITY OF MEASURES
OF PHYSICIAN COMPETENCE IN SPECIALTY CERTIFICATION

By

Steven M. Downing

This study investigated the effect of two multiple-choice item selection strategies on the discrimination of physician competence in Emergency Medicine and on several other psychometric characteristics of item subscales selected by the two different criteria. The research was carried out in the context of a field test of a library of examination materials intended for certification of specialists in Emergency Medicine.

The ninety-four subjects for this study represented four distinct groups: Residency-eligible and practice-eligible emergency physicians, second-year residents in Emergency Medicine, and fourth-year medical students. Physicians with Board eligibility represented a national stratified random sample of emergency physicians judged by their peers as very clinically skilled and, therefore, certifiable in Emergency Medicine. Residents represented a national stratified random sample of beginning second-year residents in Emergency Medicine with a wide range of competence. Fourth-year medical students were paid volunteers.

Two examination formats were investigated: 1) Objective--single best-answer, four or five option multiple-choice and pictorial-stem

multiple-choice items, and; 2) Simulated Clinical Encounters--highly structured, examiner-administered and rated patient-game simulations of typical emergency medical cases.

Four 91-item subscales were selected from the 364 items of the Objective format. Two subscales were selected for an item-difficulty criterion. Two other subscales were selected for a relevance-to-clinical-medicine criterion, which was defined for this study as item point-biserial correlation with the grand mean rating on the independent criterion measure, Simulated Clinical Encounters.

Hypotheses about criterion-group discrimination, criterion-related validity, scale reliability, mean item difficulty, proportions of identical items selected for scales using different criteria, and differences in the distributions of pictorial-stem, clinical-situational, and factual multiple-choice items were tested. Residency-eligible (n=22), resident (n=36), and student (n=22) subject groups were used to test hypotheses for the following ninety-one item objective subscales:

1. Medium Difficulty: p-value .50 to .69 with a mean p-value of .63.
2. Low Difficulty: p-value .84 to .99 with a mean p-value of .90.
3. High Clinical-Relevance: item-criterion correlations of $r = .33$ to .68 with a median $r = .38$.
4. Low Clinical-Relevance: item-criterion correlations of $r = -.23$ to .11 with a median $r = .05$.

Discriminant Analyses showed that high clinical-relevance was the best discriminator and approximately 6.7 times more effective

than medium difficulty in statistically separating known groups, but this difference was not significant at $\alpha = .05$. Both medium difficulty and high clinical-relevance were significantly more discriminating of known groups than low clinical-relevance. High clinical-relevance correctly classified 76.3 percent of subjects, while medium difficulty classified 71.2 percent correctly.

The high clinical-relevance scale had a significantly higher criterion-related validity coefficient ($r_{xy} = .90$), was significantly more reliable ($r_{xx} = .95$), and was significantly lower in mean item difficulty ($\bar{p} = .72$) than the medium difficulty scale.

There was no statistical difference in the proportion of overlap of identical items between the medium difficulty/high clinical-relevance and the medium difficulty/low clinical-relevance scales. There was also no significant difference in the distributions of pictorial-stem, clinical-situational, and factual multiple-choice items across the four scales.

These results using the practice-eligible physician group (n=14), who were not considered in any of the subscale construction analyses, withstood a small validation.

It was concluded that the relevance of multiple-choice items to simulated clinical performance is important to the valid statistical discrimination of criterion-group performance.

DEDICATION

In memory of

Florence Downing McClure

To whom I owe much for the person I am becoming

ACKNOWLEDGMENTS

I am indebted to many people and institutions for their assistance in this project. First, I am deeply grateful to Barbara Frederickson, my significant other - for her love, support, and patient understanding of my preoccupation throughout the months of this research, and for her technical assistance in data analysis and proofreading.

I wish to express my appreciation to my dissertation committee for their assistance in this project. I am honored to have had the opportunity during my doctoral program to study with Dr. Robert L. Ebel, who taught me what I know about educational measurement and directed this dissertation. Dr. Jack L. Maatsch, who directs the American College of Emergency Physicians project in the Office of Medical Education Research and Development, "brought me along" as a medical education researcher and provided the ideal empirical climate for this study. Dr. Joe L. Byers, who taught me a great deal about educational research and computers, provided support and friendship throughout my doctoral program and offered much technical assistance for this study. Dr. Leroy A. Olson, who has assisted me many times with test scoring problems, supported this study and offered many helpful suggestions.

I also wish to thank my typist, Marlene Dodge, who showed endless patience with me and the countless drafts of this report. I also appreciate technical assistance for this study from Dr. Pamela W. Wilson, Mr. Douglas Barker, and Ms. Barbara Schachenman.

Dr. Nelson H. Goud, Indiana University School of Education, encouraged my doctoral study in a very special way and supported me throughout with his friendship. Dr. Martha R. Anderson, helped me--through her special friendship, caring, and faith in me--to complete this degree program.

The American College of Emergency Physicians, the American Board of Emergency Medicine, the Office of Medical Education Research and Development at Michigan State University, and the National Center for Health Services Research (HS 02038) have all directly supported this research. I am grateful to many individuals within each of these organizations for their assistance during my four and one-half year association with the Emergency Medicine Examination.

East Lansing, Michigan
May 24, 1979

S.M.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

CHAPTER I

THE PROBLEM

INTRODUCTION

"The competence of physicians to recognize, understand, and manage the problems of their patients is a very critical element in health care" (Senior, 1976). Medical specialty boards, since the early years of this century, have attempted to measure and certify the competence of their candidates to practice in specialized areas of medicine (Hubbard, 1971). Yet, Williamson (1976) states: "Finding evidence of the relation of certification results to actual clinical performance proved to be a difficult task." And also, "....the problem in improving validity of medical specialty certification procedures is serious."

During the past decade, great pressure has been brought on medical specialty certifying bodies to demonstrate the validity of their examination procedures to predict the competence of certified physicians to deliver health care (Williamson, 1976). This issue is so critical today that a conference was devoted to this topic by the American Board of Medical Specialties (Conference on Extending the Validity of Certification, 1976). And, the American Board of Medical Specialties has a standing committee charged with studying the validity of evaluation procedures used by member boards to certify medical specialists.

While some measurement specialists may disagree that cognitive achievement examinations should predict performance (e.g., Ebel, 1961), it is clear that the medical specialty profession, governmental regulatory agencies, and medical consumer groups believe that the certification of a medical specialist should make a difference in his or her clinical performance (Williamson, 1976).

The problem of valid discriminations and predictions of clinical performance is, indeed, a thorny statistical and psychometric one. While great gains have been made in improving the psychometric quality of objective certifying examinations, few gains have been made in establishing valid criterion measures of physician clinical performance (Senior, 1976). Most specialists can not even agree on a behavioral definition of competent medical practice, much less measure this elusive characteristic.

This study will pose some empirical questions about the relative power of objective-item subscales, selected by two different item selection methods, to validly discriminate criterion-groups of subjects who were selected for their known skills in delivering health care in the specialty of Emergency Medicine. Other psychometric characteristics of these scales will be investigated, including a study of item-type contribution to scales selected by different methods.

HISTORICAL BACKGROUND OF CERTIFICATION TESTING

Historically, most specialty boards required candidates to successfully complete two to three years of post-graduate training in a specialty residency program, then to pass an essay examination over

the content knowledge of the medical specialty and a bedside oral examination, in which the candidate examined a hospitalized patient and was rated by a single examiner. By 1946, one specialty board, the American Board of Internal Medicine, introduced objective examinations to replace essay tests and many boards introduced variations of the bedside-oral examination which tended to increase the objectivity of the measurement (Hubbard, 1971).

The recent history of medical specialty certification testing has, in general, included objective examinations, patient management problems, and some type of oral or performance examination. Some boards require candidates to obtain a minimum passing score on the objective and/or patient management problem sections prior to admission to the oral or performance examination. Other boards, and all state licensing examinations, rely solely on objective examinations to certify physician competence. Passing scores are usually determined by referencing examination scores to some norm-group's performance and placing the pass/fail cutting score at some reasonable position on the scale.

The National Board of Medical Examiners in Philadelphia, Pennsylvania, has done more, perhaps, than any other single organization to improve the overall quality of specialty certification testing in the United States and Canada. In its role as consultant to many specialty boards, the National Board of Medical Examiners has moved boards from essay content examinations to high-quality objective tests, has conducted much research aimed toward the improvement of the psychometric qualities of certification examinations, and has aided many boards in the construction, administration, and scoring of their examinations

(Hubbard, 1971). Largely through the influence of the National Board of Medical Examiners, state boards of physician licensing and specialty certification boards have come to place heavy emphasis on the multiple-choice examination for measuring physician competence.

It is clear that testing procedures used to certify competence in medical specialties have improved greatly through the years of this century. The greatest gains in psychometric quality of specialty certification testing have derived from a move to more objective examination methods. But, as noted above, these examination scores do not predict physician clinical performance well (Williamson, 1976).

Essay Examinations

Prior to the 1950's, when the essay examination was used almost exclusively to measure the cognitive competence of candidates, the inter-rater reliabilities of essay scores were found to be very low (r=.50 to .60) in many studies (e.g., Hubbard and Clemans, 1961). Since the introduction of objective-format examinations to replace essay examinations, the internal-consistency reliabilities of most examinations are in the $r \geq .90$ range (Burg and Schumacher, 1979).

Oral Examinations

The oral examination has an even longer history than the essay examination in measuring the competence of physicians. Modern medical education springs from a very long history of treating the training of a physician as an apprenticeship. From ancient through medieval times, even to the present, physician skills are passed from master to apprentice in undergraduate clerkships and graduate residencies. Modern medical curricula usually divide a student's

training into two parts: the first year or two is generally devoted to study of the basic life sciences and the last two years to structured experiences in clinical settings. During the years of clinical training--both during undergraduate clinical clerkship and post-graduate clinical internships and specialty residency training--the oral examination is a highly valued tradition. Students and residents are required to present patient work-ups, during which they are orally examined over the basic and clinical science content appropriate to the particular case.

Another example of oral examination methods used to evaluate physicians-in-training is rounds. Rounds refer to the process of a master physician taking a group of students--residents and medical students--from patient to patient in a teaching hospital and orally questioning individual students about the diagnosis and management of these patients' medical conditions.

Oral Examinations in Certification

It is not, therefore, surprising that specialty certification boards adopted the oral examination as part of their examination procedures. Until the 1950's, specialty boards, in general, required a bedside-type oral examination, in which a candidate was orally examined over cases presented by one or more patients (Hubbard, 1971). As specialty boards became more aware of the psychometric limitations of the oral examination, they tended to modify the oral examination in ways intended to make the measurement more objective, or, to simply abandon the oral in favor of patient management problems. (Patient Management Problems are written, more objectively scored

simulations of a physician's ability to diagnose and manage a patient's problem.) Other boards, like the College of Family Physicians of Canada, have adopted a very structured oral format involving several different types of simulated patient interactions (Handbook for Certification in Family Medicine, 1976).

Construction of Specialty Board Examinations

The test construction methods used by the National Board of Medical Examiners typify the general practice currently used by most specialty boards in the construction of certification examinations.

Committees of specialty content experts meet several times per year to outline the content of the examination, to write and to peer-review items written by their colleagues (Hubbard, 1971). Since nearly all specialty boards interpret the scores yielded by their examinations relative to some norm-group performance, items are written to perform such that extremes of difficulty are avoided as much as possible in order to maximally discriminate levels of achievement throughout the distribution of scores.

After an objective examination is assembled and administered, it is generally scored and item analyzed twice. Items identified as poor discriminators, because they are too difficult or too easy or because of some ambiguity inherent in the wording of the item, are brought to the attention of the decision-making board. Board members generally debate the merits of these questionable items and decide, as a committee, whether to score such items (Hubbard, 1971). Standards for passing are determined by these boards after inspection of the

distribution of scores; passing scores are most often set such that the lower 15 to 20 percent of candidates fail the examination (Hechel and Bowles, 1979).

Objective-examination construction methods like those outlined above tend to produce very reliable and content valid measures. High internal-consistency reliability is achieved by producing items that perform at medium difficulty and, therefore, maximize item and test variance, which tends to maximize the internal-consistency reliability coefficient (Magnusson, 1967). Since content validity is a matter of expert judgment and concensus that the examination measures what it should measure (Standards, 1974), it follows that specialty certification examinations constructed by committees of nationally prominent medical experts in the specialty are, by definition, content valid.

NEED FOR THE STUDY

As noted above, current practice in constructing objective specialty certification examinations produces, for the most part, highly reliable and content-valid measurements. However, these examination scores do not correlate well with independent measures of clinical performance (Williamson, 1976).

Predictive-validity studies, for example, which attempt to find a correlation between scores on a certification examination and some independent measure of the quality of medical care delivered by the examinees, have failed to show any very large validity coefficients (Burg and Schumacher, 1979; Williamson, 1976).

Concurrent with the improvements in the technology of medical specialty examinations, there has been a public press toward assuring

the quality of health care provided by medical practitioners. Consumers of health care have begun demanding that their physicians provide them the best health care possible or, at least, that they have some protection against inadequate or incompetent medical practice. The public and regulatory agencies have also begun to demand that certification examination scores predict the adequacy of physician performance or that passing a certification examination, in fact, makes some difference in the quality of health care delivered by the certified physician.

Many factors may account for the low criterion-related validity coefficients of certification examinations: the reliability and validity of the criterion may be low, the range of scores may be restricted in one or both distributions of scores--thus attenuating the correlation coefficient (Magnusson, 1967), or the examination tasks may simply not be relevant to the tasks measured or rated by the criterion (Maatsch et al., 1978), since an objective examination may measure only one aspect of physician competence.

The test construction methods and philosophy outlined above--writing items to maximize internal-consistency reliability--may tend to attenuate a validity coefficient. For example, items written and selected to be of middle difficulty may be less familiar in content and, thus, less fundamentally necessary to know for the actual, day-to-day practice of clinical medicine, than items written and selected to some other criterion. The selection of middle-difficulty items to maximize item discrimination and examination internal-consistency reliability may distort the content relevance of items to the typical practice of medicine.

In summary, current objective test construction methods tend to produce examination batteries that very sharply discriminate levels of achievement in academic medical content areas which may not be totally relevant to the actual practice of medicine. Objective certification examinations do tend to yield highly reliable scores and are judged as content valid by groups of nationally recognized content experts in the specialty. However, scores from such objective examinations fail, in general, to correlate with other independent measures of the quality of health care delivered. One possible explanation for the failure of objective scores to predict clinical performance may be an inherent lack of content relevance in objective items. If items do lack relevance to the actual practice of clinical medicine, the reason may be that current item analysis criteria used in selecting these items may tend to select items that are less relevant to clinical medicine than they might be.

THE PROBLEM

The purpose of all educational achievement measurement is to discriminate those who know or can do more from those who know or can do less (Ebel, 1972). Medical specialty certification measurement has been viewed, essentially, as achievement measurement--the measurement of how successfully a candidate has mastered the content and skills of the specialty. But, at the same time, the public has come to believe that these measurements should also predict the adequacy of physician clinical performance.

The methods used by specialty boards to construct their objective certification examinations and to select items through item analysis

are taken from the literature on classroom achievement measurement (Hubbard, 1971). These methods, while clearly appropriate for their intended use (Ebel, 1972) may be less appropriate to the measurement of, or the prediction of, a medical specialist's ability to deliver adequate, safe, health care.

This study will compare certification examination subscales--selected by classical item-analysis methods and by an independent measure of item relevance to clinical practice--for item difficulty, reliability, criterion-related validity, and their discriminant validity for criterion-groups with known levels of clinical competence. The proportion of overlap of identical items selected by these two different item selection methods and the effect of pictorial-stem, factual multiple-choice, and clinical-situational[1] test items on the clinical relevance of item content will be examined.

The Emergency Medicine Examination

The Office of Medical Education Research and Development, Colleges of Human and Osteopathic Medicine at Michigan State University, developed, under contract to the American College of Emergency Physicians, a certification examination for the emerging specialty of Emergency Medicine.

This new certifying examination, which took over three years to develop, consists of the following formats:

---

[1]Clinical-situational items present clinical data about a patient--for example, signs and symptoms, laboratory data, and so on--and then ask a question about diagnosis or management. The second example item of Table 3.2 shows a clinical-situational item.

1. Objective: Multiple-Choice and Pictorial Multiple-Choice Items
2. Patient Management Problems
3. Simulated Clinical Encounters: Simulated Patient Encounters and Simulated Situation Encounters

A large library of examination materials was developed and field tested by the Office of Medical Education Research and Development and the American College of Emergency Physicians. A total of ninety-four subjects, representing four criterion-groups with known and different levels of training and experience in Emergency Medicine, were administered all examination materials under a National Center for Health Services Research Grant (HS 02038) in October, 1977 (Maatsch et al., 1978).

The four criterion-groups on whom data were collected are:

1. Residency-eligible physicians
2. Practice-eligible physicians
3. Second-year residents in Emergency Medicine
4. Fourth-year medical students

Item Selection Strategies to Investigate

The relevance of objective-item scores to the adequacy of simulated health care delivered is the major subject under investigation in this study. The effect of objective-item subscale scores, subscales which are operationally defined as high or low on the continuum of relevance to the typical practice of clinical Emergency Medicine, will be compared to subscales selected by the classical item analysis methods (Hubbard, 1971) used by most certifying boards. These empirically defined subscales will be compared regarding their

criterion-group discrimination, internal-consistency reliability, mean item difficulty, and criterion-related validity.

Definition of Clinically Relevant Knowledge

For this study, clinically relevant knowledge is defined as that knowledge which is frequently used and/or has direct utility for the accurate diagnosis and successful management of patients' medical problems as seen in typical clinical situations.

Operational Definition of Clinical Relevance

The clinical relevance of objective-item content will be operationally defined, for this investigation, as that item content which correlates most highly with the grand mean rating of the Simulated Clinical Encounters.

The twelve Simulated Clinical Encounters--eight Simulated Patient Encounters and four Simulated Situation Encounters--consist of highly structured oral simulations of typical emergency patients' clinical problems. A well-trained examiner presents the realistic problem or case to the candidate who works through orally the diagnosis and medical management of the patient or patients. The examiner then rates the candidate's performance at the conclusion of the simulation.

Objective-Item Subscales to Investigate

The objective format of the Emergency Medicine Examination consists, after the deletion of some items following an initial item analysis, of 364 four or five option, single-best answer multiple-choice items. These items sample twenty-three content categories of

Emergency Medicine and were intended to measure the essential knowledge or ability needed for the typical practice of clinical Emergency Medicine (Maatsch et al., 1976).

Four Subscales to Study

The following four 91-item examination subscales will be investigated:

1. Medium-Difficulty Subscale: 91 items selected for item difficulties closest to ideal for norm-referenced achievement tests ($.5 \leq p \geq .7$) and positive point-biserial item-total score discrimination indices.

2. Low-Difficulty Subscale: 91 items selected as having the lowest item difficulties and positive point-biserial item-total score discrimination indices.

3. High Clinical-Relevance Subscale: 91 items selected for their highest correlation with the grand mean rating on the Simulated Clinical Encounters.

4. Low Clinical-Relevance Subscale: 91 items selected for their lowest correlation with the grand mean rating on the Simulated Clinical Encounters.

Subscales one and two will be composed of independent items, as will scales three and four. However, subscales one and two will not necessarily be composed of a set of items that are completely different from the items found in subscales three and four.

RESEARCH HYPOTHESES

IA. H_1: There is a difference in the criterion-group discrimination of the medium difficulty and the high clinical-relevance subscales. The difference favors the high clinical-relevance subscale.

IB. H_1: There is a difference in the criterion-group discrimination of the medium difficulty and the low clinical-relevance subscales. The difference favors the medium difficulty subscale.

IC. H_1: There is a difference in the criterion-group discrimination of the high clinical-relevance and the low clinical-relevance subscales. The difference favors the high clinical-relevance subscale.

II. H_1: There are differences in criterion-related validity between the medium difficulty and the high clinical-relevance subscales. The high clinical-relevance subscale will have a higher validity coefficient than the medium difficulty subscale.

III. H_1: There is a difference in internal-consistency reliability between the medium difficulty and the high clinical-relevance subscales. The difference favors the medium difficulty subscale.

IV. H_1: There are differences in mean item difficulty between the medium difficulty subscale and the high and the low clinical-relevance subscales. The medium difficulty subscale and the low clinical-relevance subscale will be more difficult than the high clinical-relevance scale.

V. H_1: The proportion of overlap of identical items between the medium difficulty and the high clinical-relevance subscales will be lower than the proportion of overlap of identical items between the medium difficulty and the low clinical-relevance subscales.

VI. H_1: There are differences in the distributions of pictorial-stem, clinical-situational, and factual multiple-choice items selected for the four subscales. The high clinical-relevance subscale will have a larger distribution of pictorial-stem and/or clinical-situational items than the medium difficulty or the low clinical-relevance subscale.

SUMMARY

The techniques and methods used to objectively measure the competence of medical specialists has been greatly improved during the last twenty-five years. These improvements include specialty boards' adoption of multiple-choice formats to replace essay examinations and improvement or abandonment of the oral examination. In recent years, there has been increasing pressure both from within the medical specialty professions and from consumers of medical care

to show that specialty certification examination scores predict the quality of health care delivered by the examinee, or, at least, that certified specialists perform more adequately than non-certified specialists. Research in these areas has been largely ignored or has failed, for the most part, to demonstrate the criterion-related validity of certification examination scores.

This dissertation study will evaluate the effect of two different objective-item selection strategies on the validity of the statistical discrimination of groups of subjects with known and different levels of training and experience in the medical specialty and differing levels of competence to deliver health care in the specialty. Classical item selection methods will be empirically compared to the selection of items for clinical relevance, as defined for this study. With independent ratings of simulated clinical performance standing-in for ratings of actual clinical performance, objective subscales selected for their item difficulty or their relevance to clinical medicine will be evaluated for differences in criterion-related validity, scale reliability, and mean item difficulty. The proportions of identical items selected for subscales using the two separate item selection strategies will be evaluated. Finally, the contribution to clinical relevance of two specialized item types--the pictorial-stem and the clinical-situational item--will be assessed.

