This is to certify that the

dissertation entitled

A Comparison of Basic Science Items and Clinically Relevant Items in Measuring Physician Competence

presented by

Douglas Barker

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Educational Psychology

Walter Hapkiewicz
Major professor

Date May 22, 1987

A COMPARISON OF BASIC SCIENCE ITEMS AND CLINICALLY RELEVANT ITEMS IN MEASURING PHYSICIAN COMPETENCE

by

Douglas Barker

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling,
Educational Psychology,
and Special Education

1987

ABSTRACT

A COMPARISON OF BASIC SCIENCE ITEMS
AND CLINICALLY RELEVANT ITEMS IN
MEASURING PHYSICIAN COMPETENCE

by

Douglas Barker

This study investigated how test items which deal with basic science material perform when included in a national certification examination being administered to candidates seeking specialty certification by the American Board of Emergency Medicine.

Eighty basic science items were distributed throughout three booklets of four alternative, single best answer, multiple choice test items. These items were for experimental purposes only and did not affect the candidate's pass/fail outcome.

A total of 616 candidates sat for the examination. Candidates qualified for certification by completing five continuous years of emergency medicine practice or by successfully completing an approved emergency medicine residency program.

Four primary test item scales were of interest. The Basic Science scale, a scale of high clinical relevance (HCR) items which had been determined to correlate highly with performance on oral stimlated patient encounters (the criterion), low clinical relevance (LCR) items which had no discernible statistical correlation with the criterion, and medium difficulty (MD) items which had mid-range levels of difficulty (p values). Any items that might have been redundant (e.g., MD and LCR) were not included in the analysis.

Although each of the four scales discriminated between the two subject groups, they did differ in relative discriminating power. The results showed

that HCR items did not discriminate between levels of competency as well as was expected. Using discriminant analysis, it was found that the MD scale was superior in making discriminations. The BS scale was second best and when combined with the MD scale created the most efficient collection of items. The LCR and HCR scale made no significant contribution to the MD/BS composite.

Although the BS items proved to be slightly less correlated with the criterion than were the LCR items, they nonetheless proved to be superior to even the HCR items. It was also found that there were no differential effects of the four BS subscales (biochemistry, microbiology, pharmacology, and physiology).

To Mother:

Who somehow managed to keep a generally troublesome family intact, always urged me to do better than I was doing, and never made conditional her love and support.

ACKNOWLEDGEMENTS

I have enjoyed the support and encouragement of many people while carrying out this research. Special thanks are due the members of my dissertation committee.

I am fortunate to have had the opportunity to study with my Chairman, the late Dr. Robert L. Ebel. From him I learned measurement and test theory, the qualities of an outstanding professor, and the essentials of scholarship. I regret not having completed this work during his lifetime. I think he would approve it.

Dr. Larry Alexander provided many valuable comments and suggestions, which made the dissertation much better than it otherwise would have been. He also gave me the opportunity to work with him on an extended project in which he taught me the principles and the applications of instructional design, workshop planning and production, and how to function effectively as an educational consultant. A man of exceptional competence, insight, and patience, he often reminded me, "Sometimes it's best to be tentative." I am continuing to realize and appreciate what he was telling me.

Dr. Robert C. Craig, a first-rate scholar, writer, and thinker also chaired what was one of the outstanding educational psychology departments. His thoughtful reading of the dissertation and the suggestions he offered enabled me to make this report more readable, clearer, and better explained. His feedback also forced me to rethink what I had done and enabled me to see the subtleties, contradictions, and implications of the research. Probably the wisest decision I made during my graduate student years was asking Dr. Craig to serve on my committee. As a resource, critic, and mentor he is truly outstanding. I have had the opportunity to teach both on and off-campus

educational psychology courses. Special thanks, are due Dr. Craig for giving me the opportunity to acquire this valuable teaching experience. I am grateful for his sponsorship and support throughout the past years.

Dr. Walter Hapkiewicz occupies a special place among the many others who have helped and encouraged me. When admitted to graduate school, I was told he would be my advisor. He has since served as my academic adviser and doctoral program chairman. But those are his official titles and capabilities. More importantly, he has been a counselor, advocate, confidant, and friend. He has never wavered in his support of me; his encouragement has been equally relentless. The special opportunities, benefits, and awards I have received are a result of his efforts to represent and successfully lobby for me as one of his advisees. In addition, Dr. Hapkiewicz throughout our relationship has helped me avoid problems or entanglements with university and college rules and procedures. My graduate student life has been much easier because his vigilance and efforts ensured that I complied with all institutional requirements or, more often, found an exception to them. I am humbled and thankful that somehow fate assigned me as his advisee. The intervening years have been rich and rewarding as a result of that assignment.

My relationship with Dr. Jack L. Maatsch has benefitted me professionally, personally, and financially. Jack gave me the opportunity to work with him on the ABEM project. Being a part of that exciting and pioneering effort was something a graduate student rarely experiences. The treatment given me by Dr. Maatsch was exceptional. Besides being relatively well paid, having perquisites of office space, secretarial and computer

programming support I was expected, in turn, to function as a contributing member of a specially assembled and cohesive research team. I feel quite fortunate to have been associated with Jack's monumental test development and validation successes.

I have learned from Jack the essentials of medical and professional education, the power of social and interpersonal influences in academic decision-making, and how to avoid or resolve conflicts in an institutional setting. Jack also taught me the relative strengths of simulations, a teaching and testing format with which his name has become synonymous. I have grown in many ways under Dr. Maatsch's tutelage and appreciate all that he has done for me.

I am convinced that my committee was unusual since each of them was always accessible to me. Though they were full professors, recognized experts in their specialties, dealing with never-ending institutional obligations, and, in the cases of Dr. Craig and Dr. Maatsch, had chairmanship responsibilities, they never denied me access. They were always available and no matter how inane the reason for my intrusion they did not plead lack of time or a full calendar.

To the American Board of Emergency Medicine, it's President John Wiegenstein, M.D., and Executive Director Benson Munger, Ph.D., I owe thanks for their allowing me to collaborate with them in an effort to learn more about credentialling and test development techniques and strategies. Thanks are also due Ronald L. Krome, M.D. of ABEM who raised the issue of testing for basic science knowledge in the certification examination and advocated the question be dealt with experimentally.

Others have, of course, helped me in various ways during the years leading up to completion of the dissertation. I wish to thank Alan D. Neiberg, Ph.D., M.D., and his family who contributed substantially to me during my

graduate student years. Alan, Vicki, Maury Jo, and Forrest collectively and in their own way contributed to my personal growth and made possible the completion of this report. I thank them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Hardly any attention was paid to the issue of physician licensing and certification prior to the twentieth century; the topic was of only minor interest to aspiring physicians and those directly involved in physician training and education. Throughout the 1900's, and particularly during the past decade, interest in physician credentialling has increased substantially among political leaders, consumers, and affected health care providers. Quite likely, this increased interest is associated with this nation's intensified preoccupation with quality of life and health care delivery issues.

Responsibility for licensing physicians rests with the individual states, each of which has a board of medical examiners which determines the criteria for licensure in that state. States generally require graduation from an approved medical school, an internship or some type of supervised experience, and satisfactory performance on an examination. Although particular standards vary among the states, such variability has been reduced considerably during the past few years. In this connection the Federation Licensing Examination (FLEX) prepared by the National Board of Medical Examiners has become the norm. By 1976 "all fifty states accepted FLEX (with a standard passing score of 75) as an acceptable demonstration of competence for licensure" (Hubbard, 1978, p. 106).

Whereas an individual is licensed to practice medicine by the state, he is certified or "boarded" as a specialist by the respective specialty board. Each specialty adopts its own national standards for certifying individuals that are independent of the licensing standards of the states. The Board of

Ophthalmology, organized in 1916, was the first such board. Since that time more than twenty-three such boards have been created.

The most recently formed specialty board is the American Board of Emergency Medicine. ABEM was approved by the American Medical Association and the American Board of Medical Specialties in the fall of 1979. As part of the approval process ABEM selected the Office of Medical Education Research and Development, Michigan State University, to assist them in developing and evaluating an examination which the Board could use to certify physicians in the specialty of Emergency Medicine. Test development began in 1975. The examination was field tested in October, 1977, and was formally administered for the first time during the spring of 1980.

The purpose of this dissertation is to analyze selected results from this first official administration to determine how different item types and content may affect the reliability, validity, and discrimination power of the testing instrument. This involves testing the reliability of the field test results (Downing, 1979; Maatsch, et al., 1979) on a much larger sample. By studying the characteristics and the effects of selected content domains and subscales that were inserted into the Part I examination materials for experimental purposes, this study goes beyond those earlier findings.

NEED FOR THE STUDY

Classical test theory (Gulliksen, 1950; Lord and Novick, 1968; Magnusson, 1967) lends itself quite well to classroom achievement testing. Indeed, Ebel has authored many papers and several editions of one major text (1979) showing these practical applications. It should be observed, however,

that such achievement tests are norm-referenced and attempt only to discriminate between individuals or groups in terms of levels of knowledge or skills. Test developers and users of these instruments are primarily concerned with the reliability of the instruments. A good achievement test is one that sharply and reliably discriminates between the examinees. As a consequence, criterion-related validity (e.g., predictive validity) is typically ignored. In the few instances in which validity is explicitly considered, content validity is the primary focus. For content validity the question is whether experts agree on the basis of inspection that the test measures the knowledge it purports to. Since tests are typically prepared by teachers who assigned the subject matter or the publishers who published and distributed the subject matter, the question is inevitably answered in the affirmative. Rarely, if ever, are such tests shown to be predictive of subsequent performance in other settings. The predictive validity of most classroom achievement tests is unknown and unstudied.

Because certifying and licensing examinations form an important part of the basis for decisions which affect health practitioners and the public, predictive validity is paramount.

Public interest groups, health delivery personnel, consumers, and insurance underwriters are becoming more and more vociferous in their demands for test predictive validity. Claims of respectable content validity are doing little to silence their collective voices. Although classical test theory possesses all of the elements which could be used to improve certification and licensure examinations, greater emphasis and use will need to be made of criterion-related validity.

RATIONALE

In many ways physician certification examinations have evolved to closely correspond with the best contemporary testing practices. The internal consistency estimates of reliability are typically in the low nineties and experts agree about how to construct and score examinations. There is little room for improvement in terms of these reliabilities and content validation procedures.

Nonetheless, these examinations have come under fire since examination scores "seem to have very little relationship to quality of subsequent professional performance" (Williamson, 1976, p. 25). In general, those few predictive validity studies which have been reported have failed to show any significant relationship between test performance and independent measures of criteria of clinical competence. One notable exception has been the pioneering work of the American Board of Emergency Medicine. Using Chart Stimulated Recall performance as a criterion, the Board has designed procedures which have established the criterion-related (predictive) validity of their Certification Examination (Maatsch, et al., 1983).