The theoretical contribution of this research study will be to provide a procedural model for objective certification examination construction that will maximize the valid discrimination of clinical competence.

OVERVIEW OF THE DISSERTATION

Chapter II will review the literature on medical specialty certification testing as this literature relates to the prediction of clinical performance and the educational measurement literature as it relates to item-analysis methods used to maximize examination validity and reliability.

The procedures and methodology used to construct the Emergency Medicine Examination, the design of the field test experiment, the sampling of subjects, and the statistical methods to be used to test hypotheses will be discussed in Chapter III.

In Chapter IV, the results of the data analysis will be presented.

Chapter V will discuss the results of the statistical analyses and present the conclusions resulting from this dissertation study. Additional research suggested by this study will also be noted in Chapter V.

CHAPTER II

REVIEW OF THE LITERATURE

Chapter I stated the need for this study in relationship to a current critical problem facing medical specialty certifying boards. This basic problem--the general lack of criterion-related validity for certification examinations or lack of evidence that passing a certification examination makes much difference in physician performance--will be documented in this chapter.

The reliability and validity of some medical specialty certification examinations will be examined. Two examination construction models--one intended for achievement testing and the other intended for the prediction of performance--will be reviewed. Optimum item-analysis strategies for each type of examination will be reviewed.

CERTIFICATION EXAMINATIONS

A search of the literature on medical specialty certification examinations yields relatively few empirical studies. And, some studies that are reported are of questionable quality, such that conclusions may be of limited value.

The Relationship of Certification Results to Performance Measures

Several studies in the past thirty years have attempted to assess the relationship between physician variables--such as years and type of training, certification examination scores, and so on--and quality of health care delivered by the physician.

One of the earliest studies of correlates to physician performance was conducted for the Teamster Union (Trussell, 1962) in the 1950's. In this classic study, a team of specialists conducted a thorough chart audit of 406 hospital admissions, randomly selected from Teamster members in the New York area. Many aspects of medical care were rated by the team of specialists. This study's major conclusion was that certification status had little relation to the quality of health care delivered by the physicians in the study. The type of hospital--whether teaching or non-teaching--was more highly related to quality of care than the certification status of physician providers of health care. This study was replicated by Morehead and others (1964) and these results were confirmed.

McGuire and Williamson (1968) conducted a study for the American Heart Association in which they compared the performance of three groups of physicians--general practitioners, non-certified, and certified specialists--on three patient management problems. Results of this study showed no statistical differences in the performance of the three groups on the written simulations of physician performance.

Pawluk and others (1976) studied the relationship between scores on various formats of the Canadian Family Practice Certification Examination and physician performance on a measure of quality of care (Kessner, 1973; Sibley, 1975). The sample of subjects for this study was very small (n=15) and, thus, correlations may have been attenuated by large standard errors around r. Pawluk's data showed that the multiple-choice scores correlated -.36 with the measure of quality of care. The simulated office oral scores, correlated .42, while patient management problem scores correlated .25 with the quality of care

measure used. These researchers concluded that multiple-choice examinations were poor predictors of clinical performance.

Gonnella (1973) found low correlations between scores on multiple-choice examinations in Urology and measures of diagnostic accuracy and proper management of urinary tract infections with a sample of certified Urologists.

Some of the best evidence for the lack of criterion-related validity for medical specialty certification examinations has been presented by Beverly Payne and his associates at the University of Michigan. For example, Payne and Lyons (1972), in a large study of the correlates of physician health-care delivery in Hawaii, evaluated the adequacy of physician management of twenty common health problems typically seen in hospital and office practices. The major finding in this study was that board certification, type of specialization, years of practice, and hospital size did not correlate with process audit ratings of physician performance. However, in some specialized areas of medical practice--Pediatrics, Surgery, Internal Medicine--years of experience in the specialty did predict clinical performance; the board certification of the specialists, however, did not predict their clinical performance. Payne found only two areas in which certified specialists performed statistically better than non-certified specialists; in both of these cases, the barely significant effect seemed confounded by a disordinal interaction effect.

Rhee (1975) reanalyzed the Payne Hawaii data for a doctoral dissertation. These findings show no statistical differences in the performance of board-eligible and board-certified physicians. However, Rhee found that board-eligible and board-certified physicians performed

much better than self-claimed specialists when they were practicing in their own specialized areas.

Several other studies have examined the relationship between certification examination scores and faculty ratings of student or resident clinical competence (Hubbard, 1971; Kaplan, Freeman, and Kaplan, 1968; Kelley and Levit, 1967; Kelley, Stumpe, and Levit, 1970; Schumacher, 1964). These studies have shown low positive correlations of examination scores with faculty ratings of overall clinical performance.

None of the studies reviewed here allow a comparison of the multiple-choice content measured, its relevance to clinical medicine, or item types employed.

In summary, the literature on the relationship between board certification and physician clinical performance suggests:

1. Certification examination scores have low correlations with criterion measures of physician performance.

2. Type and length of formal postgraduate training--residency education--do correlate with subsequent measures of quality of medical practice.

Psychometric Characteristics of Some Certifying Examinations

To place the present research study on the Emergency Medicine Examination in context, it seems appropriate to review published studies on the reliability and validity of some other specialty certifying examinations. As in the previous section which reviewed the criterion-related validity of certification examinations, there are relatively few published empirical studies and they are of mixed quality.

Validity

Hubbard (1971) shows that the National Board of Medical Examiners test construction methods assure the content validity of specialty examinations from the National Board. These objective examinations, as outlined in Chapter I, are constructed by committees of experts in the specialized content. Predictive validity studies for National Board certifying examinations have failed to show any large correlations with ratings of clinical performance (Burg and Schumacher, 1979). Burg, Guerin, and Schumacher (1977) and Levine, McGuire and Nattress (1970) show some construct validities for various National Board certifying examinations. That is, these studies demonstrate that certain National Board certifying examinations yield scores that are sensitive to years of training in some specialties.

Maatsch and others (1978) have shown that the Emergency Medicine Examination yields scores, in both the Objective and Simulated Clinical Encounter formats, that were sensitive to years of training and experience in Emergency Medicine. This same study also showed the concurrent validity of the Emergency Medicine Examination in that the total Objective score is correlated with the grand mean rating on Simulated Clinical Encounters .83.

On the other hand, correlations of objective scores and ratings on an oral examination for an American Board of Orthopaedic Surgery certifying examination were .29 overall (Levine and McGuire, 1970). And, the highest concurrent validity coefficient reported by Kelley and others (1971) for the American Board of Anesthesiology Examination was $r = .54$, for scores that had been corrected for attenuation due to the unreliability of both objective and oral rating scores.

Reliability

The reliability of examination scores is defined by Ebel (1972) as "....the consistency with which a set of test scores measures whatever it does measure." Many test specialists (e.g., Mehrens and Lehmann, 1973) suggest that test reliability is the single most important index of overall examination quality.

In general, for objectively scored medical specialty certification examinations, Burg and Schumacher (1979) state that internal-consistency reliabilities are greater than r = .90. Oral examination formats, on the other hand, tend to have much lower reliability (Burg and Schumacher, 1971). Accordingly, most empirical research reported in the literature has dealt with the reliability of the Oral examination format.

Since different researchers tend to use different methods of calculating oral examination reliability, the studies reported here may not be exactly comparable. It should be noted in this context that the appropriate reliability to report for oral examinations is the inter-rater reliability coefficient; inter-rater agreement is most accurately and efficiently assessed by the interclass inter-rater reliability coefficient (Ebel, 1951a).

The generally low inter-rater reliability of oral examinations is well documented (e.g., Ebel, 1972; Mehrens and Lehmann, 1973). Yet despite much evidence for the errorfulness of oral examination ratings, medical specialty boards have used this examination format from the beginnings of the certification movement. It is interesting to note that in the classic study by Levine and McGuire (1970), which concluded that oral examinations measure something quite different than

objective examinations, the inter-rater reliability of the oral was only r = .50.

In an earlier study, McGuire (1966) reported substantial rating disagreements for oral certifying examinations, although she did not report a coefficient of rater agreement.

Other oral certifying examinations report much higher rater-agreement coefficients. For example, Carter (1962) studied the rater-agreement for the oral format of the American Board of Anesthesiology examination and found an agreement coefficient of r = .80.

For the Emergency Medicine Examination, Maatsch and others (1978) report high interclass correlation coefficients for the oral Simulated Clinical Encounters. For twelve field test cases, the inter-rater reliability coefficients for individual ratings ranged from r = .63 to .89 with an average r for all problems equal to .79. Raters for this study were, however, carefully trained to an objective rating criterion. The inter-rater reliabilities for the oral Simulated Clinical Encounters compare favorably with the internal-consistency reliability coefficients for the objective formats of the Emergency Medicine Examination. The Kuder-Richardson 20 reliability for the total pool of 103 pictorial-stem items was .89; for 261 multiple-choice items, the reliability was .94; and, for the total library of 364 objective items, the reliability was .96.

In summary, the available empirical research shows that medical specialty certifying examinations:

1. Have high internal-consistency reliability for objective formats.

2. Have low inter-rater reliabilities for oral examinations, unless this format is highly structured and raters are well trained.

3. Have low between-format correlations, from which it has been concluded that different examination formats measure different aspects of physician competence.

EXAMINATION CONSTRUCTION MODELS

Examination construction specialists have understood for years that different test construction methods are appropriate for different intended uses of test scores. For example, the test construction techniques and item selection strategies that are most efficient for classroom achievement testing may be less efficient for building examinations to validly predict successful job performance.

Achievement Versus Aptitude Test Construction

Ebel (1951b; 1956; 1967; 1972) has carefully and completely documented the most appropriate methods to use in constructing achievement examinations. These methods may be briefly summarized by the following propositions:

1. Carefully detailed test content yields content valid measurements.

2. Objective items that present novel questions or problems tend to test student understanding of the relevant and important concepts learned.

3. The achievement test may be the best operational definition of the subject content available.

4. Items of medium difficulty yield internally consistent measurements of student achievement.

5. The upper-lower (D) discrimination index, biased toward items of middle difficulty, tends to select the most efficient achievement test items.

Most test construction specialists would agree that these methods will yield valid, reliable, and objective measurements of student achievement. Henrysson (1971), for example, suggests that the point-biserial or the biserial item-total score correlation coefficient be used as an item discrimination index to select achievement test items that maximize item discrimination and test reliability. Cronbach (1951) shows that Coefficient Alpha--Kuder-Richardson 20--is the most appropriate index of internal-consistency reliability.

Procedures recommended by Ebel (1972) and most other test construction specialists tend to produce internally consistent measures of achievement that are appropriate for norm-referenced score interpretation.

While aptitude test constructors may write items that look just like achievement test items, item analysis selection strategies may be quite different. For example, if the purpose of an examination is to predict some future complex performance or status, it may be statistically beneficial to write heterogeneous test items (Guion, 1965). The logical extension of this test construction methodology may be found in non-cognitive measurement, especially in empirically-keyed instruments (Mehrens and Lehmann, 1973) wherein items are selected for a scale solely for their empirical correlation with some behavior in some sample of subjects. This strategy--to maximize criterion-related predictive validity--may be the complete opposite of the content validity strategy employed by achievement testers (Ebel, 1972). That is, items chosen to maximize predictive validity may have little or no content validity and low internal-consistency reliability (Guion, 1965).

The personnel testing situation is perhaps the best example of differences between the test construction methods of achievement and "prediction" testing. In achievement testing, well written, content valid, discriminating items of about middle difficulty will yield scores that rank-order students in accordance to their mastery of the content measured; the goals of the measuring process--to mark student achievement, validly--are accomplished. In personnel testing situations--where the goals of measurement may be to predict some future performance or to validly sort groups of subjects according to some psychological trait--test items may have little or no content relevance, but items must have empirical predictive power for the criterion of interest. Accordingly, the personnel test constructor may use some external criterion against which to correlate item scores or use multiple regression, and/or discriminant analysis techniques (Guion, 1965) to identify the most efficient items for the final form of the examination. This final form of the personnel test may have little content validity, but will likely have a very high criterion-related validity coefficient; this final form may also have a rather low internal-consistency reliability coefficient.

Magnusson (1967) points out that the prediction of future complex performance may require a test which is composed of several subtests. These subtests--to be maximally efficient--would be highly internally consistent, but would have low inter-correlations with other subscales.

Ebel (1961; 1978) has pointed out that test validity is much more a characteristic of test use than of the test itself. Measurements yielded by tests must be valid--but valid for what purpose? Is it

perhaps unreasonable to demand that a given test be both content valid and predictive of performance?

If the lack of criterion-related validity noted above for most certifying procedures is a serious problem as Williamson (1976) states, then certifying bodies must decide, more clearly than in the past, what their goals of measurement are. Is the purpose of certification testing the measurement of cognitive achievement in specialized medical content? If so, the literature on achievement testing noted here is relevant. If, however, specialty boards decide that their purpose is to protect the public from incompetent and dangerous medical practice, the literature on minimum competency testing is relevant. If boards decide that the prediction of future clinical practice is the most essential goal, then the literature of personnel and aptitude testing may be of interest.

This study can not answer the philosophical questions posed, but will attempt to address the empirical questions concerning the effectiveness of two different item selection strategies for validly and reliably discriminating groups of subjects with known levels of ability to deliver health care.

SUMMARY

Few high-quality empirical studies have been reported in the area of medical specialty certification. No studies have been reported that relate directly to the exact problem being investigated here. However, a review of the available literature on certifying examinations and two models of measurement reveals the following:

1. There is little or no evidence that scores on certification examinations in medical specialties predict the quality of medical care delivered by candidates.

2. Post-graduate residency training does predict the quality of clinical performance.

3. Objective formats of certifying examinations tend to be highly internally consistent.

4. Oral certification examination formats tend to have low inter-rater reliability coefficients, unless the oral is standardized and the raters are well trained.

5. Certifying examination formats--multiple-choice and oral--tend to have low correlations, unless multiple-choice items are written to be relevant to clinical medicine.

6. Specialty boards have not clarified the purposes of their certification measurements. If the purpose is to grant a certificate of excellence to masters of specialty content, achievement testing methods may be appropriate. If the purpose of certification testing is the valid prediction of some future clinical performance, aptitude or personnel testing methods may be appropriate.

CHAPTER III

PROCEDURES AND DESIGN

INTRODUCTION

The purpose of this research is to compare subscales selected for item difficulty and clinical-relevance for their psychometric quality in a medical specialty certification examination. Four objective-item subscales will be identified--two subscales of items selected for an item-difficulty criterion and two subscales of items selected for an external criterion of item correlation with performance on realistic clinical simulations. These four objective subscales will be compared for mean item difficulty, internal-consistency reliability, criterion-related validity, and ability to discriminate three groups with known levels of training, experience, and ability to deliver health care in the medical specialty. Further, the proportion of overlap of identical items selected by different item selection strategies will be examined. Finally, the contribution of three different objective item types--the pictorial-stem, clinical-situational, and factual multiple-choice item--to clinical relevance, as operationally defined for this study, will be examined.

This chapter includes a description of the sampling plan and rationale used to select subjects for this study, the details of examination construction for both the objective items and the clinical simulations, the design of this study, the hypotheses to be tested, and the statistical procedures to be used to test the hypotheses for this research.

SAMPLE OF SUBJECTS

A total of ninety-four subjects participated in this study. These subjects were chosen to represent four distinct groups on the dimensions of known years of training in Emergency Medicine and years of experience in practicing Emergency Medicine. The four groups of examinees were:

1. Residency-Eligible Emergency Physicians:

   n=22 subjects who were eligible to take a certification examination by virtue of graduation from an approved Emergency Medicine residency program and continuous practice in Emergency Medicine for a minimum of one year.

2. Practice-Eligible Emergency Physicians:

   n=14 subjects who were eligible to take a certification examination by meeting the requirement of five years of continuous practice in Emergency Medicine.

3. Residents in Emergency Medicine:

   n=36 subjects who were beginning their second of three years of residency training in Emergency Medicine.

4. Medical Students:

   n=22 subjects who were beginning their fourth year of pre-doctoral clinical study.

The total number of subjects selected for this study (N=94) was restricted due to the high cost of subject acquisition. The original plan was to have approximately one-hundred subjects equally divided between three groups (physicians, residents, and students), such that large-sample ($n > 30$) statistics could be used for inter-group comparisons. The final sample of subjects, as detailed below, fell considerably short of the original goal due to the programmatic constraints of subject acquisition, and subject fee and travel limitations.

Different criteria were used to select groups of subjects for this study. Sampling procedures used for each group are detailed below.

Residency-Eligible and Practice-Eligible Emergency Physicians

The American Board of Emergency Medicine was constituted in March, 1976. This new medical specialty board is ultimately responsible for graduate medical education in Emergency Medicine and the certification of specialists in Emergency Medicine. In its role as a certifying body, the Board sets certain minimum prerequisites of training and/or experience in Emergency Medicine for those who wish to take the Examination. The Board, recognizing the newness of the specialty and the short history of residency training in Emergency Medicine, has set two separate prerequisite paths to qualify for its certification examination. These two paths are: residency training in Emergency Medicine in one of its approved programs or five years of continuous practice in hospital emergency departments.

Selection of Residency-Eligible and Practice-Eligible Emergency Physicians

A combination of a peer-nomination method and a random sampling plan was used to select two groups of emergency physicians. Each group was intended to be representative of the residency-eligible and practice-eligible applicants for the examination who were clearly competent in clinical, diagnostic, and patient management skills and who were, therefore, certifiable as competent specialists in Emergency Medicine. Accordingly, the first step involved a request by the American College of Emergency Physicians to all its state affiliates for nominations of members to sit for a field test of this examination.

The criteria that were to be used for peer-nominations were personal knowledge that the nominee:

1. Provides very competent health care in the emergency department setting.
2. Maintains current knowledge of clinical, diagnostic, and patient management procedures.
3. Is eligible for the certification examination either by residency training or years of practice in Emergency Medicine.

A total of 151 emergency physicians, from throughout the United States and Canada, remained on a nomination list after a credential review by the American Board of Emergency Medicine. Thirty-two nominees had residency-eligibility and 119 had practice-eligibility. The relationship between the numbers of nominees in the two eligibility groups is roughly proportional to the percentage of membership in the American College of Emergency Physicians for each group.

The second step was to select a total of thirty-six field test subjects from these nominees. A total of twenty-two residency-eligible physicians (with ten alternates) and fourteen practice-eligible physicians (with eighteen alternates) was selected by a simple random sample of two separate nomination lists. The sample was deliberately skewed in favor of the residency-eligible physicians because of a belief that this group represented more clearly certifiable physicians then the practice-eligibles.

The emergency physician participants were reimbursed for travel and per diem expenses for their participation in the field test.

Selection of Residents in Emergency Medicine

The selection of second-year residents in Emergency Medicine also employed a two-step selection plan. The first step in the selection

of this group of thirty-six subjects involved the American College of Emergency Physicians' request to the residency directors of all twenty-four residency programs in the United States to submit a rank-ordered list of their second-year residents. This listing ranked every resident in each program from highest to lowest with respect to relative overall clinical competence in Emergency Medicine. This procedure was intended to ensure a final sample of residents that would be representative of the range of competence of second-year residents in Emergency Medicine.

The second step in selecting residents was the random selection of thirty-six subjects and alternates. This sampling was carried out by drawing random samples from each of the twenty-four residency programs in the following manner:

1. Random samples were drawn that were strafified on high, middle, and low ranges of competence within each residency program.

2. Random samples were drawn from each program such that the number of subjects selected was roughly proportional to the size of the program.

The thirty-six residents who participated in this study received a subject fee and were reimbursed for travel and per diem expenses.

Selection of Medical Students

Medical students beginning their final year of undergraduate medical education were from Michigan State University's Colleges of Human and Osteopathic Medicine. These paid-volunteer students were recruited to represent a novice or base-line group in training and and experience in Emergency Medicine. This group was not selected to be representative of fourth-year medical students at Michigan State University.

The initial procedures followed in obtaining these subjects were:

1. Random selection of subjects and alternates from each of six Michigan communities where students receive clinical training.

2. Invitations to students so selected and their alternates to participate in the field test.

This group proved to be the most difficult to obtain. The time of the field test conflicted with the clinical clerkship schedules of many students, who consequently could not serve as subjects for this study. Random selection had to be abandoned ultimately in the interest of simply obtaining sufficient numbers of student-volunteers for the study. The final group of twenty-two student subjects who participated in this study were, then, paid volunteers whose clerkship schedules permitted their participation.

Students received a subject fee, plus travel and per diem expense reimbursement. Students also received feedback on their examination performance in terms of raw and percent-correct scores and percentile ranks, based on their own group, for all Examination formats and some content categories.

EXAMINATION CONSTRUCTION

The development of the examination materials for the Emergency Medicine Examination[2] took place from January 1975 to August 1977. All items were generated by content expert members of the American

[2]The data, test formats, scoring mechanisms and all related examination development and validation procedures described in this dissertation were developed for the American College of Emergency Physicians. The American Board of Emergency Medicine, which will subsequently administer the first certification examination, reserves the right to use all or part of the test library and methodologies developed by the American College of Emergency Physicians.

College of Emergency Physicians and test construction specialists of the Office of Medical Education Research and Development, Michigan State University. The following pages will describe in detail the examination construction methods and procedures used to develop the Objective and the Simulated Clinical Encounter formats of the Emergency Medicine Examination.

Overview

The major steps involved in examination construction were:

1. Definition of the exact and proper content of Emergency Medicine by the American College of Emergency Physicians.
2. Identification of Emergency Medicine content to test, and rank-ordering of this content by its importance to test, leading to a test blueprint.
3. Development of detailed content statements on which examination materials would be based: condition sheets.
4. Assignment of item quotas to specialized task forces of American College of Emergency Physician item writers.
5. Training of American College of Emergency Physician item generators by the Office of Medical Education Research and Development.
6. Item writing, review, editing, and production.

These procedural steps culminated in the field testing of all Emergency Medicine Examination items on October 22-26, 1977, in Lansing, Michigan, with the sample of subjects noted above.

Definition of Content

The first stages of examination development required the identification and rank-ordering by importance of the content universe of Emergency Medicine. The American College of Emergency Physicians had worked prior to 1975 to gain the concensus of a certification task

force on a six-page listing of skills needed to practice Emergency Medicine and the medical conditions about which emergency physicians needed content knowledge. This Emergency Medicine Condition/Skills List (Condition/Skills List, 1976) represented the best definition of Emergency Medicine available by defining the domain of content knowledge and psychomotor skills that the emergency physician needed to have and the medical conditions about which the emergency physician needed information.

The second step toward operationalizing the definition of Emergency Medicine in an examination required the prioritization of this Condition/Skills list by a sample of the American College of Emergency Physicians certification task force members. The prioritizing of list entries was accomplished by administering a questionnaire to approximately one-hundred task force members. Each respondent marked each entry as either essential to test in a certification examination, important to test, unimportant to test, or necessary pre-condition not to be tested. The final task of each respondent was to allocate one-hundred percentage points to twenty-two broad content categories of Emergency Medicine so that the most important category received the highest percentage allocation.

These questionnaire data yielded a consensus of Emergency Medicine specialists about the proper content to test in a certification examination and the proper balance in which to test this content. The table of specifications or the test blueprint which guided the construction of the Examination followed directly out of these procedures. Summary results of the percent allocation procedure are given in Table 3.1.

TABLE 3.1

TEST ITEMS ALLOCATED TO MEDICAL CONTENT CATEGORIES

Percentage	Category
13	Cardiovascular disorders (traumatic and nontraumatic)
7	Abdominal disorders
7	Ear, nose, throat, head and neck injuries (traumatic and nontraumatic)
7	Pulmonary disorders
7	Skeletal injuries
7	Traumatic disorders
7	Urogenital disorders
6	Infancy and childhood disorders
5	Metabolic, allergic and toxicologic disorders
4	Fluid and electrolyte problems
4	Neurological disorders
3	Burn and cold exposure
3	Critical infections
3	Emergency medical services system (including disaster planning and management)
3	Eye disorders (traumatic and nontraumatic)
3	Legal-Ethical
2	Blood disorders
2	Physician/Patient skills
2	Emergency department administration
1	Dental emergencies
1	Integumental disorders
100%	

The next pre-examination construction procedure involved expanding the entires on the Condition/Skills list into content materials from which examination items could be generated. The Office of Medical Education Research and Development adopted a condition-sheet method whereby the American College of Emergency Physician task force members would complete very detailed outlines for every entry on the Condition/Skills list. The analogy of a textbook on Emergency Medicine was adopted, such that the twenty-two content categories listed in Table 3.1 became the major chapter headings and individual conditions, skills, and knowledge became major subdivisions within chapters.

Each condition sheet listed very important or essential knowledge or skills needed by a competent emergency physician for each entry on the specialty defining list. For each medical condition, typical presenting signs and symptoms, diagnostic and medical management problems frequently encountered, common errors made in diagnosing and/or managing this condition, plus complete medical references were listed in detail.

Condition-sheet writing, review and editing took nearly one-hundred emergency medical leaders approximately one year to complete. The product resulting from this task represents an encyclopedia of medical knowledge and essential skills needed for the competent practice of Emergency Medicine.

The American College of Emergency Physicians next organized five task forces of medical experts for the purpose of examination construction. These task forces were:

1. Cardio-Respiratory Task Force

2. Medicine Task Force

3. Surgery-Trauma Task Force

4. Physician-Patient Task Force

5. Administration-Systems Task Force

Item[3] quotas were assigned to each task force in accordance with the proportions noted in Table 3.1. Separate procedures were used to develop each examination format. The methods employed to construct the Multiple-Choice, Pictorial Multiple-Choice, and Simulated Clinical Encounter formats will be detailed below.

Multiple-Choice Items

A total of 372 multiple-choice items were written by the American College of Emergency Physician task force item writers for the Emergency Medicine Examination. These items required the selection of one best answer from among four or five options. Table 3.2 presents non-secure examples of the type of multiple-choice items used in this examination.

Emergency physician item writers were trained to write and review multiple-choice questions in a series of workshops. These workshops presented the basic principles of good item writing through a series of instructional materials (Downing, 1977), following closely the work of Ebel (1972); examples of well written and poorly written items were given. Task force writers then practiced writing items, had these items reviewed by a fellow physician and by a test construction specialist.

---

[3]Item is used in its widest meaning here to include not only Multiple-choice questions, but also Patient Management Problems and Simulated Clinical Encounters.

TABLE 3.2

EXAMPLE MULTIPLE-CHOICE ITEMS

Which of the following is characteristic of a normal overnight dexamethasone suppression test?

A 24 hour urinary 17 OH falls 50%

B 24 hour urinary 17 KS rises 50%

C plasma cortisol rises 50%

D plasma cortisol falls 50%

E plasma cortisol remains the same

A 35 year old female is seen in the emergency department in a comatose state. Arterial blood gases and serum electrolytes are drawn and reveal the following results: sodium 140, potassium 4.9, chloride 98, bicarbonate 10, pH 7.30, and pCO_2 24 mm Hg. Which of the following would most likely be the correct diagnosis?

A metabolic acidosis--ammonium chloride overdose

B metabolic acidosis--renal tubular acidosis

C metabolic alkalosis--duodenal fistula

D respiratory acidosis--primary

E metabolic acidosis--diabetic ketoacidosis

Which of the following procedures would give the closest index of the risk of intrauterine death for a fetus with erythroblastosis fetalis?

A direct Coomb's test of mother's blood

B indirect Coomb's test of mother's blood

C spectrophotometric analysis of amniotic fluid

D direct Coomb's test of RBC's in amniotic fluid

E spectrophotometric analysis of mother's blood

Guidelines to item writers (Maatsch et al., 1976) for the selection of item content included:

1. Frequently used general rules or principles.
2. Absolutely essential knowledge for competent emergency department practice.
3. Specific applications of knowledge to clinical Emergency Medicine.
4. Knowledge that must be remembered at all times for competent practice.
5. Frequently encountered cases and problems.

At these workshops, physician item writers received quotas of items and condition sheets from which the item content was generated. Each item written was sent to a physician reviewer for comments and criticisms, and then returned to the item author for revisions. After all items had been written, an Audit Committee of the American College of Emergency Physicians reviewed each item for content and keying and testing specialists reviewed and edited all items for form.

Pictorial Multiple-Choice Items

The Pictorial Multiple-Choice format of this Examination consisted of 136 pictorial-stem items. This item type presented some visual stimulus--an electrocardiogram rhythm strip, a color photograph of a patient, and/or a high-quality photoreduction of an x-ray--and one or more multiple-choice items based on these visual stimuli. Like the multiple-choice items, the pictorial multiple-choice questions were of the single best answer type and had four or five options.

The procedures used to construct pictorial-stem items were essentially the same as for multiple-choice items. Separate workshops,

however, were conducted to train item writers and additional criteria for item content selection were used.

Criteria for selection of visual materials and item content for this format included (Maatsch et al., 1976):

1. Visuals that test general interpretive skills
2. Visuals that typically require immediate interpretation and use in an emergency department
3. Visual materials that knowledgeable candidates can clearly see and interpret

Simulated Clinical Encounters

Twelve Simulated Clinical Encounters were developed for this Examination. Simulated Clinical Encounters are examiner-administered, highly structured simulations of realistic emergency medical problems typically seen in hospital emergency departments. The simulations developed for this Examination are of the patient-game type (Maatsch, 1974) in which a well-trained examiner presents pre-designed and standardized information about a patient, when such information is requested by an examinee, who then proceeds to diagnose and manage the patient being simulated. Realistic patient presenting signs and symptoms are described to the examinee, who may, for example, order laboratory studies, x-rays, electrocardiograms, and so on, to aid in the differential diagnosis of the patient. If laboratory studies are ordered, the examiner provides the results to the examinee at the time these data would be available during a real clinical encounter.

The Simulated Clinical Encounters are structured and standardized on a patient-game board which precisely details all data which is given, if requested by the examinee, and all examiner responses to

examinee actions. The simulated case presentation follows a logical and realistic course in which oral descriptions of the simulated patient's condition are contingent on the actions of the examinee. For example, if a patient's medical condition is deteriorating and the examinee orders a certain drug to be administered, the patient's changed condition will be reflected in data provided to the examinee. All such examiner responses to examinee actions are listed on the patient-game board which directs the administration of the simulated case.

Two separate types of Simulated Clinical Encounters were developed for the Emergency Medicine Examination. The first type, Simulated Patient Encounters, requires the examinee to manage a single simulated patient case during a fifteen minute time period. The second type of Simulated Clinical Encounter, the Simulated Situation Encounter, presents three medical cases which the examinee must manage concurrently; thirty minutes are allowed for each Simulated Situation Encounter.

The total of twelve Simulated Clinical Encounters were divided between eight Simulated Patient Encounters and four Simulated Situation Encounters. These simulations were developed by American College of Emergency Physician task force members who were teamed with educational developers from the Office of Medical Education Research and Development, Michigan State University. Prior to developing the Simulated Clinical Encounters, a total of sixty scenarios, story lines of emergency medical cases, were created from entries on the Condition/Skills list. These scenarios were prioritized by a task force of the American Board of Emergency Medicine to

ensure the proportionate sampling of the content categories (Table 3.1), the relevance of the scenario to the typical practice of Emergency Medicine, and their suitability for production as a Simulated Patient Encounter or a Simulated Situation Encounter. Each Simulated Clinical Encounter was pre-tested with an emergency physician task force member prior to final production of the case.

DESIGN

This section will present the details of administration of the Emergency Medicine Examination materials to the ninety-four subjects on October 22-26, 1977, in the Hilton Inn, Lansing, Michigan.

Subject Groups

Two administrative sections of twenty-three subjects and two sections of twenty-four subjects were formed for purposes of movement through a master schedule of testing. Subjects were randomly assigned to testing sections in proportion to the numbers in each of the four subject groups represented in this study. Subjects were assigned random identification numbers which were used to identify all responses to examination items. The four testing sections remained together throughout the twenty-two hours of testing, taking meals and breaks together to maintain isolation from all other examinee sections in order to ensure test security.

A complicated master schedule was developed to move administrative testing sections and individual subjects through all formats of this examination. Sections were randomly assigned to formats in order to avoid any testing order-effect on criterion-group scores.

Multiple-Choice and Pictorial Multiple-Choice Formats

Multiple-choice items were presented in four booklets of ninety-three items each; pictorial multiple-choice items were divided between two books of sixty-eight items each. Items were assigned to booklets for both formats by a random procedure that forced approximately proportional representation of items from each content category listed in Table 3.1 in each booklet.

Examinees answered all items on optically scanable answer sheets for computer scoring and item analysis. Each test booklet presented thorough instructions to subjects with example items; test administrators followed a standard set of instructions for each booklet administration. Each examination session was proctored by a test administrator and two assistants, with each session timed to allow exactly one minute per objective item.

Pictorial Multiple-Choice items were presented in special booklets that contained, on opposite pages, both the visual stimulus and the items related to the visual. Subjects, thus, had original visual materials--rather than printed reproductions--to examine for each pictorial question. These Pictorial Multiple-Choice booklets were reused after a thorough check for markings.

Simulated Clinical Encounters

Twenty-four emergency physician examiners conducted the Simulated Patient Encounters and Simulated Situation Encounters. Examiners had spent ten hours in training immediately prior to administering the Simulated Clinical Encounters. Developers did not administer their own cases.

Simulated Clinical Encounters were administered in small booths which had been subdivided from a large hotel ballroom. The examiner and examinee were seated across a table from each other, with the Simulated Clinical Encounter game board between them.

Sections of subjects moved through the twelve Simulated Clinical Encounters according to the master schedule and individual subject schedules. A time keeper signaled the beginning and end of each fifteen-minute Simulated Patient Encounter and each thirty-minute Simulated Situation Encounter. During every two hour Simulated Clinical Encounter block, sixteen examiners worked, with six examiners free for rest or for pairing with other administrators to verify examiner ratings. This verification was undertaken to study the inter-rater reliability of the Simulated Clinical Encounters. Second raters were randomly assigned to observe approximately twenty-five percent of Simulated Clinical Encounter administrations and independently rate examinee performance. Inter-rater reliabilities for individual ratings, computed by the interclass formula, ranged from $r = .63$ to .89 for the twelve Simulated Clinical Encounters. The inter-rater reliability of the grand mean rating on Simulated Clinical Encounters was .79 (Maatsch et al., 1978).

Examiners completed a rating form on each candidate immediately after each Simulated Clinical Encounter session. Seven separate clinical skills were rated on an eight-point scale for each case presented. It should be noted that, because of the method of examinee identification used, examiners had no knowledge of the criterion-group membership of individual subjects.

Generalizability of Results

Results of this study may be generalized to the population of emergency physicians judged by their peers to be very competent and certifiable in Emergency Medicine. For the resident sample, results may be generalized to the population of second-year residents in Emergency Medicine who have a wide range of competence as judged by their residency directors. Since the sample of students is a sample of convenience, only limited generalizations should be made to the population of fourth-year medical students at Michigan State University.

The matter of generalizability of results to specific populations of subjects is, however, not of primary concern in the present study. Since the goal of this study is to determine the relationship of item content relevance to clinical performance and the relationship of item difficulty to clinical relevance, the generalizations of most interest concern inferences about item content and types of the valid discrimination of the clinical competence of criterion-groups with known levels of clinical competence.

Subscale Development

As noted in Chapter I, four subscales will be identified for this study. Items for two of these subscales--the medium difficulty and the low difficulty subscales--will be identified using an item difficulty criterion. Items for two other subscales--the high clinical-relevance and the low clinical-relevance subscales--will be selected by using an item-criterion score (grand mean Simulated Clinical Encounter rating) correlation criterion. These four subscales are schematically diagrammed in Figure 3.1.

SCHEMATIC OF 91 ITEM SUBSCALES TO INVESTIGATE

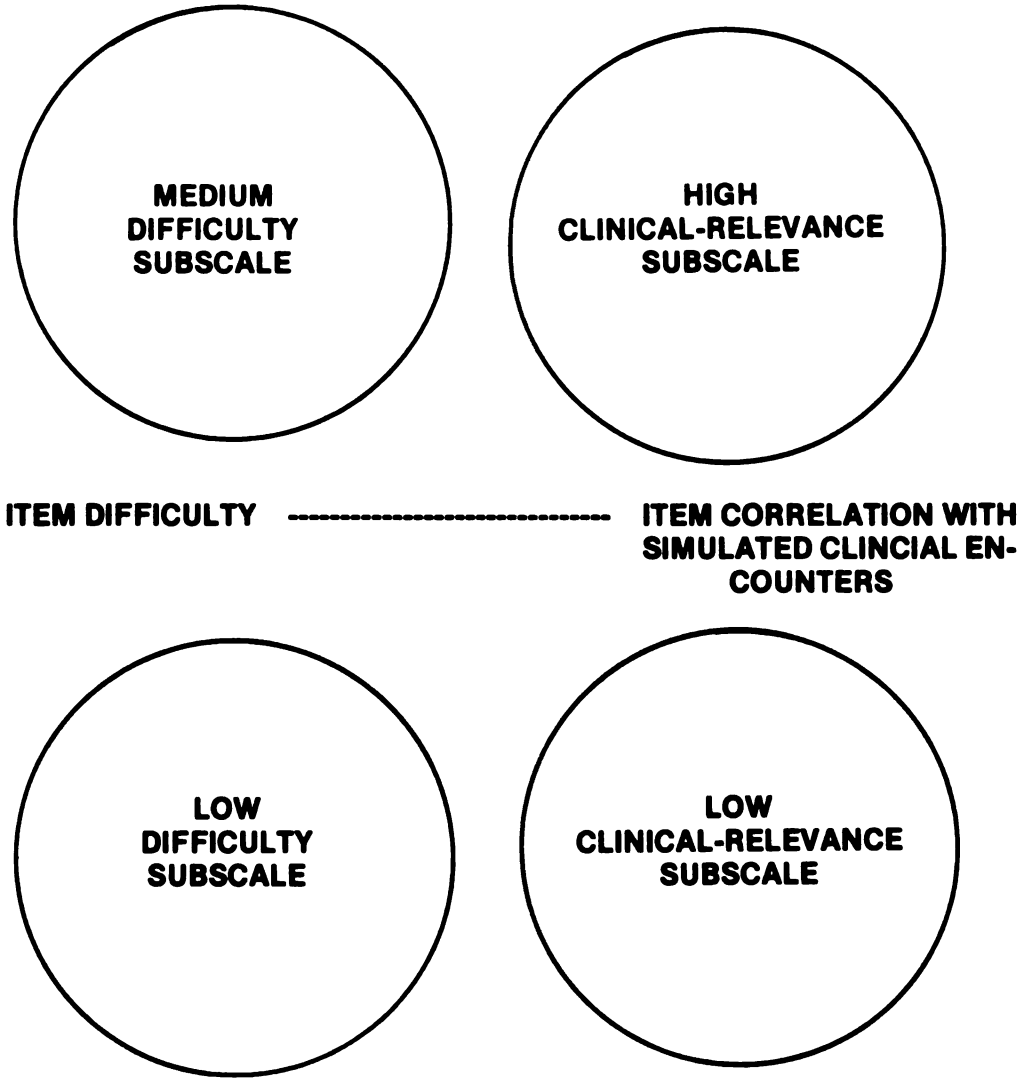

Figure 3.1

A major purpose of this research is to study the effect of item selection rules on the valid discrimination of clinical performance. The logic underlying the testable hypotheses for this study can be summarized by the following questions:

1. Does the selection of items for medium difficulty tend to distort the clinical-relevance of item content?

2. What is the effect of any such distortion on the valid discrimination of groups with known levels of clinical competence?

3. Do different item types vary with respect to clinical relevance?

The ninety-one items for each of the four subscales will be identified, using the criteria noted in Chapter I. Ninety-one items for the medium difficulty subscale and ninety-one items for the low difficulty subscale will be selected from item analysis data; subscale scores consisting of the sum of correct responses will be computed. Next, all 364 items will be correlated with the grand mean rating on Simulated Clinical Encounters; these items will then be rank-ordered and the high clinical-relevance and the low clinical-relevance items will be identified. The two clinical-relevance subscale scores will then be computed such that each subject's subscale score is the sum of the number of correct responses to the items in the scale.

All data analyses to be carried out for the hypotheses of this study will use only three criterion groups of subjects: the residency-eligible, resident, and student groups. Practice-eligible physicians (n=14) will not be considered in any of the analyses for subscale development or hypothesis testing, because--of the two physician groups--there is greater confidence that the residency-eligible group

represents true competence in Emergency Medicine. It is felt, therefore, that for the hypotheses to be tested in this study, clearer results will be obtained by omitting the fourteen practice-eligibles from all analyses. Omission of the practice-eligible group from subscale development analyses will also allow for a small validation of the results of this study by reanalysis of some hypotheses using only the fourteen practice-eligibles or the practice-eligible group in combination with residents and students.

HYPOTHESES AND ANALYSIS METHODS

IA. H_o: There is no difference in the criterion-group discrimination (residency-eligible, residents, students) of the medium difficulty and the high clinical-relevance subscales.

H_1: There is a difference in the criterion-group discrimination of the medium difficulty and the high clinical-relevance subscales. The difference favors the high clinical-relevance subscale.

IB. H_o: There is no difference in the criterion-group discrimination of the medium difficulty and the low clinical-relevance subscales.

H_1: There is a difference in the criterion-group discrimination of the medium difficulty and the low clinical-relevance subscales. The difference favors the medium difficulty subscale.

IC. H_o: There is no difference in the criterion-group discrimination of the high clinical-relevance and the low clinical-relevance subscales.

H_1: There is a difference in the criterion-group discrimination of the high clinical-relevance and the low clinical-relevance subscales. The difference favors the high clinical-relevance subscale.

These hypotheses will be tested by three separate Discriminant Analyses, using the two subscale scores noted in each hypothesis as

the discriminating variables and the three criterion groups (residency-eligible, residents, students) as the independent variable.

These analyses will identify the most discriminating subscale score by examination of the standardized discriminant function coefficients and the Wilks' Lambda statistic. It will also provide univariate F-tests of the discriminating power of each of the separate subscales (Tatsuoka, 1971). Each hypothesis will be tested by forming an F-ratio of the two univariate F's for the scales being compared.

II. H_o: There are no differences in criterion-related validity (subscale scores correlated with mean Simulated Clinical Encounter ratings) between the medium difficulty and the high clinical-relevance subscales.

H_1: There are differences in criterion-related validity between the medium difficulty and the high clinical-relevance subscales. The high clinical-relevance subscale will have a higher validity coefficient than the medium difficulty subscale.

Analysis will require computation of correlation coefficients between each of the two subscales and the criterion of Simulated Clinical Encounter mean ratings. This hypothesis will be tested by a test of the difference of two non-independent correlation coefficients (Glass and Stanley, 1970).

III. H_o: There is no difference in internal-consistency reliability between the medium difficulty and the high clinical-relevance subscales.

H_1: There is a difference in internal-consistency reliability between the medium difficulty and the high clinical-relevance subscales. The difference favors the medium difficulty subscale.

The analysis will consist of computation of Kuder-Richardson 20 reliability coefficients for each subscale. Hypothesis III will be tested by forming an F-ratio of the $MS_{persons}/MS_{total}$ for both subscales under consideration (Wilson, 1978).

IV. H_o: There are no differences in mean item difficulty between the medium difficulty and the high or the low clinical-relevance subscales.

H_1: There are differences in mean item difficulty between the medium difficulty subscale and the high and the low clinical-relevance subscales. The medium difficulty subscale and the low clinical-relevance subscale will be more difficult than the high clinical-relevance scale.

A repeated measures Analysis of Variance will be used to test Hypothesis IV. Post-hoc contrasts will test differences between the means of the medium difficulty subscale and the high clinical-relevance subscale and also between the mean of the low clinical-relevance scale and the high clinical-relevance scale.

V. H_o: The proportion of overlap of identical items between the medium difficulty and the high clinical-relevance subscales will be the same as the proportion of overlap of identical items selected for the medium difficulty and the low clinical-relevance subscales.