As noteworthy as the ABEM validity study is, it is, regrettably, an exception. The general absence of documented relationships between a credentialling examination and any credible criteria of clinical competence may be attributable to any of three reasons. First, any deficiencies in the reliability or validity of such performance criterion measures would tend to depress the validity coefficients. Oftentimes the criterion is determined by using peer review ratings or chart audit as a proxy for clinical performance. Second, the validity coefficients may be reduced by a truncated range of scores on the examinations, the criterion measures, or both. Finally,

examination performance may be only marginally related to clinical performance and low validity coefficients may not simply be the result of technical measurement and sampling. Clinical competence is comprised of a number of elements of which test performance measures only one of the interacting components. Although the examinations reliably measure content-valid material, this material may be only marginally related to the quality of health care provided in clinical situations.

Downing (1979) showed that multiple-choice items with high clinical relevance tend to more sharply discriminate among candidates than more difficult content items that maximize dispersion. In an analysis of the field test results he assessed the power of selected item types to discriminate among four subject groups known to differ in their levels of clinical competency. These groups were: fourth year medical students; second year residents; practitioners with five continuous years of practicing emergency medicine (practice-eligible); and physicians who had completed an approved emergency medicine residency program and at least one year of practice (residency-eligible). The high clinical relevance scale discriminated among three of these groups better than did the medium difficulty scale, although the difference was not statistically significant, and the medium difficulty scale discriminated more sharply among these same three subject groups than did the low clinical relevance scale ($p < .05$). The internal consistency estimates of reliability of those three scales were ordered from high to low clinical relevance.

This dissertation extends Downing's efforts. It systematically examines the relationships among test items' relevance, validity, and reliability on the one hand and their ability to validly measure clinical competence on the other. The results of this study should suggest how tests can be constructed to

maximize relevance to clinical competence. The findings should facilitate maintaining the considerable reliability certification examinations presently possess while further enhancing the predictive validity of such instruments. Improving predictive validity of credentialling examinations represents the paramount challenge facing the medical specialty boards: insuring the credibility of credentialling procedures (Webster, 1976).

NATURE OF STUDY

The Emergency Medicine Certification Examination will be more fully described in Chapter III. Of direct relevance to this study is that portion of Part I of the examination consisting of 377 multiple-choice items that include the following scales and subscales:

High Clinical Relevance Scale

This scale is comprised of items that were shown in the field test to correlate highly with mean performance ratings on 12 Simulated Clinical Encounters (the criterion). The correlations ranged from .33 to .68 with .38 being the median. These items had a mean p-value of .721, where p equals the proportion of examinees responding correctly to an item.

Low Clinical Relevance Scale

Items that correlated least with the criterion make up this scale. The correlations ranged from -.23 to .11 with the median being .05. The mean p-value of the items comprising this scale was .667.

Medium Difficulty Scale

This scale was created by selecting those items which had positive discrimination indices and p-values between .50 and .70. The mean p-value was .639.

Basic Science Scale

Eighty items randomly selected to sample basic science knowledge without regard to clinical relevance for emergency medicine were included in this scale. These items were inserted for experimental purposes only and did not affect the candidates' pass/fail outcome. This basic science scale is, in turn, made up of the following subscales: anatomy (k=3), biochemistry (k=22), histology (k=6), microbiology (k=13), pharmacology (k=15), and physiology (k=21). Unlike the three scales above, items comprising the Basic Science Scale were not used at the time of Downing's study (1979), but were included in the February, 1980, examination solely for the purpose of this study. Candidates had no prior knowledge that they would be tested on basic science knowledge. They expected all items to be clinically relevant to the practice of emergency medicine.

RESEARCH HYPOTHESES

The hypotheses in this dissertation are designed to determine if Downing's findings generalize to other groups of subjects and to extend his findings by introducing a scale of unknown utility in a certification examination. Whereas, Downing examined discriminating power between four subject groups, this dissertation deals with two groups that sat for the

examination by virtue of their eligibility for certification. Here, as in Downing's study, one group met the Board's requirements for certification since they had completed an approved emergency medicine residency program (residency). The other group (practice) consists of examinees who have not completed an approved emergency medicine residency, but instead, have had five years of experience in Emergency Medicine. There is some reason to think that this latter group will perform about the same as second year residents (Downing, 1979). That is, in the field test groups analyzed by Downing, the residents and the practice-eligible groups appear to be more like each other in levels of test performance than either group was to the residency trained physicians who scored significantly higher on the test.

The following hypotheses were tested to confirm the findings of Downing on a different group of subjects (actual candidates) in a different setting (an actual certification examination setting):

1. The high clinical relevance scale, the low clinical relevance scale, the medium difficulty scale, and the basic science scale will each discriminate between the two physician groups.
2. The high clinical relevance scale will more sharply discriminate between the two groups than will any of the other three scales.
3. Using a measure of internal consistency as an estimate of reliability, there will be a significant difference among the reliabilities of the four scales.
4. The correlation between the high clinical relevance scale and the medium difficulty scale will be greater than the correlation between any other two scales.

A second set of hypotheses dealt with the basic science scale and its subscales.

5. The basic science scale will be more difficult than any of the other scales.
6. Each of the four basic science subscales will discriminate between the two physician groups.
7. The correlation between the median basic science item and the criterion (Part II simulations) will differ from what Downing found the correlation to be between the median low clinical relevance item and the criterion.
8. The median item correlations between each of the four basic science subscales and Part II of the examination will not be zero.

OVERVIEW OF THE DISSERTATION

The relevant literature dealing with the primary nature of basic science in the study of medicine and the relationship between basic science in the study of medicine and the relationship between basic science and the practice of medicine is reviewed in Chapter II.

Chapter III describes the research procedures and methodology, including the design of the data collection instrument, subject characteristics, sampling method, and the data analysis methods.

Results of the data analysis are presented in Chapter IV.

The conclusions, implications, and suggestions for further research are discussed in Chapter V.

CHAPTER II

REVIEW OF THE LITERATURE

Chapter I introduced the problem and stated the rationale for carrying out the research of this dissertation. This chapter reviews the relevant literature with respect to three issues: (1) origins of the current practice of emphasizing basic science in the study of medicine, (2) Flexner's influence on this practice, and (3) research that has attempted to determine the role of basic science knowledge in the practice of medicine. The chapter concludes with a discussion of how measurement specialists have used and studied the concept of relevance.

THE ORIGIN OF BASIC SCIENCE IN AMERICAN MEDICAL SCHOOLS

Early in this country's history colleges were under the direct control of clergymen, and students undertook college study primarily to prepare for the ministry. The curriculum was dominated by the classics and was so rigid that the student had no opportunity to elect courses. No particular sciences---social, natural, or physical---were taught. Science was considered "natural philosophy"---a much less important part of the curriculum than moral philosophy. Such "science" was limited to a descriptive classifications of plants and animals (Harrington, 1905). James Garfield, twentieth president of the United States, described the typical curriculum of the mid-nineteenth century as follows: "In the whole program of study, lectures included, no mention whatever is made of physical geography, or anatomy, physiology, or the general history of the United States" (Nevins, 1962).

It was not until after the Civil War that college curricula became more flexible and broader based. Science began to show some promise of practical applications, which soon attracted the wealth needed for support and integration into universities (Shryock, 1956). The eventual infusion of science into college curricula enabled Bordley and Harvey (1976) to argue that the 1780 Massachusetts statute that changed Harvard from a college to a university was a misnomer because Harvard consisted at that time only of a school of Arts with three professorial chairs: Divinity, Mathematics, and Oriental Languages.

Medical schools developed under the conditions that prevailed more generally in American higher education. The first American medical school was established at the College of Philadelphia in 1765. Prior to that, one typically became a physician by studying medicine in Europe or serving as an apprentice with a practitioner who, himself, likely had no formal education in medicine either. Neither of these two forms of training were essential, however, since there were no licensing or credentialling requirements. In the many geographic areas where there were no physicians, medicine was practiced part-time by local clergyman or educated layman to supplement their income. Although four thousand individuals were estimated to have practiced medicine prior to the Revolution, only an estimated ten percent possessed medical degrees (Packard, 1932).

Many of the United States' early settlers were well-educated Scots, whose intellectual activities included: maintaining ties with intellectual leaders in their native country; establishing reading and discussion groups; and forming academies designed to prepare young men for a variety of careers including medicine. In 1746, these Scottish settlers created the College of New Jersey, which moved to Princeton ten years later and assumed the name

of that city. Medical degrees were conferred by various European schools and the University of Edinburgh attracted a substantial number of colonialists. Between 1765 and 1779, more than three hundred Americans studied medicine at Edinburgh, 112 of whom received M.D. degrees. Indeed, Americans were overrepresented among the student body, since five of the school's thirteen graduates in the class of 1765 were Americans (Packard, 1932).

Bowers (1977) suggests the factors which encouraged Americans to study in Edinburgh were the strength of the Presbyterian church in the colonies and the importance it placed on rigorous and liberal education, the Scot/Presbyterian academies, and the College of New Jersey. Moreover, Americans were impressed that the Scottish universities were beginning to liberalize their curricula without either severing ties with the church or otherwise committing heresy (Ibid.). In addition to these general conditions, more specific attractions of studying medicine at Edinburgh were the prominence of the Faculty of Medicine and the opportunity to earn an M.D. degree from the first complete medical school in the English speaking world (Bowers, 1977). Also, unlike the continental schools, some of the teaching was being done in English, with Latin being reserved for the examinations and the dissertation (Girdwood, 1977).

Tounis College, which later became the University of Edinburgh, was created in 1582. The first professor of medicine was hired in 1685, with others being added periodically over the next forty years. The University began conferring the M.D. degree in 1705; however, it was not yet a self-contained medical school because the Royal College of Physicians was responsible for examining and approving candidates prior to their receipt of the degree (Ibid.). In 1726, the Faculty of Medicine was established and the University thereby

became the first complete medical school---responsible for admitting, teaching and training, examining, and graduating students.

Just as many Americans had received the M.D. degree from Edinburgh, many of the founders, leaders and faculty of that school had earned their medical degree from the University of Leyden which was founded in 1575 and became a major medical center during the 1600's. During the period 1575 to 1875, 546 Scottish students received their medical training at Leyden. In addition to a most distinguished faculty, Leyden had several of the characteristics we see in today's American medical schools. For example, students were expected to gain practical experience in the affiliated teaching hospital and also to demonstrate mastery of then established scientific facts and principles to the satisfaction of the University faculty. Leyden's facilities for the study of science were outstanding for their time, and included an anatomy department, a chemistry laboratory, and a botanical garden. The source of this remarkably modern curriculum can be traced to the University of Padua medical school which was founded in 1212 and was the medical education center of the world for more than 300 years. Padua and all other Italian universities, however, declined in popularity during the Reformation, and their decline enabled Leyden to be founded and to flourish (Guthrie, 1959).

The Scots who created the University of Edinburgh Medical School, modeled it after Leyden, where they had studied medicine. Similarly, in many ways the early Americans recreated the University of Edinburgh when they established the first medical school in the colonies. As Guthrie (1959) observes, "The Leyden tradition was firmly transplanted to Edinburgh, and again, to Philadelphia."

Bowers (1976) contends, "The Faculty of Medicine of the University of Edinburgh is the mother of American medicine. To it we owe the first medical

schools in the colonies and the genesis of our medical profession." He contends further that the texts used in the early American medical schools were written by Edinburgh professors. Bordley and Harvey (1976) concur with Bowers and identify the University of Edinburgh as "the model" for the first American medical schools. Flexner, who later reviewed American medical education, appreciated the wisdom of the early colonists in adopting the Edinburgh model: "Our first medical school [College of Philadelphia, 1765] was thus soundly conceived as organically part of an institution of learning and intimately connected with a large public hospital" (Flexner, 1910).

The first medical schools in this country were ahead of their time in emphasizing science since historically, science, and particularly basic science, had less importance than is now the case. Several reasons have been advanced for this by Shryock (1948). Basic science was seen as having no practical value. For a young nation made up of doers, intent on conquering a continent, it was easier to borrow from abroad the science that was needed. Shryock, considered the leading authority on the history of American medicine (Curti, 1974), suggests that the church's fear of science---an additional source of resistance to science---was probably less of an impediment to its growth than is often believed. Religion was simply too diverse and factional to have been much of an influence, as evidenced by the fact that basic science quickly began to take root and grow when this country's developing class of industrialists and enterprising businessmen realized how science might serve their interests. Accordingly, the historical indifference to science began to change in the late 1800's, and this change coincides with and in part explains Flexner's finding (to be discussed in the next section) that little science was being taught to students of medicine during the first decade of the twentieth century (Flexner, 1910).

FLEXNER AND HIS REPORT

The Flexner Report (1910) is generally considered to be the seminal work that shaped the modern-day medical school, prescribing both its subject matter and organization. The purpose in reviewing the work of Flexner and his critics is to better understand how science came to occupy such a dominant position in the medical school curriculum. The Report, which was funded by the Carnegie Foundation, followed Flexner's visits to and reviews of each of 155 American and Canadian medical schools. These visits led Flexner to the conclusion that the quality of medical education was atrocious. In addition, many schools were guided by their quest for profit and few were equipped, staffed, or disposed to offer appropriate training.

Most notable among Flexner's recommendations were:

1. Prerequisite training, including science study, should be required of all incoming students.
2. There should be extensive study of science, primarily during the first two years of medical school.
3. Commercial, proprietary, and profiteering schools should be eliminated.
4. There should be minimum requirements for a school's physical facilities.
5. Medical schools should be part of a university and also should have their own teaching hospital.
6. Licensing authorities should become more active, particularly in pushing for rigorous standards.

The Report graphically exposed each school's deficiencies, and was widely hailed. One reviewer of the history of medical education concluded, "It is hard to over-emphasize the importance of this report" (Dobbs, 1957, p. 788). In a New York Times front page announcement of his death, the Report is

labeled a "bombshell" which "revolutionized medical studies in the United States" (Abraham Flexner). The requirements in contemporary medical schools of two years of basic science study and two years of clinical training have come to be called "the Flexner curriculum" (Chapman, 1974, p. 111) and Flexner has affectionately been called, "the uncle of modern American medicine" (Whipple, 1960, p. 451). While most admirers of Flexner have lauded the Report and what they see to be its desirable effects on Western medical education, one commentator argues that the work of Flexner is not yet finished and medicine must re-energize itself to become even more scientific (Engle, 1978).

Notwithstanding these many extolments, the case can now be made that there is no convincing evidence that the Flexner Report was the impetus for correcting the deficiencies he perceived. Instead, Flexner was riding a wave which had already crested.

Prior to the entry of Flexner or the Carnegie Foundation into medical education, schools were already stiffening their admission standards and graduation requirements, state licensing boards were refusing to license graduates of some schools, and schools were closing at a phenomenal rate. Bevan (1928) reported that nearly twenty-five percent of all medical schools went out of operation before Flexner. Berliner (1977) noted that prior to Flexner, "the reform of medical education was proceeding rapidly and that there was a good deal of self-reform within the profession" (p. 604). He then concluded, "the Flexner report has received attention far out of proportion to its actual contribution to medical education..." (p. 608).

Before Flexner there existed a variety of medical associations and sects, each trying to dominate the others. One such group was the American Medical Association, which though fifty-three years old in 1900, represented only 7%

of this country's healers (Berliner, 1975, p. 581). The American Medical Association, for what may have been self-serving motives, urged the Carnegie Foundation to study the field of medical education and assisted the foundation in the formulation of an evaluation plan. Several writers have voiced suspicion about Flexner's objectivity and independence. Floden (1980) has gone so far as to say, "Flexner's study was part of an effort by the AMA to reduce the number of medical schools" (p. 36).

The reduction in the number of all of the various types of medical schools began in 1904 and was largely due to the AMA's press for reforms (Markowitz and Rosner, 1973, 96). Release of the Flexner Report, however, accelerated this trend. By 1912, two years after publication of the Report, there were but ten homeopathic medical colleges and six eclectic colleges. These numbers are equivalent to the number of such schools that existed in 1880, the earliest year for which such figures are available. Physio-medical and nondescript (not regular, homeopathic, eclectic, or physio-medical) medical colleges no longer existed. Although the number of regular medical colleges had also diminished, these colleges by 1912 represented an all-time high 86.2% of the various kinds of medical schools (Medical Education, 651). Those few sectarians who survived "were usually unable to win access to hospitals or the right to prescribe drugs" (Starr, 1982, p. 127). These non-allopathic practitioners were not simply brushed aside for the moment but were permanently eliminated:

> According to a survey of nine thousand families carried out over the years 1928 to 1931, all the non-M.D. practitioners combined---osteopaths, chiropractors, christian scientists and other faith healers, midwives, and chiropodists---took care of only 5.1% of all attended cases of illness. Physicians finally had medical practice pretty much to themselves. (Ibid., emphasis added.)

An often overlooked consequence of school closings and reduced enrollment is the effect on minority persons. For example, it was much easier before publication of the Report for a woman to get into medical school and later become a physician than it was after the Report appeared. Although the number of women in all positions of health care increased after its appearance, the proportion of those who were physicians steadily declined absolutely and in comparison to male physicians (Shryock, 1950). Jews also were hard hit and continued to be disadvantaged by the imposition of quotas by the medical schools and the hospitals (Goldberg, 1939). One writer suggests that when medical school admission discrimination against Jews finally began to subside it was through no demonstrable effort of "the medical profession or its official leaders" (Jarcho, 1959, p. 371). Even more affected were blacks, who saw five of the seven negro medical schools closed while the likelihood of their being admitted to any of the surviving schools plummeted (Kessel, 1970, p. 270). At the same time, internships and residencies were being developed, entry to which was even more difficult for a black than getting admitted to medical school (Adams, 1937).

Flexner's own attitude toward the medical education of blacks in particular was, at best, patronizing. He gave two selfish and racist reasons why medical education should be provided to a few negroes who can then provide health care to the other blacks. First, disease can then be contained which otherwise might spread to white people. Also, white physician income wouldn't be effected by the presence of a few black physicians since "the practice of the negro doctor will be limited to his own race" (Flexner, 1910, p. 180). Quite likely the detrimental effect on minorities would have been much greater if Flexner would have achieved his goal of reducing the existing 155 medical schools to a mere 31 (Flexner, 1910, p. 154). The number dropped

only to 76, and that occurred nineteen years after Flexner established his objective (Jarcho, 1939, p. 356).

Writers of late are criticizing not only the substantive weaknesses in the Report, but also its methodological shortcomings. King (1984) notes, as did Kessel (1970), that Flexner was hardly a rigorous evaluator since he "raced through the inspections at a great rate" and claimed to be able to conduct an exhaustive review of a school's standards in "half an hour or less" (p. 1085). King considers the Report "an achievement in public relations and not an intrinsic contribution to medical education" (p. 1084).

Flexner's employment by and personal relationships with wealthy benefactors are cited by critics to show the social bias of his work. Berliner (1976), in a classic Marxist analysis, sees the Flexner effort as a class-inspired strategy to gain control over the health care system. Berliner notes that during the twenty years following the Report, the nation's nine largest foundations gave more than 154 million dollars to reform medical education along Flexnerian lines (Ibid., pp. 589, 590). Markowitz and Rosner (1973), in a somewhat less partisan analysis, seem to concur with Berliner's conclusion, "It [the Report] also helped to further consolidate the power and influence of the Eastern university-based elite" (Ibid., p. 101). King (1984) also cites Flexner's "strong elitist bias" (p. 1084) which "erects as its ideal a concern with knowledge, research, and intellectual training" (p. 1079). Flexner's invocation of science had the effect of distinguishing between the worthy and the unworthy practitioners and was consistent with this ideological outlook. Schudson (1974) contends that "Flexner's position is unashamedly elitist," citing Flexner's asumption that knowledge, while intrinsically abstract and so difficult it can be learned by only a few, it is nonetheless certain, universal, and classless in its applications (p. 359). Perhaps Flexner's idealism is best

captured in his own later paper titled, "The usefulness of useless knowledge" (1939).

Whereas students of Flexner disagree as to the profundity or heuristic import of his Report, they seem to agree that he brought about reforms in medical education by "generating a flow of money into many of the nation's medical schools" (Chapman, 1974, p. 11). The general success with which Flexner was able to persuade Brookings, Eastman, Rockefeller, and other rich men to support medical education has been widely noted (Banta, 1971; Chapman, 1974; Fox, 1980; Markowitz and Rosner, 1973). In addition to Whipple's remarks (1960) about Flexner's influence on Eastman to fund the development of the University of Rochester, case studies have been reported of Flexner's philanthropic genius regarding aid to Washington University (Munger, 1968) and the University of Cincinnati (Cangi, 1982).

Although Flexner's success in raising money for selected medical schools diminished long ago, his influence on the medical school curriculum continues today. The most evident effect of Flexner is the dominance of basic science in the organization of today's medical schools. To say that Flexner's views were ideological and socially-based or that he intended to make medicine a respectable field of study and practice is to miss the main point: that his Report neither established the importance of basic science per se nor the link between it and the practice of medicine. Seen another way, Flexner only told us how to educate an aspiring physician, he told us nothing about what the doctor should be able to do or what competencies should be prerequisites to practice. He contributed little to our understanding of the relationship between science and patient care.

RESEARCH ON BASIC SCIENCE IN MEDICAL TRAINING AND PRACTICE

The literature suggests that the relationship between knowledge of basic science material and health care is of greater interest to medical educators than to practitioners or regulators of medical practices. For example, a line of research has examined the amount of science studied by applicants prior to their medical school matriculation to see if that variable accounted, in any part, for medical school performance.

One of the first such studies was done by Shultz (1951), who studied the medical college admission test (MCAT), an examination prepared and administered by Educational Testing Service which continues to be required of applicants to medical school. The science section of the examination sampled materials from the basic college courses in biology, chemistry, and physics. Shultz investigated whether advanced or additional study in these disciplines would improve scores on those portions of the exam, and found that advanced study had no effect on test performance. As Shultz put it, "the results of this study would appear to offer no support for the hypothesis that taking additional courses in biology, chemistry, and physics beyond a certain minimal number leads to better scores on the MCAT test..." (p. 147). He then reasoned that the fundamental principles are learned in the introductory course and that knowledge may or may not be broadened or enhanced in advanced course work. Even if it is so enhanced, the MCAT, which measures only rudimentary knowledge, is apparently insensitive to more advanced knowledge.