H_1: The proportion of overlap of identical items between the medium difficulty and the high clinical-relevance subscales will be lower than the proportion of overlap of identical items between the medium difficulty and the low clinical-relevance subscales.

Analysis for Hypothesis V will require a count of the number of overlapping identical items in the scales noted and computation of these proportions. The hypothesis will be tested by drawing a confidence interval around the difference of the two proportions (Bacon, 1976).

VI. H_o: There are no differences in the distributions of pictorial-stem, clinical-situational, or factual multiple-choice items selected for the medium difficulty, the low difficulty, the high and the low clinical-relevance subscales.

H_1: There are differences in the distributions of pictorial-stem, clinical-situational, and factual multiple-choice items selected for the four subscales. The high clinical-relevance

subscale will have a larger distribution of pictorial-stem and/or clinical-situational items than the medium difficulty or the low clinical-relevance subscale.

Counts of the numbers of pictorial-stem, clinical-situational, and factual multiple-choice items selected for each subscale will be performed. A chi-square statistic will be used to test Hypothesis VI.

SUMMARY

Examination materials were developed as a certification examination in Emergency Medicine. This library of test items was administered to ninety-four subjects in four groups over a two and one-half day period. The two formats of most interest in this study are the Pictorial Multiple-Choice and the Multiple-Choice formats.

The study was designed to test the effect of clinically relevant item content on group-score discrimination, criterion-related validity, subscale reliability, and item difficulty and also to examine the effect of different item types on the clinical-relevance of item subscales.

Group discrimination differences for subscales of items selected by different criteria will be tested by a Discriminant Analysis procedure and an associated F-test. A Z-test of non-independent correlation coefficients will be calculated to test differences in criterion-related validity coefficients. An F-test of the difference between two reliability coefficients will be performed. A repeated measures ANOVA will test an hypothesis of equal item difficulties for three subscales. A Z-test of differences in proportions will be used to test differences in proportions of identical items selected for

subscales. And, a chi-square statistic will test differences in distributions of item types selected for each subscale.

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

CHAPTER IV

RESULTS

INTRODUCTION

This chapter presents the results of the statistical analyses that were performed to test the hypotheses of this study. Results are presented concerning differences in the ability of subscales to discriminate statistically three criterion-groups of subjects--the residency-eligible physicians, residents, and medical students. Specifically, the group discrimination of the medium difficulty subscale is compared to that of the high clinical-relevance subscale and the low clinical-relevance subscale. Differences in group discrimination between the high clinical-relevance and the low clinical-relevance subscales are also reported.

The criterion-related validity coefficients of the four subscales--correlations of subscale scores with the grand mean of Simulated Clinical Encounters--are reported and compared. An hypothesis test of the difference in criterion-related validity between the medium difficulty and the high clinical-relevance subscales is presented.

Another hypothesis for this study concerns differences in internal-consistency reliability between the medium difficulty and the high clinical-relevance subscales. Reliability coefficients for each of the four subscales are presented and the results of an hypothesis test are given.

Differences in mean item difficulty among the medium difficulty, the high and the low clinical-relevance subscales are reported. Two

post-hoc contrasts test differences in mean item difficulty between 1) the medium difficulty and the high clinical-relevance subscales, and 2) the low clinical-relevance and the high clinical-relevance subscales.

To investigate differences in the effect of item types on the statistical properties of scores, differences in the proportion of overlap of identical items between the medium difficulty/high clinical-relevance scales and the medium difficulty/low clinical-relevance scales are reported. Then, differences in distributions of item types--the pictorial-stem, clinical-situational, and factual multiple-choice item type--across the four subscales selected for this study are investigated and the results are reported.

Finally, other results of data analyses which were suggested by the findings of this study are presented. Specifically, this section presents the results of a small validation study in which some of the hypotheses tested are reanalyzed using the practice-eligible emergency physician group (n=14) alone or in combination with the resident and the medical student group.

ITEM SELECTION FOR FOUR SUBSCALES

Results of the data analyses performed to select items for the four subscales studied here are presented in Tables 4.1 to 4.4.

Medium Difficulty Subscale

Table 4.1 presents the item difficulty indices (p-value equals proportion marking a correct answer) and item-total score discrimination

TABLE 4.1

MEDIUM DIFFICULTY ITEMS
n=80

Item	p-value	Point-Biserial	Item	p-value	Point-Biserial
A12	.62	.11	D50	.54	.20
A28	.64	.30	D59	.66	.33
A30	.63	.02	D60	.65	.48
A35	.66	.38	D61	.55	.53
A36	.66	.38	D65	.56	.21
A37	.55	.29	D66	.63	.50
A40	.58	.17	D72	.66	.12
A43	.53	.40	D75	.60	.30
A45	.53	.08	D76	.53	.13
A50	.61	.30	D77	.50	.28
A57	.50	.24	D83	.64	.52
B1	.63	.31	D92	.65	.16
B2	.65	.12	E9	.61	.34
B4	.53	.51	E16	.61	.10
B6	.66	.03	E24	.55	.42
B7	.55	.24	E27	.60	.27
B21	.55	.15	E38	.63	.29
B23	.63	.10	E47	.55	.20
B33	.64	.51	E54	.54	.42
B45	.66	.25	E59	.54	.23
C7	.55	.22	E61	.59	.36
C11	.63	.19	E69	.68	.43
C14	.64	.49	E70	.55	.37
C20	.60	.59	E76	.55	.17
C24	.54	.27	E79	.63	.36
C31	.50	.35	E80	.51	.23
C33	.61	.38	E88	.66	.50
C39	.61	.10	F5	.53	.23
C59	.56	.24	F6	.68	.18
C67	.58	.33	F7	.65	.44
C70	.66	.22	F11	.64	.45
C73	.56	.36	F12	.68	.37
C74	.65	.31	F21	.60	.17
C78	.61	.26	F23	.68	.02
C84	.60	.15	F32	.68	.11
D2	.65	.46	F36	.51	.19
D7	.53	.06	F37	.61	.23
D8	.63	.13	F43	.56	.06
D12	.65	.23	F47	.66	.39
D18	.69	.06	F51	.65	.47
D23	.58	.32	F60	.61	.14
D27	.65	.37	F71	.69	.08
D28	.68	.51	F79	.61	.22
D32	.53	.20	F82	.66	.09
D37	.56	.24	F90	.60	.42
D49	.64	.43			

index (point-biserial correlation of the item score and the total correct score), for the medium difficulty subscale.

Items for the medium difficulty subscale were selected to be near the ideal difficulty recommended by some measurement specialists (e.g., Ebel, 1972) for educational achievement examinations intended to yield scores that will be interpreted relative to some norm group's performance. This ideal difficulty is a mean score on the test that is approximately midway between the chance score and the perfect score. For a ninety-one item test, with items having four options, the ideal mean score would be approximately 57 items correct. This ideal mean score corresponds to a mean p-value of approximately .63.

For the medium difficulty scale in this study, ninety-one items were selected to range in p-value from .50 to .69, for the residency-eligible, the resident, and the student groups (n=80). All items selected had positive item-total score discrimination indices. Items were selected by starting the selection process at p = .50 and continuing to select less difficult items until the quota of ninety-one items had been selected. There were five tied ranks at p = .69; two items out of five were selected at random to complete this ninety-one item subscale. The mean p-value for this scale is .603.

A subscale total score was computed such that each subject's score on the medium difficulty scale was the sum of the number of correct responses to these ninety-one items.

Low Difficulty Subscale

Item analysis data for the items selected for the low difficulty subscale are presented in Table 4.2. The ninety-one items

TABLE 4.2

LOW DIFFICULTY ITEMS
n=80

Item	p-value	Point-Biserial	Item	p-value	Point-Biserial
A30	.90	.20	C82	.91	.10
A6	.84	.30	C83	.91	.10
A33	.96	.36	D4	.85	.19
A34	.95	.07	D5	.95	.09
A44	.98	.31	D6	.89	.16
A46	.94	.10	D19	.85	.06
A49	.88	.34	D29	.88	.26
A53	.90	.49	D30	.89	.21
A55	.88	.18	D35	.93	.30
A58	.99	.32	D36	.86	.11
A59	.95	.23	D38	.90	.19
A60	.96	.08	D41	.94	.19
A65	.91	.16	D47	.94	.41
B3	.90	.08	D53	.85	.40
B15	.86	.30	D57	.85	.39
B16	.84	.40	D80	.93	.21
B18	.85	.44	D82	.94	.24
B22	.89	.18	E7	.88	.07
B24	.89	.16	E23	.89	.19
B28	.85	.23	E33	.85	.34
B29	.91	.34	E39	.84	.20
B38	.90	.30	E40	.85	.62
B39	.89	.30	E51	.86	.02
B40	.86	.30	E66	.93	.39
B44	.88	.43	E67	.85	.38
B47	.85	.23	E73	.94	.03
B48	.88	.23	E90	.90	.45
B49	.95	.13	F1	.89	.30
B58	.93	.30	F9	.96	.18
B59	.95	.33	F18	.85	.13
B63	.85	.17	F26	.84	.27
B64	.91	.33	F30	.90	.38
B66	.98	.10	F33	.89	.43
B67	.93	.33	F46	.89	.21
B68	.94	.33	F48	.94	.14
C4	.90	.05	F63	.94	.00
C21	.96.	.38	F66	.94	.26
C23	.90	.29	F69	.88	.13
C27	.89	.52	F70	.98	.18
C28	.94	.34	F75	.91	.34
C42	.91	.42	F81	.86	.33
C44	.89	.30	F89	.93	.13
C48	.96	.31	F91	.84	.06
C51	.84	.31			
C57	.98	.42			
C63	.93	.18			
C68	.94	.41			
C69	.99	.27			

selected for this subscale range in p-value from .84 to .99, with positive (or zero) item-total discrimination indices. The mean p-value for this scale is .903.

Items for this subscale were selected by choosing the ninety-one least difficult items from a listing of items ranked by p-value from easiest to most difficult. There were ten tied ranks at p = .84; six items were chosen randomly to complete the quota of ninety-one items. A low difficulty scale score was computed for each subject by summing the correct responses to these ninety-one items.

High Clinical-Relevance Items

Table 4.3 presents the items selected for the high clinical-relevance subscale and their point-biserial correlations with the criterion of grand mean ratings on the Simulated Clinical Encounters.

These items were selected by rank-ordering all items from highest to lowest item-criterion correlation and selecting the ninety-one items with the highest correlation with the criterion. Item-criterion correlations range from r = .33 to .68, with a median r = .38 and a mode of r = .33. The mean p-value for this scale is .721. There were fifteen tied ranks at r = .33; twelve of the fifteen items were randomly selected to complete this ninety-one item subscale. Correct responses to these items were summed to form a high clinical-relevance subscale score for each subject.

Low Clinical-Relevance Items

Table 4.4 presents the items selected for the low clinical-relevance subscale and their item-grand mean Simulated Clinical Encounter correlation coefficients. These items were chosen by

TABLE 4.3

HIGH CLINICAL-RELEVANCE ITEMS
n=80

Item	Point-Biserial (Item-Criterion)	Item	Point-Biserial (Item-Criterion)
A1	.58	D53	.35
A2	.68	D60	.45
A4	.42	D61	.38
A5	.36	D64	.36
A6	.33	D66	.46
A23	.37	D71	.37
A35	.40	D73	.37
A36	.37	D83	.36
A53	.33	D90	.34
A67	.39	E2	.38
B4	.48	E24	.38
B12	.34	E29	.38
B16	.37	E40	.66
B19	.44	E46	.40
B29	.31	E48	.36
B30	.33	E54	.35
B33	.40	E55	.33
B36	.42	E64	.39
B42	.39	E66	.37
B44	.36	E67	.35
B51	.38	E69	.33
B52	.56	E70	.33
B57	.43	E72	.50
B58	.33	E88	.44
B64	.37	F7	.53
B65	.40	F10	.38
B67	.36	F11	.43
B68	.37	F12	.37
C14	.45	F13	.38
C18	.33	F19	.36
C20	.56	F24	.34
C21	.35	F34	.37
C27	.50	F51	.49
C31	.33	F52	.49
C42	.33	F53	.45
C47	.34	F67	.46
C51	.45	F68	.42
C52	.42	F76	.41
C55	.36	F77	.51
C57	.34	F80	.36
C68	.38	F84	.40
D2	.33	F85	.34
D27	.34	F86	.34
D28	.45	F87	.35
D47	.46	F90	.33
D49	.40	F90	.33

TABLE 4.4

LOW CLINICAL-RELEVANCE ITEMS
n=80

Item	Point-Biserial (Item-Criterion)	Item	Point-Biserial (Item-Criterion)
A8	.11	D18	.12
A12	.11	D19	-.01
A29	.08	D30	.08
A30	.02	D36	.10
A34	-.08	D58	.07
A38	.10	D67	.00
A39	.05	D72	.06
A42	.06	D76	.07
A46	.01	D81	-.02
A51	.09	D84	.00
A52	.08	D86	-.02
B5	-.06	D92	.02
B6	.09	E7	-.05
B20	.09	E14	.02
B21	.06	E18	.04
B22	.05	E20	.05
B23	-.01	E21	.09
B25	-.03	E26	.02
B26	.02	E32	-.03
B37	-.12	E35	.10
B46	.09	E36	.05
B66	.11	E42	-.03
C4	-.01	E43	.05
C6	.05	E47	.07
C7	.11	E51	-.01
C11	.08	E56	.03
C16	.10	E58	.06
C22	-.05	E71	.11
C29	.01	E75	-.06
C32	.10	E76	.05
C35	-.07	F4	.02
C36	.09	F22	.11
C38	.08	F27	.11
C41	.01	F32	.01
C43	-.04	F46	.07
C53	-.23	F49	.10
C54	-.10	F50	-.10
C62	.04	F57	.00
C65	.10	F63	-.11
C66	-.05	F69	.10
C70	.11	F71	.06
C82	-.04	F73	.03
C83	.07	F89	.09
C84	.11		
D6	-.04		
D7	.05		
D8	.07		
D14	-.03		

selecting the ninety-one items that had the lowest item-criterion correlation. Correlations with the criterion ratings ranged from .23 to .11; there were twenty-eight negatively correlating items. The median item-criterion correlation is r = .05, with a mode of r = .11. There were no tied ranks for this subscale. The mean p-value for this scale is .667. A total correct score on these ninety-one items was computed for each subject.

STATISTICAL ANALYSIS FOR GROUP DISCRIMINATION HYPOTHESES

The first three hypotheses of this study concern the differential power of subscales selected by two different criteria to statistically separate or discriminate criterion groups with known levels of training and experience in a medical specialty. Discriminant Analysis was used to analyze data for these hypotheses. A brief description of the technique of Discriminant Analysis follows.

Discriminant Analysis is a multivariate statistical technique that weights potential discriminating variables and linearly combines these variables such that the discrimination between two or more groups of subjects is maximized. The discriminant function has the form:

$$D_i = d_{i1}Z_1 + d_{i2}Z_2 + \ldots . d_{ip}Z_p \qquad (1)$$

Where: D_i = Score on discriminant function i

d_i = Weighting Coefficients

Z = Standardized values of the p discriminating variables

The mathematics of Discriminant Analysis restrains the number of discriminant functions derived to a maximum of the number of groups minus one or to the number of discriminating variables in the analysis.

Several statistics are used to test the importance of variables to the maximum separation of known groups. For the stepwise Discriminant Analyses used to analyze data for the hypotheses of this study, the following statistics are important:

1. Eigenvalue: an index of the relative importance of the discriminant function derived. The sum of the eigenvalues is a measure of the total variance of the discriminating variables.
2. Relative Percent of Eigenvalue: Proportion of total variance of discriminating variables accounted for by the function derived.
3. Canonical Correlation: Correlation of the discriminant function and the set of g-1 dummy variables which define the g-groups discriminated. The square of the Canonical correlation coefficient defines the percentage of variance in the discriminant function explained by the criterion groups.
4. Wilks' Lambda: An inverse measure of the discriminating ability of the variables in the analysis. When Lambda is small, the discrimination is high.
5. Standardized Discriminant Function Coefficient (d_i): The coefficient which when multipled by z-scores for each subject, maximizes the discrimination of groups. These coefficients are interpreted like beta weights in a regression equation and, analytically, like factor loadings in a factor analysis.
6. Classification Analysis: A classification of predicted group membership based on the discriminant function(s) derived.

Predicted group membership is compared to actual group membership. Percentage of correct classification is an index of the ability of the discriminating variables in the analysis to validly discriminate groups.

RESULTS CONCERNING DIFFERENCES IN DISCRIMINATION: MEDIUM DIFFICULTY VERSUS HIGH CLINICAL-RELEVANCE

Hypothesis IA stated that the high clinical-relevance subscale scores would statistically discriminate the residency-eligible physicians, the residents, and the medical students better than the medium difficulty subscale scores. This hypothesis was analyzed by computing stepwise Discriminant Analyses on these data and test statistics associated with the Discriminant Analyses (Tatsuoka, 1971).

The medium difficulty and the high clinical-relevance scale scores were entered into a stepwise Discriminant Analysis[4], using a scale selection criterion that minimizes Wilks' Lambda. The F-value for inclusion of a scale in the analysis was $\alpha = .01$.

Table 4.5 shows the raw-score means and standard deviations for each criterion-group in this study. Univariate F-ratios of the scale scores indicate at $p \leq .0001$ that both the medium difficulty and the high clinical-relevance scores taken separately discriminate the three groups well. Wilks' Lambda shows fairly strong discriminating power for each scale. It should be noted that both the F-ratios

[4] An applications computer program, the Statistical Package for the Social Sciences, (Nie et al., 1970) was used for all Discriminant Analyses.

TABLE 4.5

RAW-SCORE GROUP DISCRIMINATION:
MEDIUM DIFFICULTY VS HIGH CLINICAL-RELEVANCE

	Residency-Eligible		Residents		Students		Wilks'	$F_{2, 77}$
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Lambda	
Medium Difficulty	66.77	8.67	56.83	7.89	39.77	7.14	.3686	65.94*
High Clinical Relevance	80.23	7.67	70.81	7.59	42.55	9.98	.2352	125.21*

* $p \leq .0001$

and Wilks' Lambda show relatively stronger discriminating power for the high clinical-relevance scale than for the medium difficulty scale.

Figures 4.1 and 4.2 graphically show the raw-score separation of the three criterion-groups for the medium difficulty scale and the high clinical-relevance scale, respectively. Comparing the curves of the raw-scores for the medium difficulty scale and the high clinical-relevance scales shows the relative power of the high clinical-relevance scale in the discrimination of the criterion-groups. There is considerably less overlap in curves for the three groups for the high clinical-relevance scores compared to the medium difficulty scores.

Table 4.6 presents a summary of the stepwise Discriminant Analysis performed for this hypothesis. This table shows that the high clinical-relevance scale was entered first; this first function alone yielded a small Wilks' Lambda, indicating the relatively higher group discriminating power of the high clinical-relevance scale compared to the medium difficulty scale. When the medium difficulty scale was added in the second step, Wilks' Lambda decreased only slightly. This result shows that addition of the medium difficulty scale increased the discriminating power only a small but statistically significant amount, given the discrimination accounted for by the high clinical-relevance scale.

In Table 4.7 the standardized Discriminant Function coefficients are presented. The first function weights the high clinical-relevance scale in a ratio of 6.65 : 1 relative to the medium difficulty scale.

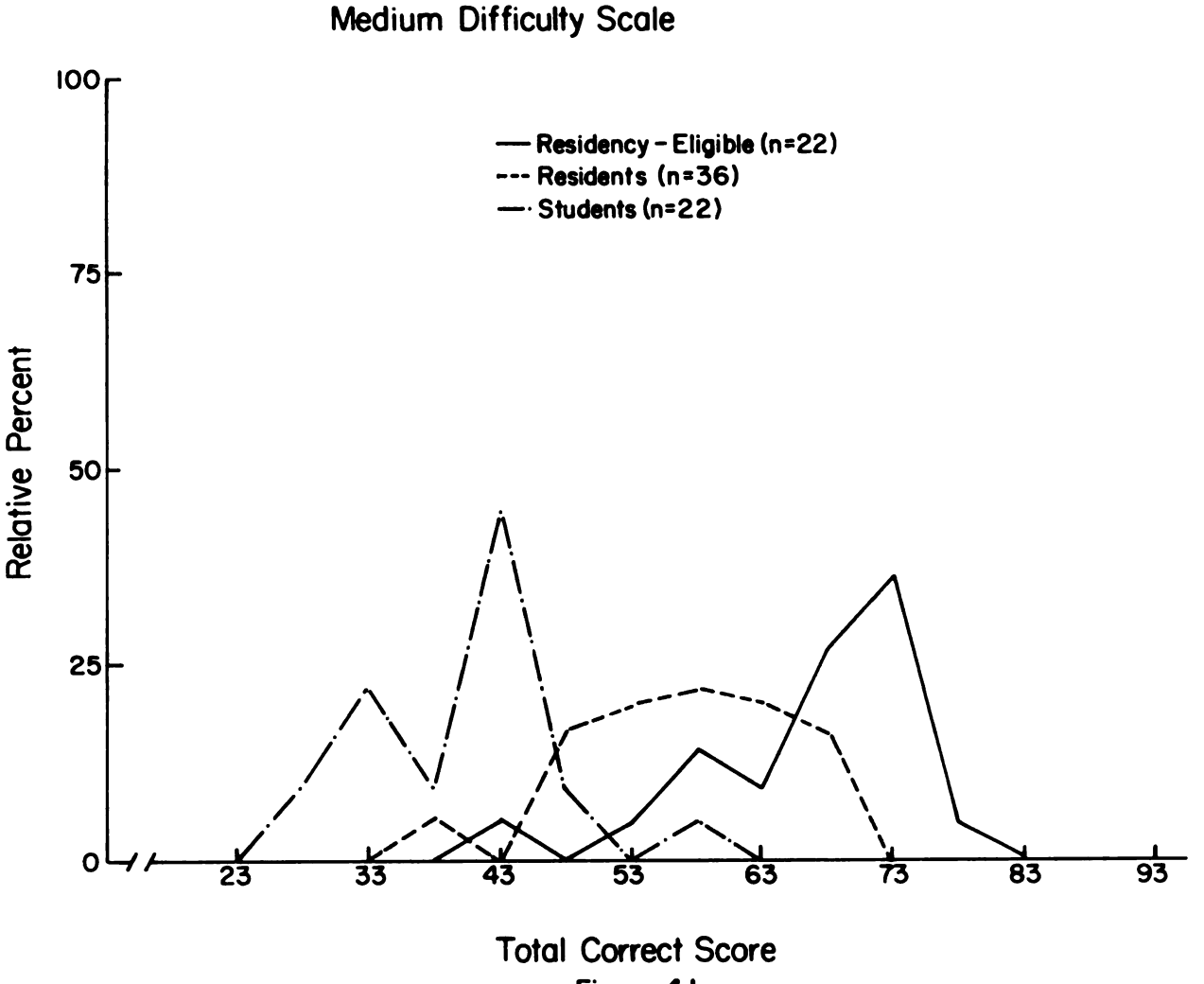
Medium Difficulty Scale
Residency - Eligible (n=22)
Residents (n=36)
Students (n=22)
Relative Percent
100
75
50
25
0
23
33
43
53
63
73
83
93
Total Correct Score

Figure 4.1

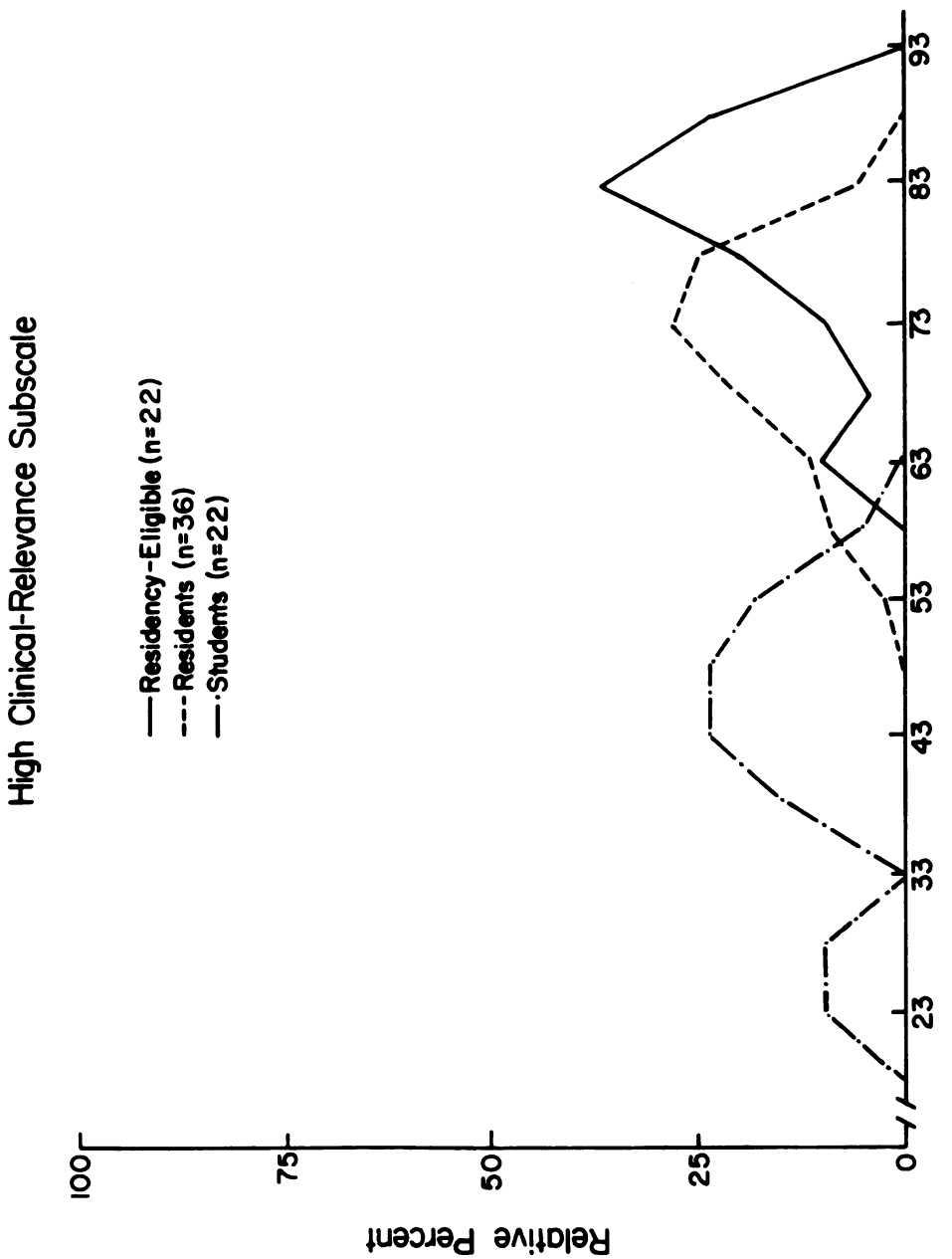
High Clinical-Relevance Subscale
—Residency-Eligible (n=22)
---Residents (n=36)
—·Students (n=22)
Relative Percent
100
75
50
25
0
23
33
43
53
63
73
83
93
Total Correct Score

Figure 4.2

TABLE 4.6

SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS:
HIGH CLINICAL-RELEVANCE VS MEDIUM DIFFICULTY

Step Number	Scale Name	F to Enter	Wilks' Lambda	p-value
1	High-Clinical Relevance	125.21	.235	$\leq$ .0001
2	Medium Difficulty	4.38	.211	$\leq$ .0001

TABLE 4.7

STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS:
MEDIUM DIFFICULTY VS HIGH CLINICAL-RELEVANCE

	Function 1	Function 2
Medium Difficulty	-0.357	-2.664
High Clinical-Relevance	2.374	2.386

The second function, the medium difficulty function, weights the medium difficulty scale only 1.12 times greater than the high clinical-relevance scale.

Table 4.8 presents other data on the relative contribution of both scales to the discrimination of groups. When both discriminant functions are used, Wilks' Lambda is small and highly significant. The Canonical correlation shows that both functions together can account for 77 percent of the known variance of group membership. A large proportion (97 percent) of the total eigenvalue is explained by both functions. When the first function, the high clinical-relevance function, is removed from the analysis, relatively small, but statistically significant, discriminating power is accounted for by the medium difficulty scale alone.

In Table 4.9, the accuracy of classifications made using the two discriminant functions of this analysis is given. A total of 81.3 percent of the subjects were accurately classified by these two discriminant functions. A chi-square statistic was calculated to test the hypothesis that the observed correct classifications were due to chance alone. The $\chi^2 = 82.66$ with 4 degrees of freedom is significant at $p \leq .0001$. The null hypothesis of chance accuracy is, therefore, rejected in favor of the alternative that correct classifications were not due to chance.

The high clinical relevance scale taken separately correctly classified 76.3 percent of subjects correctly, while the medium difficulty scale classified 71.2 percent correctly.

TABLE 4.8

RELATIVE DISCRIMINATING POWER OF
MEDIUM DIFFICULTY AND HIGH CLINICAL-RELEVANCE SCALES

Scale Functions in Analysis	Eigenvalue	Canonical Correlation	Percent of Total Eigenvalue	Wilks' Lambda	Chi-Square
High Clinical-Relevance and Medium Difficulty	3.309	.876	97.0	.211	119.08**
Medium Difficulty Alone	.101	.302	3.0	.909	7.34*

** $p \leq .0001$

* $p \leq .007$

TABLE 4.9

CLASSIFICATION ANALYSIS USING
HIGH CLINICAL-RELEVANCE AND
MEDIUM DIFFICULTY DISCRIMINANT FUNCTIONS

		Predicted[a]		
	N	Residency-Eligible	Residents	Students
Residency-Eligible	22	15(68.2)	7(31.8)	0(0)
Residents	36	6(16.7)	29(80.6)	1(2.8)
Students	22	0(0)	1(4.5)	21(95.5)

[a]Number in parentheses indicates percentage of classification for that group.

TABLE 4.10

SUMMARY OF STEPWISE DISCRIMINANT
ANALYSIS WITH MEDIUM DIFFICULTY
ENTERED FIRST

Step Number	Scale Name	F to Enter	Wilks' Lambda	p-value
1	Medium Difficulty	65.94	.369	$\leq$ .0001
2	High Clinical-Relevance	28.43	.211	$\leq$ .0001

Discriminant Analysis With Medium Difficulty Entered First

The mathematics of stepwise Discriminant Analysis restricts the second function to explaining group discrimination remaining after the first function is extracted. Therefore, to evaluate further the relative contribution to group discrimination of these two scales, another Discriminant Analysis was performed in which the medium difficulty subscale was forced to enter the analysis first. Table 4.10 summarizes this analysis.

When the medium difficulty scale is forced to enter first, Wilks' Lambda is small (.369) and statistically significant. But there is considerable difference between Lambda when the high clinical-relevance scale is entered first (.235), compared to Lambda when the medium difficulty scale is entered first. This finding supports the previous result, showing the relative power of the high clinical-relevance scale, compared to the medium difficulty scale, in statistically separating groups.

Hypothesis Test

There is no specific test of the hypothesis of no difference in group discrimination between the medium difficulty and the high clinical-relevance scale given by the Discriminant Analyses. However, a test statistic may be formed by a ratio of the two univariate F-values calculated for the separate one-way ANOVAs given in Table 4.5.

The test statistic is:

$$F_{calculated} = \frac{F_{High\ Clinical\text{-}Relevance}}{F_{Medium\ Difficulty}} \qquad (2)$$

For the hypothesis of no difference in discrimination between the high clinical-relevance and the medium difficult scales:

$$F = \frac{125.21}{65.94} = 1.90$$

The critical value at $\alpha = .05$ for 2 and 77 degrees of freedom is (conservatively) 3.15. Since F-calculated is less than the critical value, the null hypothesis can not be rejected. There is no statistical difference between the medium difficulty and the high clinical-relevance scales in their ability to discriminate criterion-groups.

RESULTS CONCERNING DIFFERENCES IN DISCRIMINATION: MEDIUM DIFFICULTY VERSUS LOW CLINICAL-RELEVANCE

Hypothesis IB stated that the medium difficulty subscale scores would discriminate the residency-eligible physician, the resident, and the medical student criterion-groups better than the low clinical-relevance subscale scores. This hypothesis was analyzed by computing stepwise Discriminant Analyses and associated statistics for these scales. All computer analyses were identical to those for hypothesis IA.

Table 4.11 reports the raw-score group discrimination for each of these subscales. The group means and standard deviations show the medium difficulty scale separates the three criterion-groups more sharply than the low clinical-relevance scale. This finding is confirmed by the Wilks' Lambda statistic and the univariate F-ratios computed for each scale.

TABLE 4.11

RAW-SCORE GROUP DISCRIMINATION:
MEDIUM DIFFICULTY VS LOW CLINICAL-RELEVANCE

	Residency-Eligible		Residents		Students		Wilks'	$F_{2, 77}$
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Lambda	
Medium Difficulty	66.77	8.67	56.83	7.89	39.77	7.14	.3686	65.94**
Low Clinical-Relevance	64.27	6.38	59.97	6.08	58.27	5.24	.8638	6.07*

** $p \leq .0001$

* $p \leq .005$

Figure 4.3 presents a plot of the low clinical-relevance scores for the three criterion-groups under investigation here. These overlapping curves show that the low clinical-relevance scores separate the groups poorly. Comparison of Figures 4.3 and 4.1 shows the relative power of group discrimination for these two subscales. (For ease of comparisons, Figure 4.4--the low difficulty subscale plot--is also presented here.)

In Table 4.12, a summary of the first stepwise Discriminant Analysis on these two subscales is presented. When the computer program was allowed to choose the most discriminating scale for entry into the analysis first, the medium difficulty scale was selected, followed by the low clinical-relevance scale. The small, but statistically significant, change in Wilks' Lambda from step one to step two indicates that the low clinical-relevance scale can contribute little to group discrimination, given the contribution of the medium difficulty scale.

The standardized Discriminant Function coefficients for the two functions and the two scales are presented in Table 4.13. The medium difficulty scale is weighted 4.66 times more than the low clinical-relevance scale in the first function. In the second function, the low clinical-relevance scale is weighted in a ratio of 2.46 : 1 relative to the medium difficulty scale.

Table 4.14 presents the relative discriminating power of the medium difficulty and the low clinical-relevance scales. Both functions together--the medium difficulty and the low clinical-relevance--account for 65 percent of the variance in groups; 98 percent of the total eigenvalue is explained by both functions. When

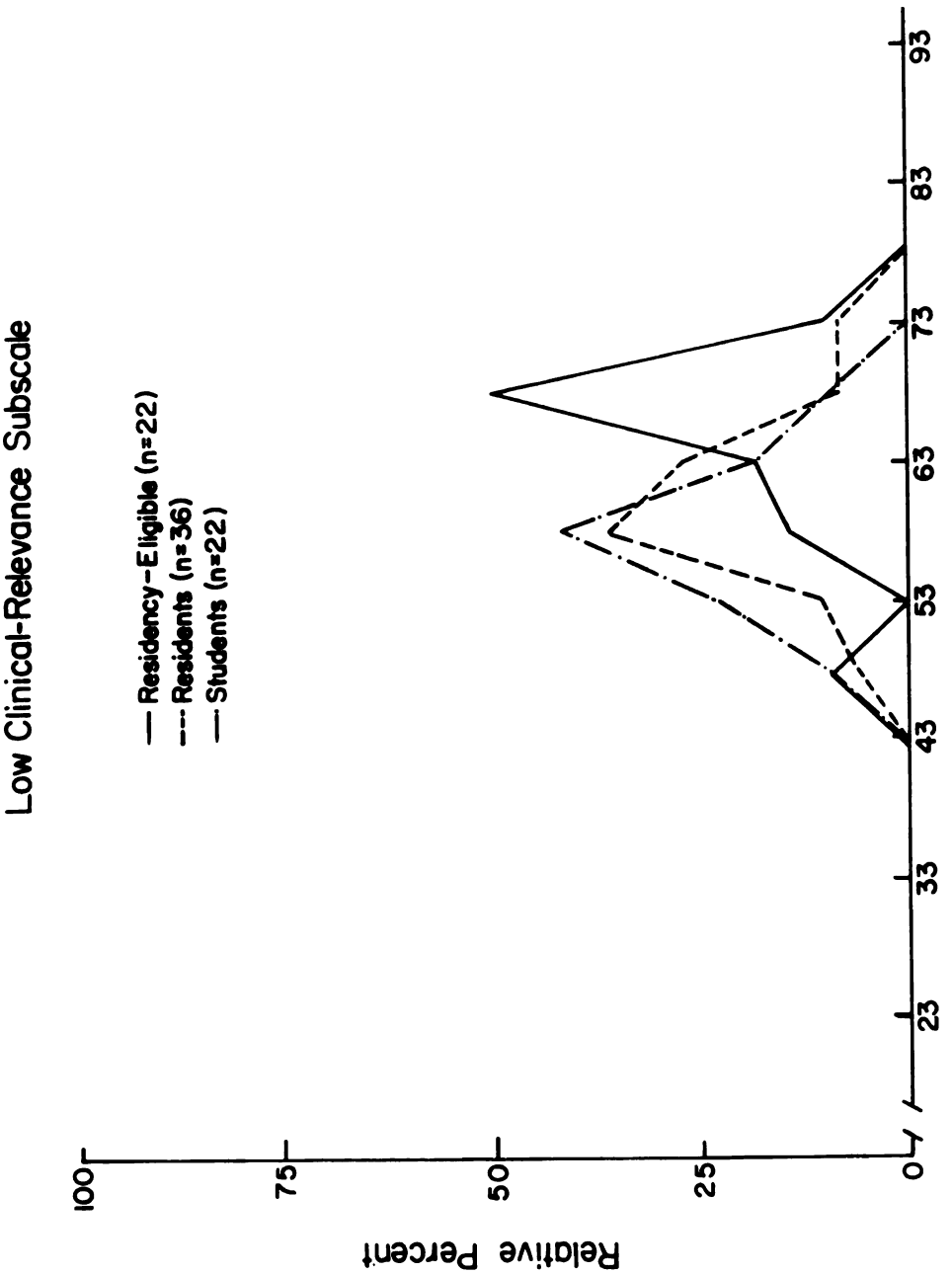
Low Clinical-Relevance Subscale
Residency-Eligible (n=22)
Residents (n=36)
Students (n=22)
Relative Percent
100
75
50
25
0
23
33
43
53
63
73
83
93
Total Correct Score

Figure 4.3

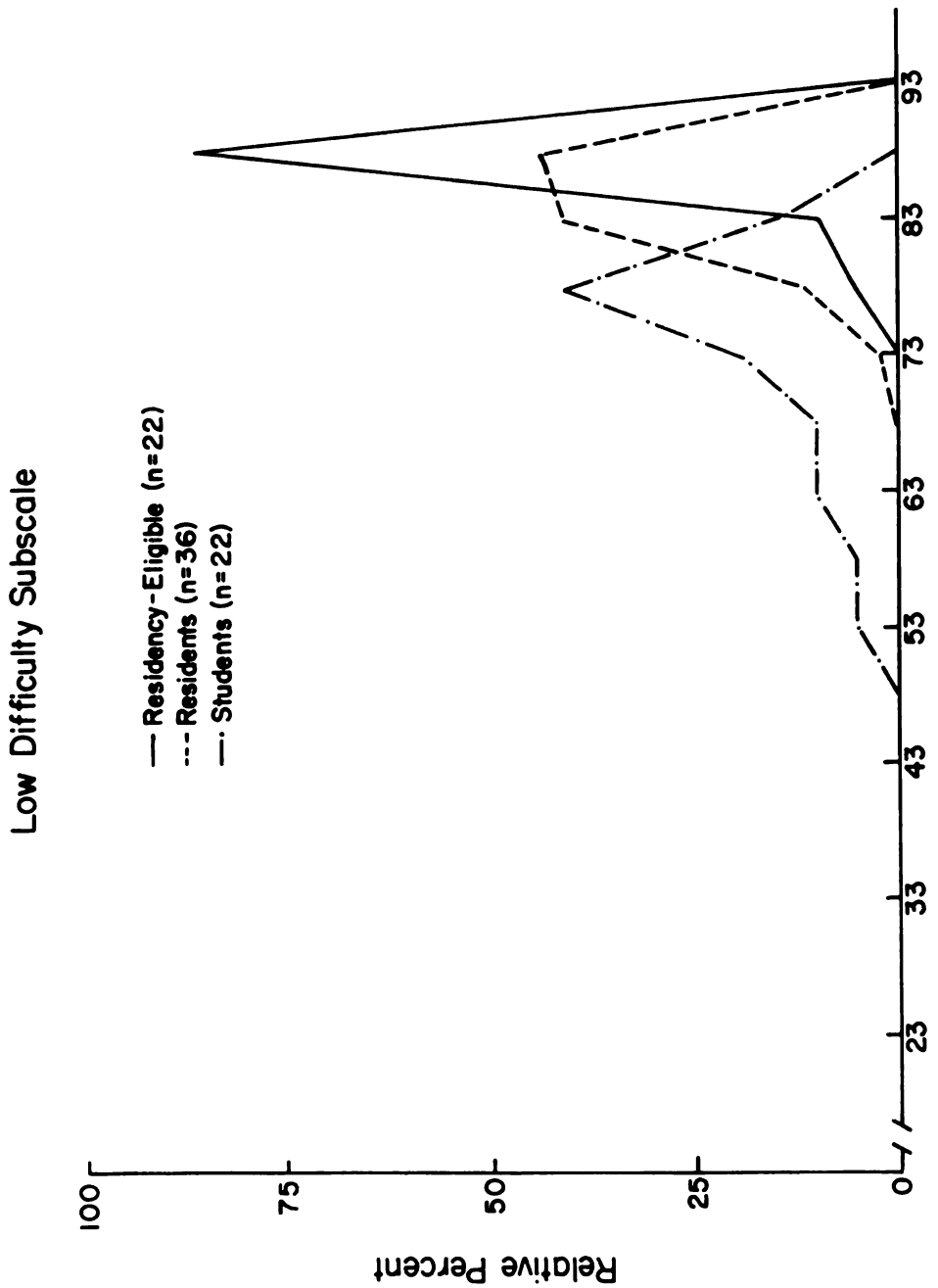
Low Difficulty Subscale
— Residency-Eligible (n=22)
--- Residents (n=36)
—. Students (n=22)
Relative Percent
100
75
50
25
0
23
33
43
53
63
73
83
93
Total Correct Score

Figure 4.4

TABLE 4.12

SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS:
MEDIUM DIFFICULTY VS LOW CLINICAL-RELEVANCE

Step Number	Scale Name	F to Enter	Wilks' Lambda	p-value
1	Medium Difficulty	65.94	.369	$\leq$ .0001
2	Low Clinical-Relevance	4.24	.332	$\leq$ .0001

TABLE 4.13

STANDARDIZED DISCRIMINANT FUNCTION
COEFFICIENTS:
MEDIUM DIFFICULTY VS LOW CLINICAL-RELEVANCE

	Function 1	Function 2
Medium Difficulty	1.879	-0.489
Low Clinical-Relevance	-0.403	1.203

TABLE 4.14

RELATIVE DISCRIMINATING POWER:
MEDIUM DIFFICULTY VS LOW CLINICAL-RELEVANCE

Scale Functions in Analysis	Eigenvalue	Canonical Correlation	Percent of Total Eigenvalue	Wilks' Lambda	Chi-Square
Medium Difficulty and Low Clinical-Relevance	1.901	.809	98.0	.332	84.44**
Low Clinical-Relevance Alone	.040	.195	2.0	.962	2.97*

** $p \leq .0001$

* $p \leq .085$

the medium difficulty scale is removed from the analysis, the low clinical-relevance scale alone can explain only 4 percent of the group membership variance and a relatively small proportion of the total eigenvalue. The medium difficulty scale is so much more powerful in its discrimination of groups that a more conservative alpha level for entry (e.g., $\alpha = .05$) would exclude the low clinical-relevance scale from further analysis.