More recently, Gough (1978) studied the MCAT's science scale in relation to the quality of the examinees' premedical science study rather than the extent of such study, as Shultz had done. Treating premedical grade-point

average (GPA) in science courses as a factor, Gough found it correlated only modestly (.21) with performance on the MCAT science scale. He also noted that neither the premedical science GPA nor the MCAT science score seemed to be related to performance during the fourth year of medical school or to faculty ratings of clinical competence. Gough used his findings to appeal for broadened admission criteria, arguing science majors applying to medical school should not be given routine preference over students of humanities. His appeal is an interesting one since a later study showed that humanities majors generally have lower premedical grades than natural science majors, and have similar MCAT scores, yet as a group are accepted to medical schools at a proportionately higher rate than their many, many more natural science counterparts (Thomae-Forgues and Erdmann, 1980). The proportion of humanities major applicants who are accepted into medical school is somewhat higher than the proportion of science applicants who are accepted. The pool of humanities majors applicants is quite small when compared to the number of science majors seeking admission. Since this study dealt only with these two groups, who together account for some 70% of medical school applicants, the relative standing of social science majors is unknown.

Dickman, et al., (1980) echoed Gough's (1978) pleas for reducing the historic antagonism admissions officers have shown toward nonscience students. They studied three classes at SUNY at Buffalo medical school and found no differences between science and nonscience majors on medical school performance measures. Neither did they find any difference in choice of residency between the two subject groups. This latter finding is inconsistent with that of Zeleznik, et al., (1983) who found that those with undergraduate nonscience degrees were more likely to enter a psychiatry residency program than were science majors who were disproportionately represented in surgery

programs. Otherwise, Zeleznik, et al., found no significant differences among the various undergraduate majors on yearly medical school grade-point averages or on any of the three parts of the National Boards.

Yens and Stimmel (1982) probably collected more data than any others who have reported on the issue of academic preparation for medical school. Their subjects were the 735 persons who comprised the nine classes admitted to Mount Sinai School of Medicine from 1972 through 1980. The authors partitioned the subjects into three groups based on undergraduate major (traditional science, social science, humanities) and then analyzed for differences among these groups on the following variables: undergraduate grade point, MCAT scores, Part I and Part II scores on the National Board of Medical Examiners Examination, medical school grades, and membership in a medical honor society. On the MCAT subscale verbal and general information, humanities, social science, and traditional science ranked 1, 2, and 3, respectively. On the MCAT science scale the ranking was traditional science, humanities, and social science. On National Board Part I Behavioral Science Score and Part II Psychiatry Score, humanities and social science group scores were superior to the traditional science members scores.

In each of the five analyses the authors fail to follow up on significant F scores by performing and reporting the result of post-hoc tests. The reader cannot determine which particular groups differ from one another. Nonetheless, their results apear to be consistent with those of Thomae-Forgues and Erdmann (1980) who also found that among those applicants admitted to medical school, traditional science undergraduate majors do no better than other majors, the lone exception being a slight increment in the MCAT science score. Science majors perform better on the science scale than

do other students who, in turn, outperform the science students on the verbal and general information scales.

Yens and Stimmel concluded that their results "suggest that the nature of the undergraduate major makes little difference in the academic performance of medical students" (p. 434). They go on to deny that the Mount Sinai curriculum is unusually hospitable to those with nonscience backgrounds. Instead, they note that the Mount Sinai curriculum does not appreciably differ from that of other medical schools, all of which have "adapted their curricula to educate a heterogeneous student body" (pp. 434-435). Given this, the nonscience major is at no disadvantage.

The research cited suggests that the nature of premedical education is independent of subsequent performance during medical school. Incoming medical students are expected to have an understanding of the rudiments of the essential sciences. More detailed premedical study is not required, nor does it give one any apparent advantage. Medical school has a homogenizing effect on the student body: those with but little study of science have no handicap; those having extensive study, no edge. The implication of these studies is that admission committees should show no preference toward particular undergraduate major fields of study. If this were done, it would not only eliminate major-based discrimination in the admission office, but might result in more diverse student bodies. Indeed, Zeleznik, et al., go so far as to say that non-preference may create physicians who can better diagnose and treat the nonbiological aspects of disease (p. 33).

Nearly all medical schools in the western world require their students to formally study the biological sciences. Normally the first two years of medical school are devoted to such study. The courses are typically provided within the framework of discrete academic departments and are taught by

discipline-based personnel. Students are expected to master the material since such mastery is a prerequisite to third and fourth year study which involves patients. Also, licensing examinations, for which students will soon sit, test for basic science knowledge.

Most research dealing with the organization, methods, and technology of teaching basic science has been conducted in a medical school context. For example, computers have been hailed as a basic science teaching tool by Essex and Sorlie (1979) who also found students performed better on test questions written by their own teachers than they did on test items written by teachers who taught at a different site (1982). Researchers from the University of Washington School of Medicine have reported there are no differences among first year students who have studied the sciences at one of the several sites in a four-state area (Cullen, et al., 1976, 1977, 1981). After comparing the students on such performance variables as final examinations in the science courses, subsequent course work, and Part I of the National Boards, Cullen, et al., stated that "decentralizing the first year of the basic science curriculum...does not place students at an academic or attitudinal disadvantage" (1981, pp. 415-416).

The amount of research which has been reported on the medical school curriculum is voluminous. Sorlie, et al., (1972, 1973) reported that they successfully compressed the normal two-year study of science into one year. At the same time, the four-year curricula was being reduced to three years (Garrard and Weber, 1974). On the other hand, Comroe, et al., (1951) demonstrate how the basic sciences should no longer be taught as separate courses, but be integrated with one another so as to better relate to clinical medicine. Other studies of basic science include Guyer, et al., (1974), who found that laboratory time was being reduced while lecture time in the

sciences remained constant, and Levy, et al., (1972), who showed how student feedback could be used to modify the curricula. To be sure, voices calling for curricular change continue to be heard.* On the other hand, the argument has also been advanced that the organization of the basic science curricula is just fine and should be left unchanged (Kendall, 1960).

To summarize, the work that has been done on basic science falls into two clusters. One cluster deals with premedical school science. Here questions such as the following are addressed: How much science should be studied prior to medical school? How well do students do who have studied more compare to those who have studied less? What are the prejudices of medical school admissions officers regarding applicants' extent of science study? How can the admission practices be changed to conform with a different view regarding premedical science? In some respects the research of this cluster has been especially illuminating. There is a level of consensus not often seen in social science research which shows that the nature of the premedical education is immaterial to success in medical school. Nonetheless, most medical students admit they majored in a science as an undergraduate because they thought that was the most likely way to get into medical school (Pellegrino, 1980). These students are probably acting wisely by disregarding the research and taking advantage of the continuing preference shown science majors by admissions officers (Dickman, et al., 1980).

The second cluster deals not with the premedical science but rather with the science to be studied during medical school, typically during the first two years. Questions include: What sciences should be taught? By whom? Using

---

* See the nine letters in New England Journal of Medicine, 1983, 308, 1230-1232.

what method? What materials? What proportion of the total curriculum should be devoted to each of the science disciplines? As the review above suggests, the questions posed have been as wide-ranging as have the research findings. There is very little consensus. These practical problems tend to be expediently resolved by blending the past, established practices with local needs and preferences in such a way that the result is practical and practicable.

Each of these two clusters is closely related to phases of education. One focuses on the premedical school period, the other, the time during which sciences are studied in medical school. It is difficult, therefore, to find a common reference point. Accordingly, this body of work says little about the relationship between basic science education and practicing physicians' competence because very little is known about how knowledge of science might interact with the quality of health care physicians provide.

One study found that practicing internists knew much less about principles of biostatistics and epidemiology than did faculty or teaching housestaff (Weiss and Samet, 1979). The authors then reviewed more than one thousand articles appearing in the respondents' specialty literature and found more than half the articles employed some biostatistical or epidemiological concept. They concluded that the level of sophistication with which the subjects were reading the literature was in doubt. The present study doesn't really determine how much basic science physicians know, but it may imply something about how well they are able to stay abreast and understand the evolving and ever-growing body of basic science literature.

More central to this dissertation, Machotka, et al., (1971) found that second and third year pediatric residents believed they had largely forgotten the basic science they studied five years earlier. Nonetheless, they scored

higher on course examinations of basic science knowledge than did medical school freshmen who had just completed their study of the material tested.

Kennedy, et al., (1981) report data that both confirm and conflict with the findings of Machotka, et al. (Ibid.). They found that a general decline in basic science knowledge occurs between the end of the second year and the end of the fourth year of medical school. They then downplay the decline, claiming it is statistically, not practically, significant. They note that the decline was variable among individual disciplines: knowledge loss was greatest in biochemistry, followed by anatomy, physiology, and microbiology. In the case of the behavioral sciences, pathology, and pharmacology, gains (not declines) were observed.

This general finding is consistent with the finding of Dubois, et al., (1969) that the amount of basic science knowledge retained was inversely related to the number of years which had elapsed since the second year of medical school. Moreover, they suggested that this finding is independent of the subject's activity in the interim. Most of their subjects went on to finish medical school and then undertake residency training after completing the second year of medical school. Since this advanced training didn't seem to reduce the diminution of science knowledge, the time was apparently not being spent learning science or reinforcing science earlier learned. The authors suggest that science knowledge is diminished by forgetting and also by the tendency of basic science to rapidly become out-dated.

The analysis of Dubois', et al., of the basic science subscale results are generally inconclusive. In one analysis of the National Board Part I physiology scale they found second year student candidates performed no better than first year student noncandidates. In another analysis (based on only seven residents from unspecified specialties) they found a general reduction in basic science

knowledge and hints of subscale interactions, "but the number of residents examined was too small to allow firm conclusions" (p. 1040).

Dubois, et al., are unabashed proponents of continuing education for medical practitioners. The primary point of their paper is to show that basic science knowledge fades after mastery but can be restored by a 156-hour lecture graduate course. To keep abreast of the basic sciences, the course should be completed "once every three years following completion of his second year of medical school and thereafter throughout his career" (pp. 1042-1043). Dubois, et al., then suggest two options which are equivalent to the course. One would require life-long home study; the other calls for spending "five to ten days away from practice once or twice a year" undergoing intensive study. In these ways, they conclude, keeping abreast of basic science can be done painlessly.

Unfortunately, the authors beg the question of the role of basic science in clinical practice. They assume that maintaining mastery level knowledge of basic science among physicians throughout their careers is a desirable objective. At no time, however, do they even suggest that the quality of health care provided is related to the amount of basic science known. Nor do they imply that patient care is improved by the physician who has completed the course and restored his science knowledge.

Instead, in a paragraph in their Discussion section they attempt to establish the connection between basic science knowledge and medical practice as follows:

> One can rightly ask what role basic sciences play in medical practice, and whether Part I of the National Board Examinations contains questions which a practicing physician should be able to answer. It was found that basic science questions prepared by the faculty who taught this Correlated Basic Science Course, for purposes of interim examinations of the students, resembled the questions asked on the National Board examinations. Thus, the basic science information

> which the graduate faculty felt graduate doctors needed to know resembled basic science material which medical students are taught, and on which they are examined as candidates for medical licensure (Ibid., p. 1042).