The accuracy of group classification using the medium difficulty and the low clinical-relevance scales is given in Table 4.15. The percentage of subjects correctly classified by these two Discriminant functions is 76.3. A chi-square test statistic was calculated to test the hypothesis that accuracy of classification was due to chance. A χ^2 value of 66.31 with 4 degrees of freedom is significant beyond $\alpha = .0001$. There is strong support for the alternative hypothesis that accuracy of classification is not due to chance.

The low clinical-relevance scale, taken separately, classified only 45 percent of subjects accurately.

Discriminant Analysis With Low Clinical-Relevance Entered First

As in the previous hypothesis, a second Discriminant Analysis was performed on these two scales. In this analysis, the low clinical-relevance subscale was forced to enter the analysis first. Table 4.16 shows a summary of the Discriminant Analysis with the low clinical-relevance scale entered first. Comparing the Wilks' Lambda when each scale is forced to the analysis first, a large difference in Lambda (.395) favoring the medium difficulty scale is observed, indicating the strength of the medium difficulty scale relative to the low clinical-relevance scale in its power to discriminate groups.

TABLE 4.15

CLASSIFICATION ANALYSIS USING MEDIUM DIFFICULTY AND LOW CLINICAL-RELEVANCE DISCRIMINANT FUNCTIONS

		Predicted[a]		
	N	Residency-Eligible	Residents	Students
Residency-Eligible	22	15(68.2)	6(27.3)	1(4.5)
Residents	36	6(16.7)	26(72.2)	4(11.1)
Students	22	0(0)	2(9.1)	20(90.9)

[a]Number in parenthesis indicates percentage of classification for that group.

TABLE 4.16

SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS WITH LOW CLINICAL-RELEVANCE ENTERED FIRST

Step Number	Scale Name	F to Enter	Wilks' Lambda	p-value
1	Low Clinical-Relevance	6.07	.864	$\leq$ .004
2	Medium Difficulty	60.98	.332	$\leq$ .0001

Hypothesis Test

A test of Hypothesis IB--that the medium difficulty scale discriminates the criterion-groups better than the low clinical-relevance scale--is given by an F-ratio of the two univariate F-statistics presented in Table 4.11.

For this hypothesis, the test statistic is:

$$F_{calculated} = \frac{F_{Medium\ Difficulty}}{F_{Low\ Clinical\text{-}Relevance}} \qquad (3)$$

For the hypothesis of no difference in group discrimination between the medium difficulty and the low clinical-relevance scales:

$$F = \frac{65.94}{6.07} = 10.86$$

For $\alpha = .001$, the critical value for 2 and 77 degrees of freedom is (conservatively) 7.76. Since F-calculated is larger than the critical value, the hypothesis of no difference in group discrimination between the medium difficulty and the low clinical-relevance scale is rejected. These is a statistical difference in criterion-group discrimination and this difference favors the medium difficulty scale.

RESULTS CONCERNING DIFFERENCES IN DISCRIMINATION: HIGH CLINICAL-RELEVANCE VERSUS LOW CLINICAL-RELEVANCE

Hypothesis IC stated that the high clinical-relevance scale would discriminate the three criterion-groups (residency-eligible, residents, medical students) better than the low clinical-relevance subscale. This hypothesis was analyzed by computing stepwise Discriminant Analyses and associated statistics using these scales as the discriminating variables and the three criterion-groups as the independent variable.

The raw-score discrimination of these scales is presented in Table 4.17. The scale means and standard deviations and the univariate F-ratios, plus the Wilks' Lambda, all show the high clinical-relevance scale to be much more powerful than the low clinical-relevance scale in its discrimination of the three criterion groups.

Comparison of the plots of scores for the high clinical-relevance and low clinical-relevance scales (Figures 4.2 and 4.3) shows the relative power of the high clinical-relevance scale to discriminate the known groups in this study. Low clinical-relevance scores for the three groups almost totally overlap while high clinical-relevance scores separate groups more adequately.

Table 4.18 summarizes the stepwise Discriminant Analysis of these two subscales. The computer program, choosing the best discriminator, entered the high clinical-relevance scale first. This first step produced a low Lambda (.235), indicating a high degree discriminating power for this scale. When the low clinical-relevance scale was entered in the second step, only a small decrease (.025) was noted in Lambda. This indicates that the low clinical-relevance scale could add only a small, but statistically significant, amount of discrimination to that already accounted for by the high clinical-relevance scale.

The standardized Discriminant Function coefficients for these two scales and the two functions derived are presented in Table 4.19. The first function weights the high clinical-relevance scale 7.03 times more heavily than the low clinical-relevance scale. In the second function, low clinical-relevance is weighted only 2.96 times heavier than the high clinical-relevance scale. This difference in weighting

TABLE 4.17

RAW-SCORE GROUP DISCRIMINATION:
HIGH CLINICAL-RELEVANCE VS LOW CLINICAL-RELEVANCE

	Residency-Eligible		Residents		Students		Wilks'	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Lambda	$F_{2, 77}$
High Clinical-Relevance	80.23	7.67	70.81	7.59	42.55	9.98	.2352	125.21**
Low Clinical-Relevance	64.27	6.38	59.97	6.08	58.27	5.24	.8638	6.07*

** $p \leq .0001$

* $p \leq .005$

TABLE 4.18

SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS:
HIGH VS LOW CLINICAL-RELEVANCE

Step Number	Scale	F to Enter	Wilks' Lambda	p-value
1	High Clinical-Relevance	125.21	.235	$\leq$ .0001
2	Low Clinical-Relevance	4.57	.210	$\leq$ .0001

TABLE 4.19

STANDARDIZED DISCRIMINANT FUNCTION
COEFFICIENTS: HIGH VS LOW
CLINICAL-RELEVANCE

	Function 1	Function 2
High Clinical-Relevance	2.221	-0.383
Low Clinical-Relevance	-0.316	1.132

shows the relative importance to group discrimination of the high clinical-relevance scale compared to the low clinical-relevance scale.

Table 4.20 shows more clearly the relative contribution to group discrimination of these two scales. When both the high and the low clinical-relevance scales are in the analysis, 78 percent of the variance of known groups is accounted for. When only the low clinical-relevance scale is used to discriminate groups, only 6 percent of known group variance can be accounted for. The change in Wilks' Lambda from a highly significant .210 for both scales taken together to .945 for the low clinical-relevance scale alone indicates the relative strength of the high clinical-relevance scale's contribution to group discrimination.

Table 4.21 presents the accuracy of group classifications made by predictions from the two discriminant functions derived from this analysis. The prediction accuracy was 80 percent using both discriminant functions noted here. A chi-square test of the hypothesis that this accuracy was random, versus the alternative hypothesis that the accuracy of classification was not due to chance, was computed. A $\chi^2 = 78.4$ with 4 degrees of freedom is significant at less than $\alpha = .0001$. The null hypothesis is, therefore, rejected in favor of the alternative that the accuracy of classification was not due to chance.

Discriminant Analysis With Low Clinical-Relevance Entered First

When the low clinical-relevance scale is forced to enter the analysis first, Wilks' Lambda is high (.864), but statistically significant, as a discriminator of the criterion-groups, as shown

TABLE 4.20

RELATIVE DISCRIMINATING POWER OF HIGH VS LOW CLINICAL-RELEVANCE SCALES

Scale Functions in Analysis	Eigenvalue	Canonical Correlation	Percent of Total Eigenvalue	Wilks' Lambda	Chi-Square
High Clinical-Relevance and Low Clinical-Relevance	3.501	.882	98.4	.210	119.42**
Low Clinical-Relevance Alone	.058	.235	1.6	.945	4.34*

** $p \leq .0001$

* $p \leq .037$

TABLE 4.21

CLASSIFICATION ANALYSIS USING HIGH AND LOW CLINICAL-RELEVANCE DISCRIMINANT FUNCTIONS

		Predicted[a]		
	N	Residency-Eligible	Residents	Students
Residency-Eligible	22	13(59.1)	9(40.9)	0(0)
Residents	36	5(13.9)	30(83.3)	1(2.8)
Students	22	0(0)	1(4.5)	21(95.5)

[a]Numbers in parentheses indicate percentage of classification for that group.

TABLE 4.22

SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS WITH LOW CLINICAL-RELEVANCE ENTERED FIRST

Step Number	Scale Name	F to Enter	Wilks' Lambda	p-value
1	Low Clinical-Relevance	6.07	.864	$\leq$ .004
2	High Clinical-Relevance	118.36	.210	$\leq$ .0001

in Table 4.22. The addition of the high clinical-relevance scale reduces Lambda to .210, showing that the high clinical-relevance scale adds very significantly to the statistical discrimination of these groups. Comparing Wilks' Lambda when the high and the low clinical-relevance scales are entered first in separate analyses, a large difference in Lambda (.629) is noted. This difference in Lambda favors the high clinical-relevance scale and indicates the relative contribution of high clinical-relevance to the statistical separation of groups, when compared to the low clinical-relevance scale.

Hypothesis Test

An F-test statistic was formed by the ratio of the univariate F's for each scale in this hypothesis, such that:

$$F_{calculated} = \frac{F_{High\ Clinical\text{-}Relevance}}{F_{Low\ Clinical\text{-}Relevance}} \qquad (4)$$

For the hypothesis of no difference in criterion-group discrimination between the high clinical-relevance and the low clinical-relevance scales:

$$F = \frac{125.21}{6.07} = 20.63$$

For 2 and 77 degrees of freedom, the critical value for rejection of H_o at $\alpha = .001$ is (conservatively) 7.76. Since F-calculated is greater than 7.76, the hypothesis of no difference between the high and the low clinical-relevance scales in group discrimination is rejected. There is a statistical difference in the discrimination power of these two scales; this difference favors the high clinical-relevance scale.

RESULTS CONCERNING CRITERION-RELATED VALIDITY

Alternative Hypothesis II for this study stated that the criterion-related validity--the correlation of subscale scores with grand mean ratings on the independent Simulated Clinical Encounters--of the high clinical-relevance scale would be higher than the validity coefficient for the medium difficulty scale. The method of item selection for the high clinical-relevance scale forced this subscale to have a high criterion-related validity coefficient. However, the validity of the medium difficulty scale--selected by a different criterion--could be equal to or lower than the validity of the high clinical-relevance scale.

Table 4.23 presents the criterion-related validity coefficients of all four scales in this study. Hypothesis II was tested by a Z-test of the difference of two non-independent correlation coefficients as presented by Glass and Stanley (1970). The Z-test statistic calculated to test the hypothesis of no difference against the hypothesis that the high clinical-relevance scale has a larger validity coefficient than the medium difficulty scale is given here:

$$Z_{calculated} = -3.60$$

For a one-sided (upper) hypothesis test, the critical value for rejection of the null hypothesis is 2.33 at $\alpha = .01$. Since the decision rule is: Reject H_o in favor of H_1 if Z>/C/, Hypothesis II is rejected in favor of H_1.

It may be concluded that an $r_{xy} = .895$ for the high clinical-relevance scale is greater than the $r_{xy} = .797$ for the medium difficulty scale.

TABLE 4.23

CRITERION-RELATED VALIDITY COEFFICIENTS:
SUBSCALE SCORE CORRELATION WITH
MEAN SIMULATION RATINGS
n=80

Subscale	Grand Mean Simulated Clinical Encounters
High Clinical-Relevance	.895
Medium Difficulty	.797
Low Difficulty	.774
Low Clinical-Relevance	.214

Although this hypothesis concerned differences in validity coefficients between the high clinical-relevance and the medium difficulty scales, it is interesting to note observed differences in validity for the other scales, as presented in Table 4.23. The low clinical-relevance scale has the lowest validity coefficient of the four scales; this result was anticipated, since items for this scale were chosen for their lowest correlation with the criterion. The low difficulty scale's validity coefficient (.774) is only slightly lower than the r_{xy} = .797 for the medium difficulty scale; this result is some surprising since it was anticipated that the low test score variance of this scale would attenuate the scale's correlation with the criterion.

RESULTS CONCERNING INTERNAL-CONSISTENCY RELIABILITY

Hypothesis III of this study stated that the internal-consistency reliability coefficient of the medium difficulty subscale would be higher than the reliability coefficient of the high clinical-relevance scale. Table 4.24 presents the reliability coefficients computed for each of the four subscales in this study.

Much research in educational and psychological measurement suggests that internal-consistency reliability will be maximized by selecting test items that cluster as closely as possible to p = q = .5. The reason for this phenomenon is that when p = .5, item variance is maximized (s_i^2=pq=.25) and test variance is therefore maximized, which tends to produce high internal-consistency reliability.

The hypothesis of no difference in scale reliability between the medium difficulty and the high clinical-relevance scale was tested

by a ratio of two F-values associated with each reliability coefficient (Wilson, 1978).

The test statistic is given by:

$$F_{calculated} = \frac{F_{High\ Clinical\text{-}Relevance(HCR)}}{F_{Medium\ Difficulty(MD)}} \qquad (5)$$

Where: $$F_{HCR} = \frac{Mean\ Square_{persons}}{Mean\ Square_{total}}$$

$$F_{MD} = \frac{Mean\ Square_{persons}}{Mean\ Square_{total}}$$

The logic underlying this test statistic is derived from the formula for Alpha or Kuder-Richardson 20 reliability, given by:

$$\alpha = \left(\frac{K}{K-1}\right) 1 - \frac{(\Sigma s_i^2)}{s_x^2} \qquad (6)$$

Where: K = Number of test items

s_i^2 = Item Variance

s_x^2 = Test Variance

Formula 6 can be computed from an Analysis of Variance of items and subjects (Hoyt, 1941), such that:

$$\alpha = \frac{MS_p - MS_r}{MS_p} \qquad (7)$$

Where: MS_p = Mean Square for Persons

MS_r = Mean Square Residual

A test statistic for the difference between two reliability coefficients is, therefore, given by:

$$F_{calculated} = \frac{MS_p/MS_t(HCR)}{MS_p/MS_t(MD)} \qquad (8)$$

Where: MS_t = Mean Square Total

TABLE 4.24

INTERNAL-CONSISTENCY RELIABILITY OF SUBSCALES
n=80

Subscale	Kuder-Richardson 20
High Clinical-Relevance	.954
Medium Difficulty	.878
Low Difficulty	.869
Low Clinical-Relevance	.581

TABLE 4.25

MEAN SQUARE VALUES FOR HIGH CLINICAL-RELEVANCE AND MEDIUM DIFFICULTY SCALES

Subscale	Mean Square Persons	Mean Square Total
High Clinical-Relevance	3.160	.201
Medium Difficulty	1.822	.239

Table 4.25 gives the Mean Square pieces for calculation of this F-test statistic.

For this hypothesis, then:

$$F_{calculated} = \frac{3.160/.201}{1.822/.239} = \frac{15.721}{7.623} = 2.062$$

Alternative Hypothesis III stated that the medium difficulty scale would be more reliable than the high clinical-relevance scale. Since the opposite direction was observed in the data, a two-sided hypothesis test is appropriate. Accordingly, for a two-sided test at $\alpha = .05$ with 79 and 79 degrees of freedom, the critical value (conservative) is 1.53. Since F-calculated is larger than the critical value, the null hypothesis is rejected in favor of an alternative that states that there is a difference in reliability between the medium difficulty and the high clinical-relevance scales. The high clinical-relevance scale is statistically more reliable than the medium difficulty scale.

RESULTS CONCERNING MEAN ITEM DIFFICULTIES

Hypothesis IV stated that both the medium difficulty subscale and the low clinical-relevance subscale would be more difficult than the high clinical-relevance subscale. The logic underlying this research hypothesis is that information that is relevant to the everyday practice of clinical medicine is used frequently and, therefore, remembered better than less frequently used knowledge.

Since mean item difficulty is a function only of the test mean and the number of items, the subscale means can be used to test an hypothesis about differences in mean item difficulty.

Table 4.26 presents means, standard deviations and mean p-values (mean proportion correct) for the four subscales of this study. Inspection of this table shows that the three subscale means are ranked in the order predicted by this hypothesis. That is, the medium difficulty scale is most difficult, followed by the low clinical-relevance and the high clinical-relevance scales. The low difficulty scale is not considered in this hypothesis since the item selection criteria used to select items for this subscale force it to be the least difficult. The medium difficulty scale is considered here because it acts as a difficulty reference point for the two scales selected by a different, independent criterion.

A repeated measures ANOVA of the three subscales of hypothesis IV reveals, in Table 4.27, a significant F-ratio. Tukey post-hoc analyses of the two contrasts of interest here show that the medium difficulty mean is significantly lower than the high clinical-relevance mean. Also, the low clinical-relevance mean is significantly lower than the high clinical-relevance mean.

These analyses support accepting alternative hypothesis IV that the medium difficulty and the low clinical-relevance scales are each statistically significantly more difficult than the high clinical relevance scale.

RESULTS CONCERNING OVERLAPPING ITEMS IN SUBSCALES

Hypothesis V for this study stated that the proportion of overlap of identical items between the medium difficulty and the high clinical-relevance subscales would be lower than the proportion of

TABLE 4.26

SUBSCALE MEAN ITEM DIFFICULTY
n=80

Subscale	Mean	Standard Deviation	Mean p-value
Low Difficulty	82.14	7.51	.903
High Clinical-Relevance	65.63	16.96	.721
Low Clinical-Relevance	60.69	6.32	.667
Medium Difficulty	54.86	12.88	.603

TABLE 4.27

REPEATED MEASURES ANOVA OF MEDIUM DIFFICULTY, HIGH AND LOW CLINICAL-RELEVANCE SUBSCALES
n=80

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between People	28624.73	79	362.34	
Within People--Between Measures	4632.71	2	2316.35	35.37*
Residual	10345.96	158	65.48	
TOTAL	43603.40	239	182.44	

*$p \leq .0001$

TUKEY POST-HOC ANALYSIS

Contrast	Difference of Means	$q_{3,77} \cdot SE$	Confidence Interval	Significance of Contrast
$\hat{\psi}_1 = \bar{X}_{HCR} - \bar{X}_{MD}$	10.75	3.08	$7.67 \leq \psi_1 \geq 13.83$	$p \leq .05$
$\hat{\psi}_2 = \bar{X}_{HCR} - \bar{X}_{LCR}$	4.94	3.08	$1.86 \leq \psi_2 \geq 8.02$	$p \leq .05$

overlap of identical items between the medium difficulty and the low clinical-relevance subscales. The logic of this research hypothesis is related to the logic of Hypothesis IV. That is, if medium difficulty items tend to be low in relevance to clinical medicine because of a lower frequency of use of information and knowledge--then, it is expected that there will be a smaller overlap of identical items between the high clinical-relevance and medium difficulty scales than between the low clinical-relevance and the medium difficulty scales.

Table 4.28 presents the number and proportion of identical items found between the scales noted in this hypothesis. The proportion overlap of identical items is slightly higher between the medium difficulty/high clinical-relevance scales than between the medium difficulty/low clinical-relevance scales.

This hypothesis was tested by drawing a 95 percent confidence interval around the difference of the two proportions (Bacon, 1976) of the form:

$$(p_1-p_2) \pm Z_{\alpha/2} \sqrt{\bar{p}_1\bar{q}_1/n_1 + \bar{p}_2\bar{q}_2/n_2} \qquad (9)$$

Where: p = proportion overlap

$$\bar{p} = \frac{p_1n_1 + p_2n_2}{n_1 + n_2}$$

n = number of items for proportion

$\bar{q} = 1-\bar{p}$

Testing this hypothesis, then:

$$95\%C = .055 \pm 1.96 \sqrt{.004}$$

$$= .055 \pm .123$$

TABLE 4.28

OVERLAP OF IDENTICAL ITEMS

Subscales	Identical Items	Proportion Overlap
Medium Difficulty and High Clinical-Relevance	24	.264
Medium Difficulty and Low Clinical-Relevance	19	.209

So that,

$-.068 \leq (p_1-p_2) \geq .178$

Since this confidence interval includes zero, the hypothesis of no difference in proportions can not be rejected. There is no statistically significant difference between the proportions of overlap in identical items between the medium difficulty/high clinical-relevance scales and the medium difficulty/low clinical-relevance scales.

RESULTS CONCERNING THE DISTRIBUTION OF ITEM TYPES IN SUBSCALES

Hypothesis VI for this study stated that there would be differences in the distributions of pictorial-stem, clinical-situational, and factual multiple-choice items selected for each of the four subscales studied here. It was expected that the high clinical-relevance scale would have a larger distribution of pictorial-stem and clinical-situational items than the medium difficulty or the low clinical-relevance scales.

Data analysis for this hypothesis required a classification and a count of the numbers of pictorial-stem, clinical-situational, and factual multiple-choice items for each of the four scales. The pictorial-stem items were easily classified; the only criterion used for this classification was the presence or absence of a visual stimulus with the item. If the item had a visual with it, it was classified as pictorial.

For the clinical-situational items, all multiple-choice items were inspected by two raters (the author and an educational specialist).

All items that met the following criteria were classified as clinical-situational:

1. The item stem contained clinical data about a patient's presenting signs or symptoms and/or other clinical data from a physical examination, laboratory studies, or any other information relative to a patient's presenting complaint.

2. The stem of the item ended with a question (or statement) asking for a diagnosis, a management strategy, or the next appropriate action to take for the patient(s).

The two raters disagreed on the classification of seven items; each of these disagreements was resolved by a third rater.

Figure 4.5 gives the distribution of item types by subscale. A chi-square test of independence was performed on the data given in Figure 4.5 to test the hypothesis of no difference in frequencies of item types by subscale.

The χ^2 test statistic computed for this hypothesis is given by:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E} \qquad (10)$$

Where: O = Observed frequency

E = Expected frequency

For these data, $\chi^2 = 11.22$ is less than the critical value of 12.59 for six degrees of freedom at $\alpha = .05$. Therefore, the hypothesis of no difference in item-type distributions can not be rejected. There is no statistically significant difference in the distributions of item types across the four subscales. However, the trend of the distribution of item types is that which was predicted by the research hypothesis.

OBSERVED DISTRIBUTIONS OF ITEM TYPES
BY SUBSCALE

	Pictorial Multiple-Choice[a]	Clinical-Situational	Factual Multiple-Choice
Medium Difficulty	20(22)	17(19)	54(59)
Low Difficulty	35(39)	22(24)	34(37)
High Clinical-Relevance	28(31)	21(23)	42(46)
Low Clinical-Relevance	22(24)	19(21)	50(55)

[a]Numbers in parentheses indicate percentage of item types in subscale.