Their argument can be summarized as follows:

1. Practicing physicians should be able to answer the basic science questions which comprise Part I of the National Boards.
2. The faculty who taught the 156 lecture basic science graduate course prepared test questions for interim examinations in the course and those test questions "resembled the questions" asked on Part I.
3. The faculty writing these test items intended them to deal with information which the practitioners needed to know.
4. Since these items are related to practice, the basic science items which resemble them are also related to practice.

Thus the role of basic science in medical practice is established through some curious, perhaps tautological, reasoning, because the conclusion begs the question.

Although Dubois', et al., research is methodologically sound, it does not establish the intended conclusion. They set out to discover how much basic science knowledge the practicing physician has. Quite simply, the answer is, "not as much as he once did," or, more specifically, "a little less each year." Had the paper ended there, it would have served a purpose. Unfortunately, the authors went beyond the data by assuming knowing basic science is good for the practitioner and knowing more is even better. This proposition, in turn, then establishes the need for their course which optimizes this knowledge. Stretching their data even further in an effort to show the relevance of science and Part I to practice (as in the paragraph quoted above) detracts from this otherwise lucid paper.

The Dubois team (Ibid.) begin their paper by admitting they "could find no information concerning a physician's knowledge of basic sciences after his graduation from medical school" (p. 1035). That statement is nearly as true today as it was fifteen years ago. The present study was conducted to learn more about how much basic science physicians know and how that amount of knowledge is related to their competence to provide health care.

TESTING EXPERTS' STUDY OF RELEVANCE

The notion of relevance has been given little direct attention in the measurement literature. Perhaps the topical area of educational and psychological measurement has been disproportionately concerned with psychological tests rather than educational tests. Tests of an individual's psychic make-up are probably not intended to appear relevant; indeed, they might lose their usefulness were they to do so. For example, one's precision in counting dots does not seem immediately related to the probability of keeping an appointment (Buros, 1970), and it is not apparent that the degree of group loyalty is indicated by how quickly one responds to the word-association stimulus "green" (Ibid.). Undesirable personality traits or dispositional quirks would not likely be revealed by examinees responding to a test of obvious relevance.

Accordingly, relevance is probably of more interest to the educational psychologist who is concerned with intellectual achievement and occupational credentialling than it is to the clinical psychologist whose primary emphasis is with the examinee's psychological composition, disposition, and intention.

The few early educational measurement specialists who discussed relevance generally suggested that relevance and reliability were the two

components of validity. Cureton (1951) attempted to give the concept of relevance statistical elegance in his argument for five different types of "relevance." Remmers and Gage (1955) distinguished only two types of relevance: logical relevance, which could be demonstrated by expert subjective consensus; and empirical relevance, which was inferred from statistical predictive studies.

Neither of these advocates of relevance appear to have had any appreciable influence on testers. Their work is not widely cited and did not precipitate any published research efforts or commentary by their contemporaries.

Ebel, an educational pragmatist, argued that the test constructor, who from his point of view was typically a classroom teacher, could prepare better tests if the items were classified by their degree of relevancy (Ebel, 1953). These categories were content detail, non-functional, vocabulary, fact, generalization, understanding, and application. These categories reflected common teaching objectives and a hierarchy within which test items could be assigned. Ebel's hierarchy anticipated the well-known taxonomy of educational objectives which appeared three years later (Bloom, 1956).

If the utility of a classification scheme lies in the rate of agreement among its users, Ebel's relevance categories are apparently well conceived. A colleague of Ebel's reported that the two of them were able to assign items to categories with 70% to 80% agreement (Cook, 1960).* This compares favorably with the rate of consensus among persons assigning items to their appropriate level of Bloom's taxonomy (Stanley and Bolton, 1957).

---

* Although Cook's paper first appeared in 1959, the 1960 version is cited herein since it is more widely available to the interested reader. Both, however, are referenced in the attached Bibliography.

It would seem to follow that the practitioner who constructs his test by using the relevance categories as a guide probably constructs a better test, at least along this particular dimension, because he is forced to sample more broadly along this scale than he otherwise would be. But tests that widely sample among relevance categories are not more discriminating. In the one reported effort to study this issue, Cook (1960) assigned each of 943 items to their appropriate relevance category on Ebel's six-level hierarchy of relevance. He found that the items of low relevance were more discriminating than were the high relevance items. Perhaps from disbelief, Cook then collapsed the six category scale into two categories: fact items and interpretive items. The results were in the same direction, with fact items proving to be superior discriminators to their more relevant counterparts. Additionally, discrimination did not interact with the difficulty of the items where no real differences or pattern emerged.

The little interest the measurement experts showed in relevance during the 1950's quickly diminished. Writers of the 1960's and the 1970's neither discussed nor indexed the topic (Cronbach, 1970; Cronbach and Gleser, 1965; Guilford, 1967; Magnusson, 1967; Nunnally, 1967).

Downing (1979), however, recently resurrected the issue of relevance and used the concept in a novel way. Downing's predecessors used the notion of relevance to denote the relationship between test material and the subject matter earlier taught. Unlike those before him, Downing's notion of relevance dealt with the relationship between two contemporaneous phenomena, one of which emulates the real world. Specifically, he was interested in the utilitarian nature of the test material. Downing studied relevance in a statistical and controlled way; he operationally defined his high and low clinical relevance test items according to their statistical strength of

association with examiner ratings of physician performance on Simulated Patient Encounters, a higher fidelity test of clinical competence.

This statistical-semantic difference in the way researchers have treated relevance may account for the apparent difference between the results of Downing and Cook. This dissertation follows the lead of Downing and keeps intact the empirical assignment of relevance he made to the test items.

SUMMARY

The conclusions of the topical areas reviewed above can be stated rather simply:

1. Basic science material was incorporated into the study of medicine initially to enhance the integrity of the teaching institution and later to lend credibility to the emerging profession.
2. No evidence can be found---indeed, no studies have been undertaken---which show the relationship between physician's knowledge of basic science material and the quality of patient care delivered.
3. Background breadth and quality of basic science study is heavily weighted in the medical school admission process. Once in medical school, the student spends the first two years primarily studying basic science.
4. The concept of relevance has never attracted much attention from testing specialists.

CHAPTER III

PROCEDURES AND DESIGN

INTRODUCTION

This report is based on the results of the Part I multiple-choice items and is designed to test directly the generalizability of Downing's (1979) field test findings to unique subject groups. Moreover, this study attempts to determine the extent to which basic science content might serve as appropriate testing material when the purpose of testing is to discriminate among levels of clinical competence. The study is designed in such a way that comparative statements can be made between the basic science scale and the other scales of interest.

This chapter contains a discussion of the subjects, how they were sampled, their relationship to the greater population, and the criterion by which the two subject groups were created and individuals so assigned. The testing materials are described as are the relevant scales and subscales which were an integral part of the test battery. The data gathering procedures are discussed, the hypotheses of interest are stated in testable form, and the statistical procedures used to test these hypotheses are described. The results are presented in the following chapter.

SUBJECTS

National specialty boards are solely responsible for certifying candidates within respective specialty areas. Physicians involved in this study applied to

the American Board of Emergency Medicine (ABEM) to sit for the certification examination and were required to meet either Residency requirements or Practice requirements to take the ABEM examination. Residency qualified physicians had completed an approved Emergency Medicine residency program. Practice qualified physicians met the Board's requirement by completing five continuous years practice of Emergency Medicine. Both groups met the additional credential requirements of the Board.

The total number of subjects, 616, consisted of 134 Residency and 482 Practice physicians. There were fewer Residency than Practice in the sample because Emergency Medicine is a new specialty. There were thirty-six approved residency programs at the time the Certification Examination was first offered. Relatively few practitioners of Emergency Medicine had the opportunity to complete a residency program. Since Emergency Medicine had only been recently organized and residency programs continued to be created, ABEM elected to establish the Practice qualifications to permit experienced emergency physicians to sit for the examination without requiring them to complete residency training.

Although selection was not random, these samples were presumed to be representative of their respective populations based on the expert judgment of Board members and project staff. Also, self-selection probably was a characteristic of the population since anyone who ever takes the exam will do so voluntarily.

EXAMINATION CONSTRUCTION

The Emergency Medicine Certification Examination was developed by Office of Medical Education Research and Development (OMERAD), Michigan State University personnel between February, 1975, and August, 1977. The organization and substantive details of the developmental processes have been reported by Dr. Jack L. Maatsch, the project director, and his associates (Maatsch, et al., 1976; Maatsch and Elstein, 1979).

The testing materials were field tested in October 1977 on samples of examinees drawn from four different populations: Residency physicians, Practice physicians, second-year Emergency Medicine residents, and fourth-year medical students. The primary purpose of the field test was to determine which test items and test formats adequately discriminated among these subject groups and to identify materials that did not discriminate so that they might be deleted from the item bank. The field test results have been reported by Downing (1979), Maatsch, et al., (1978), and Maatsch and Elstein (1979). Selected test materials that the field test demonstrated to be adequate comprised the first version of the ABEM certifying examination.

The examination was administered in two parts. Part I was administered in February, 1980, and served as a screen for Part II, which was adminstered in May, 1980. Part I consisted of multiple-choice questions, some with accompanying visual stimulus material, and Patient Management Problems which were latent image problem-solving tasks that required the candidates to show evidence of diagnostic and treatment capabilities. Unlike the paper-and-pencil format of Part I materials, Part II consisted of a series of examiner-administered and examiner-scored Simulated Patient Encounters. Simulated

Patient Encounters are considered to be one of the highest fidelity test formats of all the presently available standardized oral examination types.

Three hundred and seventy-two multiple-choice items (MCQ) and 136 pictorial multiple-choice questions were field tested in October of 1977, each of which had an associated visual or pictorial stimulus. Hence, these 136 items were called pictorial multiple choice questions (PMCQ's) unlike the MCQ's which had no pictorial correlates. Ineffective items were deleted based on the field test results, resulting in a library of 261 MCQ's and 102 PMCQ's. Materials were created from this pool for the first official administration conducted during February, 1980.

From the pool of 261 MCQ's, 197 items were selected and randomly assigned to one of three test booklets. Another pool was created of 80 experimental basic science items that had no apparent or intended relevance to clinical medicine in general, or emergency medicine in particular. The experimental basic science items were for research purposes only and in no way affected pass/fail decisions. These 80 items were also randomly distributed throughout the three test booklets. Finally, two PMCQ booklets were created by selecting 100 PMCQ's from the PMCQ library of 102 items. These 100 items were randomly assigned to one of two booklets of equal length. The resulting multiple-choice examination materials consisted of five test booklets.

DESCRIPTION OF SCALES

These combined test booklets contained several scales, some of which were created during analysis of the field test results and others which were

formed as part of the test materials used for the first official administration. A listing and description of these scales follows.

High Clinical-Relevance (HCR) Scale

The rationale for the creation of this scale is straightforward. Of the various testing formats currently being used, the simulated encounter is thought to be the best proxy for real-world performance. The material that is used in the simulation cases and the manner in which it is presented, more closely approximates the real-world than do paper and pencil tests, oral examinations, chart audits, etc. It follows that if simulation performance is clinically relevant, then we can rank-order multiple-choice items according to how highly they correlate with simulation performance. Those which correlate highest can be assigned to a high clinical-relevance scale. Correspondingly, those with the lowest such correlations can be assigned to a low clinical-relevance scale, etc.

Item-criterion point-biseral correlations were computed for each of the 363 multiple-choice items in the test library, using the grand mean ratings on the 12 simulated encounters as the criterion (Downing, 1979). The 91 items having the highest such correlations were then assigned to this HCR scale. These correlations ranged from .33 to .68 with a median of .38. The mean p-value or difficulty level (proportion of examinees correctly responding) for these 91 items was .721. Of these 91 items that composed the initial HCR scale, 76 were included in the 297 "non-basic science" items which were administered in February, 1980. The number of such items randomly distributed throughout Booklets 1 through 5 were 17, 15, 18, 16, and 10, respectively.

Low Clinical-Relevance (LCR) Scale

Just as the HCR was composed of those 91 items which correlated highest with the criterion, the LCR scale consisted of those 91 items which had the lowest such correlation. These correlations ranged from -.23 to .11 with a median of .05 and a mean p-value of .667.

The LCR scale was composed of 70 items randomly sampled from the 91 items in the library. These 70 items were randomly distributed throughout Booklets 1 through 5 with the resulting distribution being 14, 16, 18, 9, and 13, respectively.

Medium Difficulty (MD) Scale

Downing (1979) constructed the MD scale by selecting 91 items which had p-values which ranged from .50 to .69. The mean p-value of items comprising this scale was .603. Seventy of these 91 items were contained in the five test booklets with the respective number of items being assigned to Booklets 1 through 5 being 20, 17, 13, 12, and 8. This scale included items from the HCR and LCR scales as well as other items in the library meeting the selection criterion.

Basic Science (BS) Scale

Eighty (80) experimental basic science items were randomly distributed throughout the 197 "scorable" MCQ items. These items were obtained by random sampling from the item library of a standing test committee in the College of Human Medicine, Michigan State University. The items in the committee's library are submitted by selected faculty from the basic science departments in the College and are continually reviewed for psychometric quality and content relevancy. The items are designed to be administered to

second-year College of Human Medicine students and have no apparent relationship to the clinical practice of medicine. The items are written by discipline-based basic scientists, not practicing physicians. Moreover, they are designed to test undergraduate medical students' knowledge of basic science concepts and principles. Indeed, since second-year students have had little, if any, clinical experience, it would be inappropriate to test their clinical competence. Such skills would be acquired later in the curriculum and during graduate training.

All items in the committee's item library are coded by content domain which, in this case, is the discipline submitting the item and whose subject-matter knowledge the item measures. Accordingly, basic science subscales were formed, corresponding to each of the disciplines. The distribution of items by discipline and subscales is shown in Table 1.

Seven-hundred of each of the three Multiple-Choice Question Booklets were constructed so that each candidate could use a new Booklet in working through the MCQ's. Three-hundred and fifty of each of the two Pictorial Multiple-Choice Question Booklets were constructed by carefully assigning items from the PMCQ item library whose high psychometric qualities were demonstrated in field test item analysis results. These two PMCQ forms were assumed to be equivalent. Since the production costs of the PMCQ Booklets were substantially greater than the costs of the MCQ Booklets and the field test results indicated the PMCQ items performed no better than MCQ items, each candidate was scheduled to complete only one of the two PMCQ Booklets. One-half of the candidates were randomly assigned to Booklet 4, and the other half to Booklet 5. The composition of the five Booklets is shown in Table 2.

TABLE 1

TOTAL NUMBER OF ITEMS PER SCALE FIELD-TESTED AND ADOPTED

SCALE NAME		NUMBER OF ITEMS FIELD TESTED IN OCTOBER, 1977	NUMBER OF ITEMS IN THE FEBRUARY, 1980, MATERIALS
High Clinical-Relevance		91	76
Low Clinical-Relevance		91	70
Medium Difficulty		91	70
Basic Science		None	80
Basic Science Subscales:			
Anatomy	3		
Biochemistry	22		
Histology	6		
Microbiology	13		
Pharmacology	15		
Physiology	21		

TABLE 2

TOTAL NUMBER OF ITEMS PER SCALE WITHIN EACH TEST BOOK

		TOTAL	NUMBER OF ITEMS COMPRISING			
BOOKLET	TYPE	ITEMS	HCR	LCR	MD	BS
1	MCQ	92	17	14	20	26
2	MCQ	92	15	16	17	30
3	MCQ	93	18	18	13	24
4	PMCQ	50	16	9	12	0
5	PMCQ	50	10	13	8	0
	TOTALS	377	76	70	70	80

TESTING PROCEDURE

Individuals who wished to sit for the Emergency Medicine Certification Examination made application to ABEM prior to December 31, 1978. Applicants who were eligible for certification were advised by letter of the date, time, place, costs, accomodations, and format of the examination.

Part I of the examination was administered on Wednesday, February 20, 1980, in Philadelphia, Chicago, and Los Angeles. Candidates were assigned to the site that was nearest their current mailing address. At each of the sites, candidates were randomly assigned to one of two groups. The order in which each of the two groups was tested differed in the morning sessions with one group completing MCQ 1, MCQ 2, and then a pictorial booklet and the other group completing a pictorial booklet first and then taking MCQ 1 and MCQ 2 (See Table 3).

Creating two groups at each of the three sites was intended to facilitate the administration of the examination. It enabled the distribution and collection of materials to be done more expeditiously, made proctoring more manageable, and also made it possible for the same PMCQ booklets to be used by different candidates while allowing time for the ABEM staff to inspect each PMCQ booklet during the interim period.

The Chairman of the ABEM Test Committee was responsible for the overall coordination of the examination, which included dealing with any proctoring decisions or questions from chief examiners at the three sites. Assigned to each site was a chief examiner, a chief staff person, board member proctors, and staff proctors, each of whom had been charged with specific assignments and responsibilities. Central administrative ABEM

TABLE 3

EXAMINATION SCHEDULE

WEDNESDAY, FEBRUARY 20, 1980

TIME	ACTIVITY		
	Group A	Both Groups	Group B
Morning			
6:30 - 8:00		Registration	
8:00 - 8:10		Instructions to Candidates and distribution of materials	
8:10 - 9:10	PMCQ 4 or 5		
9:10 - 9:20	Instructions		
9:20 - 12:20	MCQ 1 and 2		
8:10 - 11:10			MCQ 1 and 2
11:10 - 11:20			Instructions
11:20 - 12:20			PMCQ 4 or 5
Afternoon			
12:20 - 1:30		LUNCH	
1:30 - 1:50		Afternoon Registration	
1:50 - 2:00		Instructions	
2:00 - 3:20		MCQ 3	
3:20 - 3:30		Instructions	
3:30 - 5:00		Patient Management Problems	

personnel had prepared a manual that was distributed and discussed with the site personnel shortly before the date of the examination.

In addition to the general instructions contained in the manual, specific instructions were provided to the candidates at the beginning of each testing session. Appendix A contains the cover sheet that appeared in Booklets 1, 2, and 3. These instructions directed the candidates on the use of the MCQ test materials and contained a sample question. Slightly different instructions prefaced the PMCQ Booklets and these, along with a sample question, are attached as Appendix B. Candidates were instructed to answer all items in each of the five multiple-choice booklets at a rate of approximately one item per minute.

DESIGN AND ANALYSIS PROCEDURE

The examinees had been instructed that there was only one correct answer to each item and that they should answer each item with their single best response. No correction for guessing was used. Each examinee was provided a separate machine scorable op-scan answer sheet with each of the three multiple-choice test booklets. All answer sheets were collected and centrally scored; the results of that scoring are the data base from which the analyses for this report were made.

The organization of the three test booklets is shown in Tables 4, 5, and 6, corresponding to Test Booklets 1, 2 and 3, respectively. The numbers 1 through 92, (93 in the case of Booklet 3) denote the item number and its serial position within the test booklet. No entry opposite that number indicates that item was not a member of any of the scales under investigation. Low Clinical Relevant items are abbreviated LCR; High Clinical Relevant items,

TABLE 4

COMPOSITION OF TEST BOOK I

1.
2.
3. BS 1 (micro)
4. HCR 1
5. BS 2 (pharm)
6.
7. LCR 1
8.
9. BS 3 (physio)
10. BS 4 (physio)
11. BS 5 (bioch)
12. HCR 2/MD 1
13.
14. LCR 2
15. LCR 3
16. BS 6 (hist)
17. BS 7 (bioch)
18. HCR 3
19.
20.
21.
22. BS 8 (anat)
23. LCR 4/ MD 2
24. BS 9 (hist)
25. BS 10 (micro)
26.
27. HCR4/ MD 3
28. MD 4
29. BS 11 (physio)
30.
31. LCR 5
32. LCR 6
33. BS 12 (physio)
34.
35. BS 13 (hist)
36. HCR 5
37. LCR 7
38. MD 5
39. MD 6
40. BS 14 (bioch)
41. BS 15 (micro)
42. LCR 8/MD 7
43. BS 16 (physio)
44. LCR 9
45. LCR 10
46. BS 17 (pharm)
47. HCR 6/MD 8
48. HCR 7
49. LCR 11
50.
51. BS 18 (micro)
52. MD 9
53. MD 10
54. LCR 12
55.
56. BS 19 (pharm)
57.
58. MD 11
59. LCR 13
60. HCR 8
61. BS 20 (pharm)
62.
63. BS 21 (physio)
64. MD 12
65. HCR 9
66. MD 13
67.
68. HCR 10/MD 14
69. BS 22 (pharm)
70.
71. HCR 11
72.
73.
74. BS 23 (physio)
75. BS 24 (bioch)
76.
77. MD 15
78. LCR 14
79. MD 16
80. HCR 12
81. BS 25 (hist)
82. HCR 13
83.
84.
85. MD 17
86.
87. HCR 14
88. HCR 15
89. BS 26 (physio)
90. HCR 16/MD 18
91. HCR17/ MD 19
92. MD 20

TABLE 5

COMPOSITION OF TEST BOOK 2

1. HCR 18
2. HCR 19
3.
4. HCR 20/MD 21
5. LCR 15/ MD 22
6. LCR 16
7. MD 23
8. MD 24
9. HCR 21
10. BS 27 (micro)
11. LCR 17
12.
13. MD 25
14. MD 26
15. MD 27
16. LCR 18
17. BS 28 (pharm)
18. HCR 22
19.
20.
21. BS 29 (physio)
22. HCR 23/MD 28
23. LCR 19
24. BS 30 (bioch)
25. BS 31 (bioch)
26.
27. HCR 24
28. LCR 20/MD 29
29. BS 32 (micro)
30. LCR 21
31. HCR 25
32.
33. BS 33 (bioch)
34. HCR 26
35. HCR 27/MD 30
36. BS 34 (physio)
37. BS 35 (micro)
38. BS 36 (physio)
39.
40.
41. BS 37 (physio)
42. BS 38 (physio)
43. MD 31
44. MD 32
45. BS 39 (bioch)
46. BS 40 (physio)
47.
48.
49. LCR 22
50. BS 41 (pharm)
51.
52.
53. BS 42 (physio)
54. MD 33
55. BS 43 (anat)
56. HCR 28/MD 34
57.
58. LCR 23
59. MD 35
60. BS 44 (bioch)
61. BS 45 (micro)
62.
63. LCR 24
64. HCR 29
65. BS 46 (hist)
66. LCR 25
67. LCR 26
68. HCR 30
69. BS 47 (physio)
70. LCR 27
71. LCR 28
72. BS 48 (pharm)
73. BS 49 (pharm)
74. MD 36
75. LCR 29
76.
77. BS 50 (physio)
78.
79.
80. BS 51 (micro)
81.
82.
83. HCR 31
84. BS 52 (bioch)
85.
86. LCR 30/MD 37
87. BS 53 (pharm)
88. BS 54 (bioch)
89. BS 55 (micro)
90.
91. HCR 32
92. BS 56 (bioch)