FIGURE 4.5

Additional Results Concerning Differences in Proportions of Item Types

Figure 4.6 presents the combined percentage of pictorial-stem and clinical-situational items and the percentage of factual multiple-choice items for each of the four subscales under investigation here.

It is interesting to note that the high clinical-relevance scale has a slightly higher, but not statistically significant at $\alpha = .05$, proportion of pictorial and clinical-situational items than the medium difficulty scale. The low difficulty scale has the highest proportion of pictorial and clinical-situational items. This difference in percentage between the low difficulty and the high clinical-relevance scale is also not statistically significant at $\alpha = .05$, using a confidence internal procedure (Bacon, 1976) for differences in proportions.

The proportions of clinical-situational items alone (Figure 4.5) selected for the high clinical-relevance scale (.23) shows a slightly higher proportion of clinical-situational items in the high clinical-relevance scale, compared to the medium difficulty scale (.19). This difference in proportions is, however, not statistically significant at $\alpha = .05$.

Another analysis compared differences in proportions of combined pictorial-stem and clinical-situational items to factual multiple-choice items within each subscale. The results of these analyses are presented in the right-hand columns of Figure 4.6. These confidence intervals indicate that the low difficulty scale has significantly more pictorial-stem and clinical-situational items than

CONFIDENCE INTERVALS AROUND
DIFFERENCES IN PROPORTIONS OF ITEM TYPES
BY SUBSCALE

	Pictorial-Stem and Clinical-Situational	Factual Multiple-Choice	Difference	95% Confidence Interval Limits
Low Difficulty	.63	.37	.26	.11 to .41
High Clinical-Relevance	.54	.46	.08	-0.07 to .23
Low Clinical-Relevance	.45	.55	.10	-0.05 to .25
Medium Difficulty	.41	.59	.18	.03 to .33

FIGURE 4.6

factual multiple-choice items. And, the medium difficulty scale has significantly more factual multiple-choice items than pictorial-stem and clinical-situational items.

SUMMARY OF RESULTS FOR TESTS OF HYPOTHESES

Statistical analyses performed to test the hypotheses for this study may be summarized by the following:

1. Residency-eligible physicians, residents, and medical students are statistically discriminated by:
   a. The high clinical-relevance scale and the medium difficulty scale. The high clinical-relevance scale is more discriminating than the medium difficulty scale, but this difference in discrimination is not statistically significant at $\alpha = .05$. A total of 81.3 percent of subjects were correctly classified using both scale Discriminant Functions.
   b. The medium difficulty and the low clinical-relevance scale. The medium difficulty scale is statistically significantly (at $\alpha = .05$) more powerful than the low clinical-relevance scale in discriminating groups. The low clinical-relevance scale does not discriminate residents from students. The total of correct classifications for these two scales was 76.3 percent.

c. The high clinical-relevance and the low clinical-relevance scale. The high clinical-relevance scale is statistically significantly more powerful (at $\alpha = .05$) in discrimination than the low clinical-relevance scale. Eighty percent of subjects were correctly classified by these two scales.

2. Criterion-related validity: The correlation of the medium difficulty scores with the grand mean of Simulated Clinical Encounters ($r_{xy}=.878$) is statistically significantly lower than the criterion-related validity coefficient for the high clinical-relevance scale ($r_{xy}=.954$).
3. Internal-consistency reliability: The high clinical-relevance scale ($r_{xx}=.954$) is statistically significantly more reliable than the medium difficulty scale ($r_{xx}=.878$).
4. Mean item difficulty: The medium difficulty scale is statistically significantly more difficult than the high clinical-relevance scale. The low clinical-relevance scale is significantly more difficult than the high clinical-relevance scale.
5. Overlap of identical items in scales: There is no statistically significant difference in the proportions of overlap of identical items between the medium difficulty and high clinical-relevance scales and the medium difficulty and the low clinical-relevance scales.
6. Distribution of item types in scales: There is no statistically significant difference in the distributions

of pictorial-stem, clinical-situational, or factual multiple-choice items across the four subscales of this study.

RESULTS OF ADDITIONAL ANALYSES

The results of the data analysis performed to test the hypotheses for this study suggest several additional analyses. This section will present the findings of various additional analyses carried out to further explore these data. This section will be divided into three parts: 1) Results concerning the criterion-group discrimination of all four subscales taken together, 2) Results concerning the criterion-group discrimination of subscales using the practice-eligibles, residents, and students as the criterion-groups, and; 3) A validation of the findings about subscale criterion-related validity, reliability, and mean item difficulty with the n = 14 practice-eligible physician group.

Results Concerning the Criterion-Group Discrimination of Four Subscales

Several stepwise Discriminant Analyses using the medium and low difficulty, the high and the low clinical-relevance subscales as the discriminating variables and the residency-eligible, residents, and student criterion groups were performed. Table 4.29 compares the raw-score discriminating power of all four subscales. These data clearly show the rank-ordering of subscales according to their ability to discriminate known groups. This finding and the standardized Discriminant Function coefficients displayed in Table 4.30 show the power of the high clinical-relevance scale (relative to the other three

TABLE 4.29

RAW-SCORE GROUP DISCRIMINATION OF FOUR SUBSCALES

	Residency-Eligible		Residents		Students		Wilks' Lambda	$F_{2,77}$
	Mean	S.D.	Mean	S.D.	Mean	S.D.		
High Clinical-Relevance	80.23	7.67	70.81	7.59	42.55	9.98	.235	125.21
Medium Difficulty	66.77	8.67	56.83	7.89	39.77	7.14	.369	65.94
Low Difficulty	87.00	3.02	84.53	3.78	73.36	8.00	.457	45.75
Low Clinical-Relevance	64.27	6.38	59.97	6.08	58.27	5.24	.864	6.07

TABLE 4.30

STANDARDIZED DISCRIMINANT FUNCTION
COEFFICIENTS: FOUR SUBSCALES

	Function 1	Function 2
Medium Difficulty	0.209	1.689
Low Difficulty	0.305	-0.687
High Clinical-Relevance	-2.664	-1.118
Low Clinical-Relevance	0.252	0.549

TABLE 4.31

CLASSIFICATION ANALYSIS USING
TWO DISCRIMINANT FUNCTIONS
DERIVED FROM FOUR SUBSCALES

		Predicted[a]		
	N	Residency-Eligible	Residents	Students
Residency-Eligible	22	14(63.6)	8(36.4)	0(0)
Residents	36	4(11.1)	31(86.1)	1(2.8)
Students	22	0(0)	0(0)	22(100)

[a]Numbers in parentheses indicate percentage of group classified.

subscales) to discriminate groups with known levels of training and experience in Emergency Medicine. This finding is consistent with results presented for Hypotheses IA to IC.

Table 4.31 gives the classification analysis results, using the two Discriminant Functions derived for all four subscales. A total of 83.7 percent of the cases were correctly classified. A chi-square test was calculated for these data. With four degrees of freedom, $\chi^2 = 91.51$ is statistically significant at $p \leq .0001$.

Subscale correlations are presented in Table 4.32. The highest correlation is observed to be between the medium difficulty and the high clinical-relevance scales, while the lowest correlation is between the high clinical-relevance and the low clinical-relevance scale scores. Moderately high correlations are observed between medium difficulty and low difficulty and between high clinical-relevance and low difficulty.

In order to further investigate the inter-relationships of these subscales, first and second-order partial correlations were computed. The only zero-order correlation that is seriously decreased by controlling for other scale correlations is the high clinical-relevance --low clinical-relevance correlation. When the correlation with the medium difficulty scale is controlled in a first partial correlation, r decreases from .458 to -.214. There is also a large decrease of the zero-order correlation of high and low clinical-relevance when the correlation with the low difficulty scale is controlled. When the correlation with both the medium and the low difficulty scales is controlled in the second partial correlation, r drops to -.336 from

TABLE 4.32

SUBSCALE ZERO-ORDER CORRELATIONS
n=80

	Medium Difficulty	Low Difficulty	High Clinical-Relevance
Medium Difficulty	1.000		
Low Difficulty	.772	1.000	
High Clinical-Relevance	.922	.879	1.000
Low Clinical-Relevance	.571	.465	.458

r = .458. This finding suggests that the observed moderate correlation between the high and low clinical-relevance scales is spurious and due to the correlation of each of these scales with the medium difficulty and the low difficulty scales.

Results Concerning the Criterion-Group Discrimination of Subscales Using Different Criterion Groups

Hypotheses IA, IB, and IC were reanalyzed using the same subscales as the discriminating variables in the analyses, but substituting the practice-eligible group for the residency-eligible group as the independent variable. These analyses were carried out to attempt a partial validation of the previous results of this study with a group of subjects who were not considered in the item analysis criteria used to select subscale items.

Table 4.33 summarizes the results of these analyses. When practice-eligible physicians are substituted for residency-eligible physicians in Discriminant Analyses:

1. The four subscales are rank-ordered in discriminating power in exactly the same manner as in the earlier analyses using residency-eligibles (Table 4.29). That is, high clinical-relevance is the best discriminator, followed by medium difficulty, low difficulty, and low clinical-relevance.

2. Most of this discrimination occurs between the resident and student groups; none of these scales statistically separates the practice-eligible group from the resident group. In fact, the practice-eligible group has a slightly lower mean on the high clinical-relevance and the low difficulty scales than the resident group.

Results Concerning Criterion-Related Validity: Practice-Eligible Group

Correlations between the subscales of this study and the grand mean of the Simulated Clinical Encounters were computed for the

TABLE 4.33

RAW-SCORE DISCRIMINATION USING THE PRACTICE-ELIGIBLE GROUP:
FOUR SUBSCALES

	Practice-Eligible		Residents		Students		Wilks' Lambda	F2, 69
	Mean	S.D.	Mean	S.D.	Mean	S.D.		
High Clinical-Relevance	69.71	10.09	70.81	7.59	42.55	9.98	.312	75.93
Medium Difficulty	58.57	7.40	56.83	7.89	39.77	7.14	.456	41.22
Low Difficulty	82.64	5.94	84.53	3.78	73.36	8.00	.566	26.43
Low Clinical-Relevance	63.57	4.24	59.97	6.08	58.27	5.24	.897	3.98

practice-eligible group (n=14) alone. Table 4.34 presents these criterion-related validity coefficients. The rank-ordering of these validity coefficients is different than the rank-ordering of the validities for the n=80 sample of residency-eligible, residents, and students (Table 4.23).

For the practice-eligible group, the low clinical-relevance scale has the highest validity coefficient (r_{xy}=.49), while the high clinical-relevance scale had the highest validity for the n=80 sample (r_{xy}=.90). The high clinical-relevance scale's validity was ranked second for practice-eligible physicians, while medium difficulty ranked second for the larger group. Other reversals also are noted in Table 4.34. These inconsistencies are likely the result of large standard errors around r computed for a small sample of subjects.

Results Concerning Internal-Consistency Reliability of Subscales: Practice-Eligible Group

Table 4.35 presents a comparison of the Kuder-Richardson 20 reliabilities computed for the practice-eligible group alone and for the residency-eligible, residents, and students (n=80) for all four subscales in this study.

Only two reversals of reliability ranks are noted in Table 4.35. That is, the reliability of the low difficulty scale is second in rank for the practice-eligible group and third for the previous (n=80) group. Medium difficulty is third in rank for the small group and second in rank for the larger group.

The F-test of the difference between the high clinical-relevance and medium difficult reliability coefficients yielded:

TABLE 4.34

COMPARISONS OF
CRITERION-RELATED VALIDITY COEFFICIENTS

	Grand Mean Simulated Clinical Encounters	
	Practice-Eligible n=14	Others n=80
Low Clinical-Relevance	.487	.797
High Clinical-Relevance	.461	.895
Low Difficulty	.441	.774
Medium Difficulty	.244	.797

TABLE 4.35

COMPARISON OF INTERNAL-CONSISTENCY RELIABILITY OF SCALES:
PRACTICE-ELIGIBLE AND PREVIOUS SAMPLE

	Practice-Eligible n=14	Others n=80
High Clinical-Relevance	.868	.954
Low Difficulty	.796	.869
Medium Difficulty	.644	.878
Low Clinical-Relevance	.117	.581

$$F_{calculated} = \frac{1.118/.179}{.601/.230} = 2.39$$

At 13 and 13 degrees of freedom, the (conservative) critical value at $\alpha = .05$ is 2.69. Since F-calculated is less than the critical value, the hypothesis of no difference between the reliability of the high clinical-relevance and medium difficulty scales for this group can not be rejected.

Results Concerning Mean Item Difficulties: Practice-Eligible Group

Table 4.36 presents a comparison of the means, standard deviations, and mean p-value for the practice-eligible group and the group of residency-eligible, residents, and students (n=80). This table shows that the means of subscales are ranked in exactly the same order for the practice-eligible group as for the larger group of eighty subjects.

In Table 4.37, the results of a repeated measures Analysis of Variance for the medium difficulty, the high and low clinical-relevance scales for the practice-eligible group (n=14) are given. There are statistical differences in these three means revealed by the significant F-ratio. Tukey post-hoc analysis of differences between the high clinical-relevance and medium difficulty means and the high clinical-relevance and low clinical-relevance means shows that these contrasts of means are significantly different at the $\alpha = .05$ level. Thus, for the practice-eligible group, the high clinical-relevance scale is significantly easier than the medium difficulty scale and the low clinical-relevance scales. This finding is the same as the earlier result with the residency-eligible, resident, and student groups.

TABLE 4.36

COMPARISON OF SUBSCALE MEAN ITEM DIFFICULTY:
PRACTICE-ELIGIBLE GROUP (n=14) VS PREVIOUS GROUP (n=80)

SUBSCALES	PRACTICE-ELIGIBLE (n=14)			OTHERS (n=80)		
	Mean	S.D.	Mean p-value	Mean	S.D.	Mean p-value
Low Difficulty	82.64	5.94	.908	82.14	7.51	.903
High Clinical-Relevance	69.71	10.09	.766	65.63	16.96	.721
Low Clinical-Relevance	63.57	4.24	.699	60.69	6.32	.667
Medium Difficulty	58.57	7.40	.644	54.86	12.88	.603

TABLE 4.37

REPEATED MEASURES ANOVA OF MEDIUM DIFFICULTY, HIGH AND LOW CLINICAL-RELEVANCE SCALES: PRACTICE-ELIGIBLE GROUP n=14

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between People	1629.90	13	125.38	
Within People-Between Measures	872.19	2	436.10	17.78*
Residual	637.81	26	24.53	
TOTAL	3139.90	41	76.58	

*$p \leq .0001$

TUKEY POST-HOC ANALYSIS

Contrast	Difference of Means	$q_{2,11} \cdot SE$	Confidence Interval	Significance of Contrast
$\hat{\psi}_1 = \bar{X}_{HCR} - \bar{X}_{MD}$	11.14	5.06	$6.08 \leq \psi_1 \geq 16.20$	$p \leq .05$
$\hat{\psi}_2 = \bar{X}_{HCR} - \bar{X}_{LCR}$	6.14	5.06	$1.08 \leq \psi_2 \geq 11.20$	$p \leq .05$

SUMMARY RESULTS OF ADDITIONAL ANALYSES

Several additional analyses were suggested by the results of the hypothesis tests reported here. The results of these additional data analyses are summarized below:

1. A Discriminate Analysis of all four subscales of this study shows that the high clinical-relevance subscale is the best discriminator of the residency-eligible, resident, and student groups. The medium difficulty subscale is ranked second in its power of discrimination for these criterion-groups. The low difficulty and the low clinical-relevance scales are ranked third and fourth, respectively.

2. All four subscale scores discriminate groups statistically. Only the low clinical-relevance scale fails to statistically separate residents and students.

3. Use of all four subscale Discriminant Functions classifies 83.7 percent of subjects correctly; use of just the high clinical-relevance and the medium difficulty scales classifies 81.3 percent of cases correctly.

4. These four subscales are moderately to highly intercorrelated. Partial correlations show that the moderate correlation of the high and the low clinical-relevance scales is a function of scale correlations with the medium difficulty and the low difficulty scales.

5. Attempts to partially validate the previous findings of this study by reanalyzing data substituting the practice-eligible group (that was not considered in subscale development analyses) for the residency-eligible group showed the following:

    a. None of the subscales discriminates the practice-eligible group from residents.

    b. Discriminant Analysis ranks the discriminating power of the four scales in the same order as did the primary analyses of this study. However, nearly all this discrimination occurs between students and residents, groups that were included in the primary study.

    c. Scale criterion-related validity coefficients for the practice-eligible group taken separately are rank-ordered differently than for the residency-eligible, resident, and student groups of the primary study.

d. There is no statistically significant difference in internal-consistency reliability between the high clinical-relevance and the medium difficulty subscales for the practice-eligible group taken separately.

e. The mean item difficulties of these four subscales are ranked identically for the practice-eligible group and the larger group used for hypothesis testing in the primary study.

Chapter V discusses these findings, draws conclusions and makes suggestions for further research.

CHAPTER V

SUMMARY AND CONCLUSIONS

This chapter summarizes the present research study and draws conclusions based on the results. These conclusions are discussed and suggestions are made for future research, based on the findings of this study.

SUMMARY OF FINDINGS

This research was designed to study the effect of two different item selection strategies on the psychometric properties of four objective-item subscales. This study was carried out in the context of a field test of an item library for certification of specialists in Emergency Medicine. Items were selected for two subscales using an item difficulty criterion (medium and low difficulty scales); two additional ninety-one item scales were selected using an empirically defined clinical-relevance criterion (high and low clinical-relevance scales). For this study, item relevance to clinical Emergency Medicine was defined empirically as high or low item score correlation with an independent criterion measure of simulated clinical performance.

Ninety-four subjects, representing four distinct groups in training, experience, and known ability to deliver emergency medical health care, were administered all examination materials. The four groups were:

1. Residency-eligible emergency physicians
2. Practice-eligible emergency physicians
3. Second-year residents in Emergency Medicine
4. Fourth-year medical students

These four subscale scores were statistically analyzed to test hypotheses about:

1. The relative strength of subscales selected for item difficulty or clinical-relevance to statistically discriminate known groups of subjects.

2. Differences in subscale criterion-related validity--correlation of subscale scores with the grand mean rating on Simulated Clinical Encounters.

3. Differences in internal-consistency reliability between subscales of items selected for item difficulty and clinical-relevance.

4. Differences in mean item difficulty among subscales of items selected by the two different methods.

5. Differences in the proportions of overlapping identical items in subscales selected by the two different item selection strategies.

6. Differences in the distributions of factual multiple-choice, pictorial-stem, and clinical-situational items across the four subscales of this study.

Discriminant Analyses showed that the high clinical-relevance scale was better than the medium difficulty scale in discriminating the residency-eligible physician, the resident, and the student groups. However, this difference in discrimination was not statistically significant at an alpha level of .05. But, the high clinical-relevance scale was significantly better (at $\alpha = .05$) in the discrimination of the three groups of subjects than the low clinical-relevance scale.

Additional Discriminant Analyses showed that the medium difficulty subscale was significantly more effective (at $\alpha = .05$) in statistically separating these three groups than the low clinical-relevance subscale.

The criterion-related validity of the high clinical-relevance subscale ($r_{xy} = .90$) was statistically higher (at $\alpha = .01$) than the

validity of the medium difficulty scale (r_{xy} = .80). This result was anticipated, since high clinical-relevance items were selected for their high correlation with the same independent criterion of grand mean ratings on Simulated Clinical Encounters as was used for computing criterion-related validity coefficients. The low difficulty scale also had a high validity coefficient (r_{xy} = .77). The low clinical-relevance items--selected for their low correlation with the Simulated Clinical Encounter criterion--as anticipated, had the lowest validity coefficient (r_{xy} = .21).

The high clinical-relevance subscale had the highest internal-consistency reliability of the four subscales (r_{xx} = .95). The difference between the reliability coefficients for the high clinical-relevance and the medium difficulty scale (r_{xx} = .88) was significant at an alpha level of .05.

A one-way repeated measures ANOVA of the high and low clinical-relevance and the medium difficulty scale scores showed that the medium difficulty scale was more difficult than the high clinical-relevance scale and the low clinical-relevance scale was more difficult than the high clinical-relevance scale. This finding supports an hypothesis that clinically relevant knowledge is used more frequently and is, therefore, retained better by physicians than isolated factual knowledge.

There were no significant differences in the proportions of over-lapping identical items between the medium difficulty/high clinical-relevance scales and the medium difficulty/low clinical-relevance scales. It had been hypothesized that there would be a larger overlap

of identical items selected for the medium difficulty/low clinical-relevance scales than for the medium difficulty/high clinical-relevance scales. The rationale for this hypothesis was: if the selection of certification examination items for a criterion of middle difficulty lessens the relevance of item content to the practice of clinical medicine, then there should be a larger number of identical items shared by a medium difficulty and a low clinical-relevance scale than by a medium difficulty and a high clinical-relevance scale. This hypothesis was not supported by this research. There were only slightly fewer low clinical-relevance items (21 percent) than high clinical-relevance items (26 percent) selected for the medium difficulty scale.

There was no statistical difference in the distributions of factual multiple-choice, pictorial-stem, or clinical-situational item types observed across the four subscales for this study. However, a trend was observed in the data which suggested that pictorial-stem and clinical-situational items were selected at a slightly higher frequency for the high clinical-relevance and the low difficulty scales than for the low clinical-relevance and the medium difficulty scales.

A Discriminant Analysis of all four subscales taken together showed the high clinical-relevance scale to be the best discriminator of the residency-eligible, resident, and student groups. The medium difficulty, low difficulty, and low clinical-relevance scales

follow in their relative discriminating ability.[5] Each of the four subscales statistically discriminated these groups. However, the low clinical-relevance scale failed to statistically separate residents from students.

Noting that the inter-scale correlation coefficients are inflated by auto-correlation (since there is some overlap of identical items between scales), the four subscales are moderately to highly inter-correlated (r = .46 to .92). This finding is consistent with previous research which showed a large general competence factor across items and formats of this Examination (Maatsch et al., 1978).