TABLE 6

COMPOSITION OF TEST BOOK 3

1.
2. HCR 33
3. BS 57 (physio)
4. HCR 34
5.
6. HCR 35
7. LCR 31
8. BS 58 (pharm)
9.
10.
11. HCR 36
12. LCR 32
13.
14. BS 59 (hist)
15.
16.
17. BS 60 (bioch)
18. HCR 37
19. HCR 38
20. LCR 33
21. BS 61 (physio)
22.
23.
24.
25. HCR 39
26. BS 62 (micro)
27. BS 63 (bioch)
28. BS 64 (micro)
29. LCR 34
30.
31.
32. LCR 35
33. BS 65 (pharm)
34. MD 38
35.
36. LCR 36
37. BS 66 (bioch)
38. BS 67 (bioch)
39. HCR 40
40.
41. MD 39
42. MD 40
43. HCR 41/MD 41
44. HCR 42
45. HCR 43
46. BS 68 (physio)
47.
48. LCR 37
49. LCR 38
50. HCR 44
51. LCR 39/MD 42
52. MD 43
53. BS 69 (bioch)
54. LCR 40
55. HCR 45
56.
57. LCR 41
58. LCR 42
59. HCR 46/MD 44
60. BS 70 (bioch)
61. LCR 43
62. HCR 47
63. LCR 44
64.
65. LCR 45
66.
67.
68. BS 71 (pharm)
69.
70.
71. LCR 46/MD 45
72. BS 72 (micro)
73.
74.
75. BS 73 (bioch)
76. HCR 48/MD 46
77. MD 47
78. BS 74 (anat)
79. BS 75 (pharm)
80.
81.
82. HCR 49/MD 48
83. LCR 47
84. BS 76 (bioch)
85.
86. BS 77 (pharm)
87. MD 49
88. HCR 50/MD 50
89. BS 78 (bioch)
90.
91. BS 79 (bioch)
92. BS 80 (physio)
93. LCR 48

HCR; Medium Difficulty items, MD; and Basic Science items, BS. Items having two such entries (e.g., item 12 in Booklet 1) appeared on two scales. Parentheses identify the Basic Science sub-scale of that item. Finally, the serial numbers were for identification, cross-checking, and referencing to the item library and have no other function.

Tables 4, 5, and 6 are summarized in Table 7 which shows the frequency of the various item types found in each of the three test booklets.

Since each of the 277 multiple-choice items were randomly assigned to one of the three test booklets, the booklets were considered to be equal in difficulty. To check on that and to determine whether fatigue or any other factors might have affected performance as candidates worked their way through the booklets, accuracy rates were computed for the 197 scorable items.

The entries in Table 8 are means of the p-values (proportion of examinees responding correctly), excluding the eighty Basic Science items and ignoring the scale assignment of the 197 items. Candidate performance remained stable throughout each booklet and throughout the day (see Table 8). This finding justified pooling the data from the three MCQ Books for all subsequent analyses.

An earlier analysis (Downing, 1979) showed there was no functional difference between MCQ items and PMCQ's. Downing found that the pictorial items correlate highly (.84) with the non-pictorial items. Indeed, the magnitude of that correlation is equivalent to the internal reliability of the MCQ test itself. Hence there was good reason to pool the MCQ's and the PMCQ's; namely, they measured the same underlying competency factor. However, the scales of interest were of different lengths in the two pictorial booklets. Moreover, all candidates didn't sit for these PMCQ books as they did

TABLE 7

COMPOSITION OF MCQ TEST BOOKS

	Book I	Book 2	Book 3	TOTAL
Basic Science (BS)	26	30	24	80
High Clinical Relevance (HCR)	17	15	18	50
Low Clinical Relevance (LCR)	14	16	18	48
Medium Difficulty (MD)	20	17	13	50

TABLE 8

PROPORTION OF CORRECT RESPONSES PER MCQ BOOK
(n = 616)

	Book I	Book 2	Book 3	TOTAL
Items 1 - 31	.76	.73	.76	.75
Items 32 - 63	.74	.74	.77	.75
Items 64 - 92 or 93	.74	.72	.76	.74
TOTAL	.75	.73	.76	

for the three MCQ books. Of the 616 subjects who completed the three MCQ books, 260 then completed PMCQ Book 4 and the other 356 candidates completed PMCQ Book 5. Since the scales of interest were comprised of different items in the two PMCQ books and also were of different lengths (see Table 2), the decision was made to eliminate those potential sources of confound by pooling the results of only one of the two pictorial books with the data from the three MCQ books.

Book 5 was randomly selected to be included. The composition of Book 5 is shown in Table 9. The number of scorable items comprising each of the scales when Books 1, 2, 3, and 5 are combined is as follows:

Items which appeared only on the LCR scale	51
Items which appeared only on the MD scale	32
Items which appeared only on the HCR scale	44
Items which appeared on both the LCR and MD scale	10
Items which appeared on both the HCR and MD scale	16

To eliminate the redundancy of the twenty-six joint items and to create scales of equal length, the following manipulations were performed prior to the data analysis:

1. One of the fifty-one LCR items was randomly eliminated.
2. The ten LCR/MD items were removed from the LCR scale and were assigned exclusively to the MD scale.
3. Eight of the sixteen HCR/MD items were randomly assigned solely to the MD scale, six to the HCR scale, and two were eliminated from the data base.

TABLE 9

COMPOSITION OF TEST BOOK 5

1.		26.	
2.	HCR 67	27.	
3.		28.	LCR 63
4.	LCR 58	29.	
5.	LCR 59	30.	LCR 64
6.		31.	HCR 74/MD 65
7.		32.	LCR 65/MD 66
8.	HCR 68	33.	
9.		34.	
10.		35.	MD 67
11.	HCR 69	36.	LCR 60
12.	HCR 70	37.	LCR 67
13.	HCR 71	38.	MD 68
14.		39.	
15.		40.	
16.	HCR 72	41.	
17.	HCR 73	42.	
18.	LCR 60	43.	LCR 68
19.	LCR 61	44.	LCR 69/MD 69
20.	LCR 62/MD 63	45.	MD 760
21.		46.	
22.		47.	HCR 75
23.	MD 64	48.	HCR 76
24.		49.	
25.		50.	LCR 70

These procedures yielded three independent scales of equal item length. After randomly removing thirty of the eighty items from the BS scale, each of the four independent scales could be directly compared to one another.

HYPOTHESES

1. H_0: Not all of the four scales (HCR, LCR, MD, BS) discriminate between the two physician groups.

   H_1: Each of the four scales discriminate between the two physician groups.

2. H_0: The HCR is not the most discriminating scale.

   H_1: The HCR is the most discriminating scale.

   Note: Downing found the HCR scale to be more discriminating than the LCR scale and the MD scale, although the difference was not statistically significant in the latter case. His findings, however, were based on subject groups which represented a broader range of clinical competence (residency qualified physicians, residents, and medical students) than the two groups used in the present study (residency qualified and practice qualified physicians). On the other hand, the sample size of the two groups in the present study is much larger.

3. H_O: Using a measure of internal consistency as an estimate of reliability, there is no difference among the reliabilities of the four scales.

   H_1: Not all of the scales reliabilities are equal.

4. H_O: The correlation between the HCR scale and the MD scale is no greater than the correlation between any other two scales.

   H_1: The correlation between the HCR scale and the MD scale is greater than the correlation between any other two scales.

5. H_O: The BS scale is not the most difficult scale.

   H_1: The BS scale is the most difficult scale.

6. H_O: Not all of the four BS subscales discriminate between the two physician groups.

   H_1: All four BS subscales discriminate between the two physician groups.

7. H_O: The median item correlation between the BS items and Part II does not differ from what Downing found the median item correlation to be between the LCR items and Part II.

   H_1: The median item correlation between the BS items and Part II does differ from the median correlation Downing found between the LCR items and Part II.

8. H_O: The median item correlations between each of the four BS subscales and Part II are zero.

   H_1: The median item correlation between each of the four BS subscales and Part II are not zero.

ANALYSIS METHODS

To determine which, if any, of the four scales discriminated between the two physician groups four one-way analyses of variance were performed. These univariate F tests were used to test hypothesis 1.

Hypothesis 2, which dealt with the relative discriminating power of the scales, was tested by the SPSS program Discriminant Analyses (Nie, 1970).

Cronbach Alphas were computed for each of the four scales and were then transformed using Fisher's Z transformation of r and z-tests were performed to test Hypothesis 3. This same method was also used to test for differences between the correlations in Hypothesis 4 and 7.

Hypothesis 5 was tested by testing the differences between proportions. t-tests were computed to test Hypothesis 6 and tests to determine if correlations differ from zero were performed to test Hypothesis 8. The results of these tests are presented in the following chapter.

CHAPTER IV

RESULTS

Items were scored right or wrong and no corrections or adjustments were made to the raw scores. Proportions of correct responses for each subject were computed for each scale. Those results, collapsing across subject groups, are shown in Figures 1, 2, 3, and 4.

The distribution of the two candidate groups are shown in Figures 5, 6, 7, and 8. Means and standard deviations for each group and each scale are shown in Table 10 along with the reliabilities of each scale.

Null hypothesis number one stated that not all of the four scales discriminate between the two physician groups. To test this hypothesis discriminant functions analysis was performed using the computer program of Nie, et al (1970). That program allows for a one-way analysis of variance to be performed on each factor of interest. The results of those analyses are reported as Fs in Table 11. Since each is statistically significant, null hypothesis number one is rejected, the conclusion being that each of the scales do discriminate between the two subject groups ($p < .0001$).