A partial validation of the results of the present study, using the practice-eligible group (n=14)--who were not considered in subscale development analyses--yielded mixed results. Discriminant Analysis of the four subscales, using the practice-eligible, resident, and student groups, ranked the discriminating ability of the four scales in the identical order as in the primary analysis. However, none of the sub-scales statistically discriminated practice-eligible physicians from second-year residents. This finding is also consistent with previous research on this Examination item pool (Maatsch et al., 1978).

Criterion-related validity coefficients for practice-eligibles taken separately were ranked differently than in the primary analysis. The low clinical-relevance subscale ranked first in criterion-related

[5]Differences in relative discriminating power of the four subscales are not due to differences in reliability of these scales. In one analysis, unreported here, scores of the low clinical-relevance and the high clinical-relevance scales (scales that differ most in reliability) were corrected for attenuation due to unreliability and reanalyzed. The result of this analysis was identical to the result of the analysis of uncorrected raw scores.

validity (r_{xy} = .49) for the practice-eligibles alone, while the high clinical-relevance validity coefficient (r_{xy} = .46) ranked second. These differences in scale validities were likely due to the large sampling error around r for n = 14. The generally lower validity coefficients observed for the practice-eligibles taken separately exemplify the well known validity coefficient shrinkage phenomenon of cross-validation (Magnusson, 1967), and the possible effect on correlation of reduced score variance from small samples.

The high clinical-relevance subscale was most reliable for the practice-eligible group taken separately, as it was for the n = 80 group in the primary study. However, in contrast to the finding in the primary analysis, there was no statistical difference in the internal-consistency reliability coefficients of the high clinical-relevance and the medium difficulty scales for the practice-eligible group taken separately.

The mean item difficulties of the four subscales were ranked in the identical order for the practice-eligible group taken separately as they were in the primary study.

CONCLUSIONS

1. All four multiple-choice subscales selected for study here statistically discriminate groups of subjects with known capabilities to deliver health care. A subscale selected for low clinical-relevance, however, does not discriminate residents from fourth-year medical students. Neither subscales selected for high or low clinical-relevance nor subscales selected for medium or low item difficulty

adequately discriminate practice-eligible physicians from second-year residents in Emergency Medicine.

2. A clinically relevant multiple-choice subscale, composed of items selected for their high correlation with independent ratings of simulated clinical performance, yields a score which, in the context of a medical specialty certification examination, is the best discriminator of independent criterion groups with known levels of training, experience, and ability to deliver health care.

   a. Clinically relevant subscale scores are substantially more discriminating of criterion groups than subscales of items selected for their low difficulty or their low clinical-relevance.

   b. Clinically relevant subscale scores are slightly more discriminating of criterion groups than medium difficulty subscale scores.

   c. Clinically relevant subscale scores are substantially higher in criterion-related validity than subscales of items selected by a middle or low difficulty or a low clinical-relevance item selection strategy.

   d. Clinically relevant subscale scores are substantially more reliable than subscales of items selected by a middle or low difficulty or a low clinical-relevance criterion.

   e. Clinically relevant subscale scores are substantially less difficult than either a medium difficulty or a low clinical-relevance subscale.

3. Use of a medium difficulty item analysis strategy to choose certification examination items tends to select relatively small and approximately equal proportions of high and low clinical-relevance items for the subscale.

4. Neither the pictorial-stem nor the clinical-situational item type is definitively associated with subscales selected by an item difficulty or a clinical-relevance criterion.

5. Use of a medium difficulty item selection strategy tends to choose substantially more factual multiple-choice than pictorial-stem and clinical-situational items for a scale. A low difficulty item selection criterion tends to choose substantially more pictorial-stem and clinical-situational items than factual multiple-choice items.

   a. Pictorial-stem and clinical-situational items tend to be less difficult than factual multiple-choice items.

   b. There are only slight differences in the proportions of pictorial-stem and clinical-situational items selected by the high and the low clinical-relevance criterion.

   c. Item type can not be used as a definitive indicant of item relevance to the practice of clinical medicine, but trends suggest that slightly higher proportions of pictorial-stem and clinical-situational item types, rather than factual multiple-choice items, will be chosen by a high clinical-relevance item selection strategy.

6. Subscales of items selected for high clinical-relevance and for medium difficulty have moderate to high positive correlations.

7. Several findings of this research withstand validation using a small group of subjects not considered in any of the scale construction analyses.

Generalizations of the results of this study are limited to the populations of subjects randomly sampled for this research. A conservative statistical view suggests that the results of this study should be generalized only to the population of residency and practice-eligible emergency physicians who are judged by their peers as "certifiable" and to second-year residents in Emergency Medicine. However, using the arguments advanced by Cornfield and Tukey (1956), the reader may make inferences to populations of subjects who have similar characteristics to subjects sampled here. Thus, cautious generalizations of these results could be made to candidates for certification in other medical specialties and to other certification examinations that are similar in format and content to the Emergency Medicine Examination.

DISCUSSION

Non-Significant Results

Several null hypotheses for this study could not be rejected by the statistical tests used. Specifically, the hypothesis of no difference in discriminating power of the medium difficulty and the high clinical-relevance scales could not be rejected. Although the data indicated that the high clinical-relevance scale was most powerful in its discrimination of groups, the F-test used to test the null hypothesis was not powerful enough (at n=80) to reach a conventional level of significance. A larger sample of subjects might have shown

the high clinical-relevance scale to be a significantly better discriminator of known groups than the medium difficulty scale.

The hypothesis concerning overlapping identical items was also not rejected. It had been expected that if selection of items for medium difficulty distorts the clinical-relevance of item content, then the medium difficulty scale would share more identical items with the low clinical-relevance scale than with the high clinical-relevance scale. The anticipated effect was not found in these data. One reason for this failure to reject this null hypothesis may be that item writers for this Examination produced clinically relevant questions that tested knowledge judged as absolutely essential to the competent practice of Emergency Medicine and, consequently, the item pool may have been too homogeneous with respect to clinical relevance to detect any effect. The size of the criterion-related validity coefficients for all four scales supports this rationale. It is especially interesting to note the value of criterion-related validity coefficient for the low clinical-relevance scale (r_{xy} = .21). Even the subscale of items chosen for their lowest correlation with the criterion produces a validity coefficient that is at least low positive and that is not much lower than the highest validity coefficients found in some previous studies (e.g., Levine and McGuire, 1970).

The hypothesis of no differences in the distributions of pictorial-stem, clinical-situational, and factual multiple-choice items across subscales selected by different criteria could not be rejected. The trend of these data shows a few more pictorial-stem and clinical-situational items chosen for the high clinical-relevance

than for the low clinical-relevance or the medium difficulty scales. However, most pictorial-stem and clinical-situational items were chosen for the low difficulty scale. Thus, the generally low difficulty of all items may be confounding an effect. The point-biserial correlation coefficient is influenced by item difficulty and favors medium difficulty items (Henrysson, 1971). Since pictorial-stem and clinical-situational items tend to be lower in difficulty than factual items, a bias may have been introduced that systematically reduced the numbers of pictorial-stem and clinical-situational items selected for the high clinical-relevance scale and, thus, masked an effect. Another possible explanation for failure to reject this null hypothesis may be that clinically based items were defined too narrowly. That is, this research examined the effect of items that had visual stimuli or were classified as clinical-situational. If the definition of clinical-situational items had been broadened to include all items that had any content base in clinical medicine, then the anticipated effect would almost certainly have been detected.

Significant Results

There was a significant difference in internal-consistency reliability of the high clinical-relevance and the medium difficulty scales in this study. The high clinical-relevance scale (r_{xx} = .95) was significantly more reliable than the medium difficulty scale (r_{xx} = .88). This result was not hypothesized, since classical psychometric theory suggests that middle difficulty items will yield the most reliable measurements (Magnusson, 1967). This finding is most

easily explained by the following: the high clinical-relevance scale's items were chosen for their highest correlation with the criterion of Simulated Clinical Encounter mean ratings. This selection procedure forced the items of this scale to be very homogeneous with respect to whatever the Simulated Clinical Encounters measure and, therefore, to correlate highly with each other. High homogeneity of items produces high Alpha reliability (Cronbach, 1951); therefore, it is not surprising that the high clinical-relevance scale has the highest internal-consistency reliability.

The difference in criterion-related validity between the high clinical-relevance and the medium difficulty scales was statistically significant. Since the high clinical-relevance scale was chosen for its high item correlation with the criterion, it was not surprising that this scale had the highest criterion-related validity coefficient. It is interesting to note, however, that both the medium and low difficulty scales also had relatively high ($r_{xy} \tilde{=} .80$) criterion-related validity coefficients. This finding is consistent with the earlier research of Maatsch and others (1978), but is inconsistent with much other research (e.g., Williamson, 1976) showing the lack of concurrent or predictive validity for various certification examinations.

There are several possible explanations for this finding:

1. All objective items of the Emergency Medicine Examination may be more clinically relevant than the items of other certification examinations studied. The $r_{xy} = .21$ for the low clinical-relevance scale supports this hypothesis.

2. The criterion used for the present study may be more reliable (inter-rater r_{xx} = .79) than criteria used in other studies (e.g., LeVine and McGuire, 1970).
3. The criterion used for the present study may be more valid than criteria of clinical performance used for other studies (e.g., Burg and Schumacher, 1979; Williamson, 1976).
4. The range and variance of scores in both the objective scales and the criterion measure is greater than observed in other studies (e.g., Pawluk, 1976). The medical student group included in the present study increased the variance in both score distributions, and thus, permitted higher scale-criterion correlations than observed in other studies.

The wider range of clinical competence sampled in the present study may limit the usefulness of these results for other specialty certification boards. The reader must also note that this study sampled only ninety-four subjects who were carefully selected primarily to field test and calibrate a pool of items for a new specialty certification examination in Emergency Medicine. The Cornfield-Tukey (1956) argument, therefore, may have only limited usefulness for some readers; all inferences to other populations should be very cautious and limited.

Implications of This Study

The major goal of the present research was to determine the multiple-choice item-selection strategy that would produce the highest quality psychometric scale for a certification examination. Another goal of this research was to learn something about how to create

multiple-choice items that would perform well in medical specialty certification examinations. This second goal met with only limited success.

The results of this study indicate that objective items for certification which are clinically relevant perform very well: these items discriminate known criterion groups best, are most reliable, have high criterion-related validity (by definition), are less difficult than many items, and consist of relatively fewer factual multiple-choice items than other scales which perform less well.

Theoretically, then, this study has shown:

1. Strong support for the view that well written objective test items that measure important applications of knowledge to novel situations (Ebel, 1972) are useful, efficient, and cost-effective methods of measuring important outcomes of specialized professional education.

2. That selection of examination items for medium difficulty is not the most efficient strategy to employ to maximize the valid discrimination of known groups, criterion-related validity, or internal-consistency reliability.

3. That construction of objective items to maximize content validity may not be efficient to maximize discriminant and criterion-related validity.

However, practically, the empirical methodology used to select the high clinical-relevance items for this study may have only limited usefulness for constructors of other certification examinations. Many examination constructors do not have a valid and reliable independent measure of simulated clinical performance available to them against which objective items can be calibrated.

Insofar as generalizations of these findings can be made to other certification situations, the following points of advice to multiple-choice item writers seem appropriate from the present study:

1. Choose item content that attempts to measure the most essential aspects of specialized knowledge which is clearly related to typical clinical practice in the specialty.

    a. Items that present clinical situations in the stem and ask examinees to select the most appropriate diagnosis or treatment may be slightly more clinically relevant than items that ask for factual information.

    b. Items that ask for interpretations of visual stimuli may be slightly more clinically relevant than factual items.

2. Selection of items for middle difficulty may not be the most efficient item analysis criterion to use.

    a. Calibrate certification items against a valid and reliable measure of clinical performance, if possible.

    b. If a clinical-performance criterion is unavailable, select items that appear to be related to clinical medicine and are of reasonable difficulty.

    c. Items that are highly related to a clinical performance criterion and of middle difficulty may be most psychometrically powerful.

FUTURE RESEARCH

Much additional research is suggested by the findings of the present study. First, this study should be replicated with larger samples of subjects and with samples of other specialty populations. The results of this study withstood, to a limited degree, a small validation study using the practice-eligible group; this finding suggests that some of the results may withstand replications, validation, and cross validation.

Further research is needed in the larger area of validation of certification procedures in medicine. Concurrent and predictive validity studies, which attempt to discover the real-world correlates

to certification scores, are needed. Further, research should be undertaken that will definitively assess the types of certification methods most useful to the prediction of actual clinical performance.

Studies of the relationship of content validity of certification examinations to the prediction of clinical competence are needed. Most certification examinations in medical specialties rely heavily on the content validity of their measurement procedures to discriminate levels of physician competence. This study and previous research (e.g., Williamson, 1976) suggest that a content valid examination does not necessarily best predict clinical performance. Yet, what are the implications of seriously distorting the content validity of a certification examination in order to gain criterion-related validity (concurrent and/or predictive)? What are the legal implications of such a test construction strategy?

This study has shown that clinically relevant items are most discriminating of levels of simulated clinical performance. Research replicating this finding with actual clinical performance is needed. Before such studies can be carried out adequately, other research must be done to discover valid and reliable methods to define and measure the criterion--competent clinical performance.

Other practical research must be done to discover the rules for specifying the proper content of high-quality clinically relevant objective items. The present research could not adequately define for future item writers specific content-selection rules for producing clinically relevant items. What general rules can be discovered about the content being measured by the high clinical-relevance scale of

this study? Are the correlations of items with the criterion in this high clinical-relevance scale meaningful or merely accidental? How, specifically, can certification examination item writers increase the probabilities of producing items that are highly related to clinical medicine?

BIBLIOGRAPHY

BIBLIOGRAPHY

Bacon, F.R. Statistical Concepts in Business: A Scientific Problem Solving Approach, Columbus, Ohio: Grid Publishing Co., 1976.

Burg, F.D.; Guerin, R.O.; Schumacher, C.F. Use of Pre-Test and Post-Test Examinations to Evaluate Cognitive Knowledge for Pediatric Residents, Chicago, Illinois: American Board of Pediatrics, 1977.

Burg, F.D.; Schumacher, C. "Standardized Tests as Measures for Medical Certification", Professions Education Researcher, Vol. 1, No. 1, 1979, pp. 13-17.

Carter, H.D. "How Reliable are Good Oral Examinations?" California Journal of Educational Research, Vol. 13, No. 4, 1962, pp. 147-153.

Conference on Extending the Validity of Certification, Chicago, Illinois, American Board of Medical Specialties, 1976.

Cornfield J.; Tukey, J.W. "Average Values of Mean Squares in Factorials," The Annals of Mathematical Statistics, Vol. 27, 1956, pp. 907-949.

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests," Psychometrika, Vol. 16, 1951, pp. 297-334.

Downing, S.M. Multiple-Choice Item Writing Handbook, East Lansing, Michigan: Michigan State University, Office of Medical Education Research and Development, March, 1977.

Ebel, R.L. Essentials of Educational Measurement, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

_____. "Estimation of the Reliability of Ratings," Psychometrika, Vol. 16, 1951a, pp. 407-424.

_____. "Must all Tests be Valid?" American Psychologist, Vol. 16, 1961, pp. 640-647.

_____. "Obtaining and Reporting Evidence on Content Validity," Educational and Psychological Measurement, Vol. 16, 1956, pp. 269-282.

_____. "A Proposed Solution to the Validity Problem," Address at the 1978 National Council on Measurement in Education Annual Meeting, Toronto, Ontario, Canada, 1978.

_____. "The Relation of Item Discrimination to Test Reliability," Journal of Educational Measurement, Vol. 4, 1967, pp. 125-128.

_____. "Writing the Test Item," in Educational Measurement, ed. E.F. Lindquist, 1st edition, Washington, D.C.: American Council on Education, 1951b.

"Emergency Medicine Condition/Skills List," Journal of the American College of Emergency Physicians, August, 1976.

Flexner, J.T. Medical Education in the United States and Canada. New York: Carnegie Foundation for the Advancement of Teaching, Bulletin No. 4, 1910.

Glass, G.V.; Stanley, J.C. Statistical Methods in Education and Psychology, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1970.

Gonnella, J.S. "Evaluation of Patient Care: An Approach," Journal of the American Medical Association, Vol. 214, No. 11, 1973, pp. 2040-2043.

Guion, R.M. Personnel Testing, New York: McGraw-Hill Book Co., 1965.

Handbook for Certification in Family Medicine, Revised Edition, Willowdale, Ontario: The College of Family Physicians of Canada, 1976.

Hechel, H.; Bowles, L.T. "Specialty Certification in North America: A Comparative Analysis of Examiantion Results," Journal of Medical Education, Vol. 54, 1979, pp. 69-74.

Henrysson, S. "Gathering, Analyzing, and Using Data on Test Items," in Educational Measurement, Second Edition, ed. R.L. Thorndike, Washington, D.C.: American Council on Education, 1971.

Hoyt, C.J. "Test Reliability Estimated by Analysis of Variance," in Principles of Educational and Psychological Measurment, eds. W.A. Mehrens and R.L. Ebel, Chicago, Illinois: Rand McNally and Company, 1967.

Hubbard, J.P.; Clemans, W.V. Multiple-Choice Examinations in Medicine: A Guide for Examiner and Examinee. Philadelphia, Pennsylvania: Lea and Febiger, 1961.

Hubbard, J.P. Measuring Medical Education. Philadelphia, Pennsylvania: Lea and Febiger, 1971.

Kaplan, H.E.; Freedman, A.M.; Kaplan, H.S. "The Evaluation of Psychiatric Residents by Objective Multiple-Choice Examinations," American Journal of Psychiatry, Feb. 1968, pp. 128-132.

Kelley, P.R.; Levit, E.J. "A Three Year Study of the Internship in Air Force Hospitals condu-ted by NBME," Washington, D.C.: Report to the Surgeon General, Department of the Air Force, June, 1967.

Kelley, P.R.; Matthews, J.H.; Schumacher, C.F. "Analysis of the Oral Examination of the American Board of Anesthesiology," Journal of Medical Education, Vol. 46, Nov. 1971, pp. 982-988.

Kelley, P.R.; Stumpe, A.R.; Levit, E.J. "A Four-Year Study of the Internship in United States Air Force Hospitals: An Objective Measure of Gain in Clinical Competence," Military Medicine, Vol. 135, 1970, pp. 537-545.

Kessner, D.M. A Strategy for Evaluating Health Services: Contrasts in Health Status. Washington, D.C.: National Academy of Science, Institute of Medicine, 1973.

Levine, H.G.; McGuire, C.H.; Nattress, L.W. "The Validity of Multiple-Choice Achievement Tests as Measures of Competence in Medicine," American Educational Research Journal, Vol. 7, 1970, pp. 69-82.

Levine, H.G.; McGuire, C.H. "The Validity and Reliability of Oral Examinations in Assessing Cognitive Skills in Medicine," Journal of Educational Measurement, Vol. 7, No. 2, 1970, pp. 63-74.

McGuire, C.H. "The Oral Examination as a Measure of Professional Competence," Journal of Medical Education, Vol. 41, 1966, pp. 267-274.

McGuire, C.; Williamson, J.W. "Consecutive Case Conference: An Educational Evaluation," Journal of the American Medical Association, Vol. 43, No. 10, 1968, pp. 1068-1074.

Maatsch, J.L.; et al. "The Emergency Medicine Specialty Certification Examination (EMSCE)," Journal of the American College of Emergency Physicians, July 1976.

_____. An Introduction to Patient Games: Some Fundamentals of Clinical Instruction, East Lansing, Michigan: Michigan State University Office of Medical Education Research and Development, 1974.

_____; et al. "Toward a Testable Theory of Physician Competence: An Experimental Analysis of a Criterion-Referenced Specialty Certification Test Library," in Proceedings of the Seventeenth Annual Conference on Research in Medical Education, Washington, D.C.: Association of American Medical Colleges, 1978.

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

Mehrens, W.A.; Lehmann, I.J. Measurement and Evaluation in Education and Psychology, New York: Holt, Rinehart and Winston, Inc., 1973.

Morehead, M.A.; Trussell, R.E.; Ehrlich, J. A Study of the Quality of Hospital Care Secured by a Sample of Teamster Families in New York City, New York: Columbia University Press, 1964.

Nie, N.H.; et al. Statistical Package for the Social Sciences, Second Edition, New York: McGraw-Hill Book Co., 1970.

Pawluk, W.; et al. "Concurrent Validity of the Canadian Certification Examination in Family Medicine," in Proceedings of the Fifteenth Annual Conference on Research in Medical Education, Washington, D.C.: Association of American Medical Colleges, 1976.

Payne, B.C.; Lyons, T.F. Method of Evaluation and Improving Personal Medical Care Quality, Office Care Study for Hawaii Medical Association, Ann Arbor, Michigan: University of Michigan School of Medicine, 1972.

Rhee, S. "Factors Determining the Quality of Physician Performance in Patient Care," 1975, A Dissertation at Governors State University.

Schumacher, C.F. "A Factor-Analytic Study of Various Criteria of Medical Student Accomplishment," Journal of Medical Education, Vol. 39, 1964, pp. 192-196.

Senior, J.R. Toward the Measurement of Competence in Medicine. Philadelphia, Pennsylvania: National Board of Medical Examiners, 1976.

Sibley, J.C. "Quality of Care Appraisal in Primary Care: A Quantitative Method," Annals of Internal Medicine, Vol. 83, 1975, pp. 46-50.

Standards for Educational and Psychological Tests, Washington, D.C.: American Psychological Association, Inc., 1974.

Tatsuoka, M.M. Multivariate Analysis: Techniques for Educational and Psychological Research, New York: John Wiley and Sons, Inc., 1971.

Trussell, R.E. The Quanity, Quality and Costs of Medical and Hospital Care Secured by a Sample of Teamster Families in the New York Area, New York: Columbia University Press, 1962.

Williamson, J.W. "Validation by Performance Measures," in Conference on Extending the Validity of Certification, Chicago, Illinois: The American Board of Medical Specialties, 1976.

Wilson, P.W. "A Comparison of Two Student Instructional Rating Forms," 1978, A Dissertation at Michigan State University.