It was hypothesized that the HCR is not the most discriminating scale. This is null hypothesis number two which contrasts with the alternate hypothesis which stated that HCR is the most discriminating scale. To determine the relative discriminating power of the four scales, Wilk's Lambda, a measure of discriminating ability, was computed for each scale. Being an inverse measure, the smaller the Lambda value, the greater the discrimination. These direct comparisons serve as a test of hypothesis number two to determine if HCR is the superior scale. The scales are presented in

FIGURE 1

HIGH CLINICAL RELEVANCE SCALE

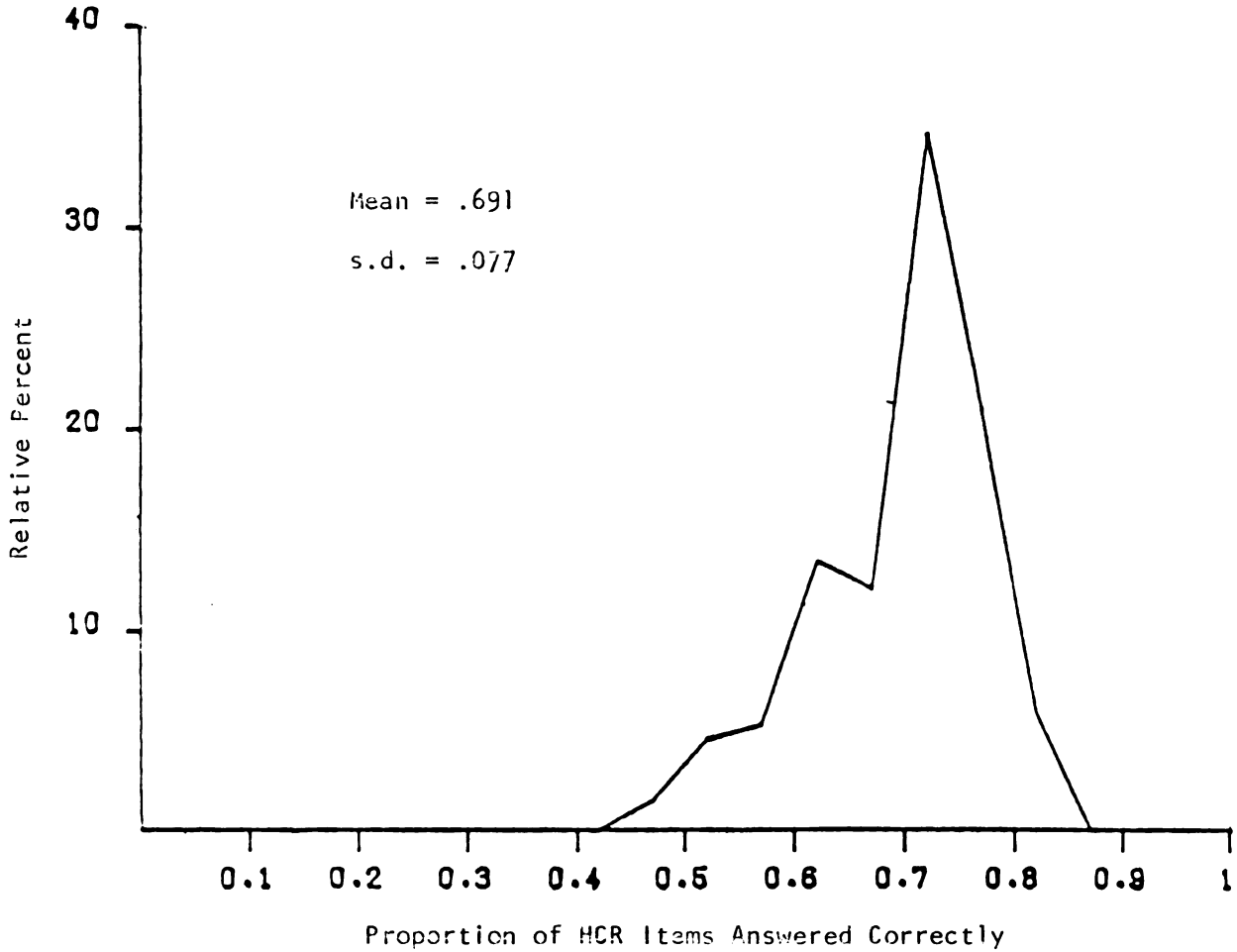

FIGURE 2

BASIC SCIENCE SCALE

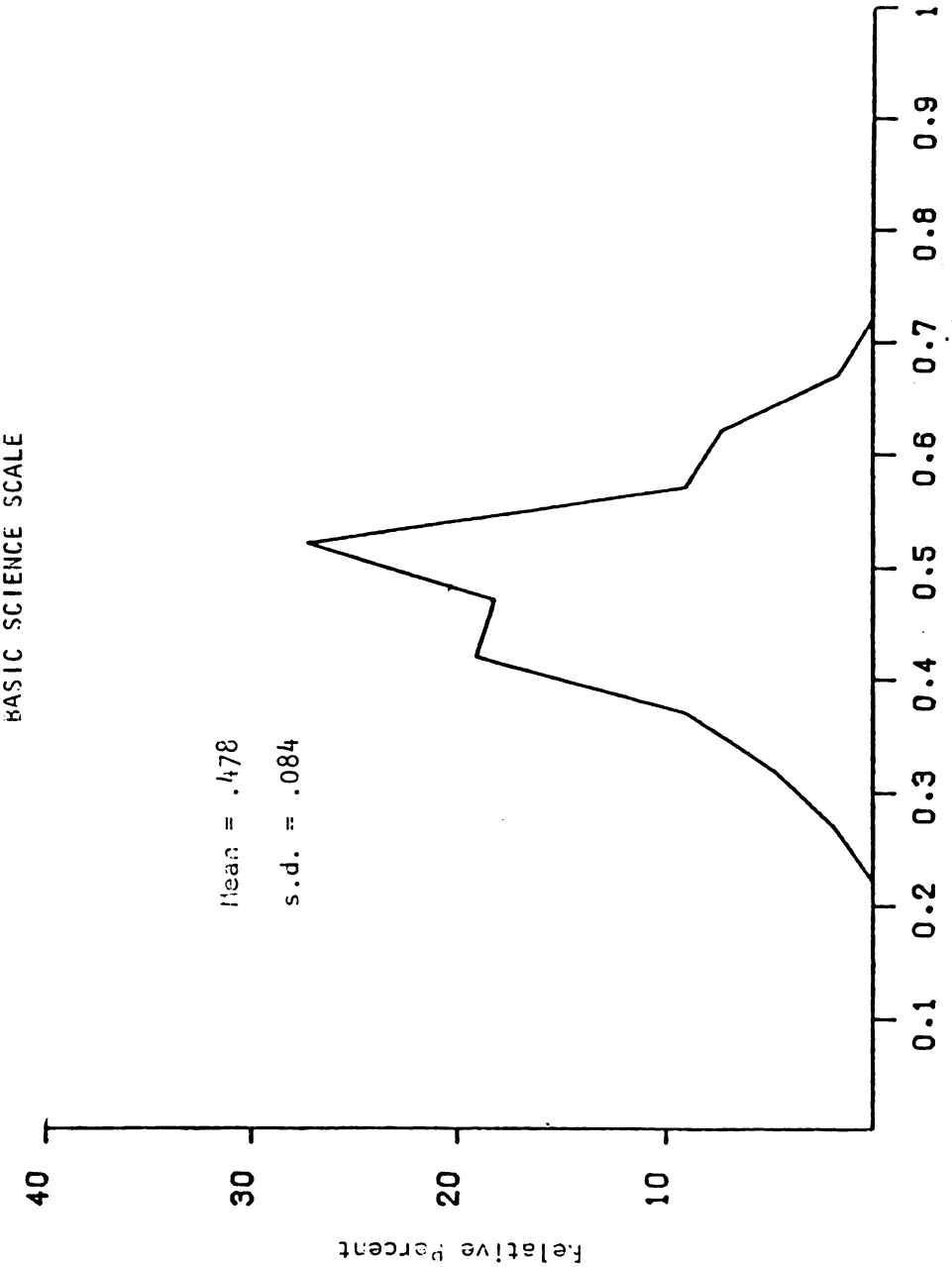

FIGURE 3

LOW CLINICAL RELEVANCE SCALE

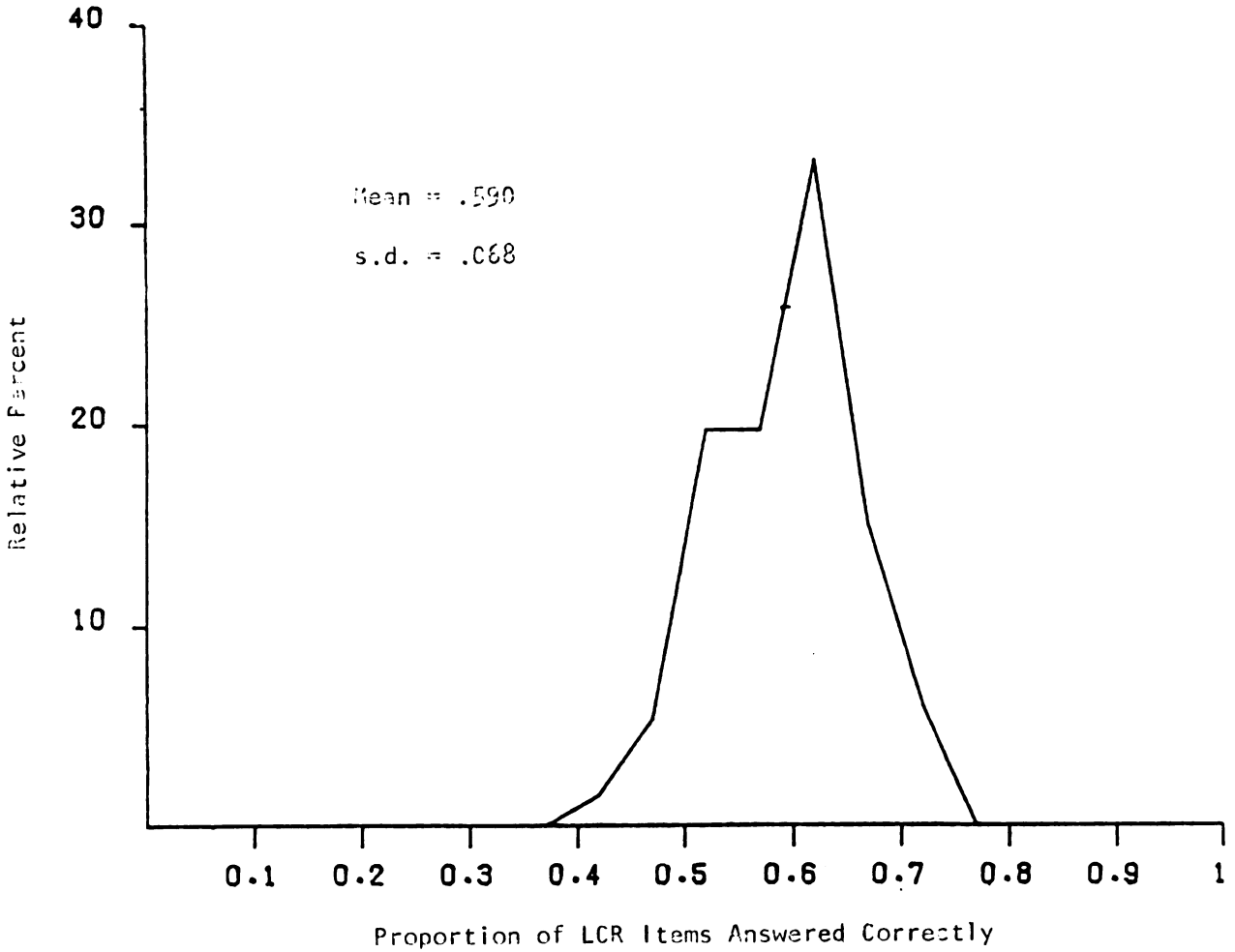

FIGURE 4

MEDIUM DIFFICULTY SCALE

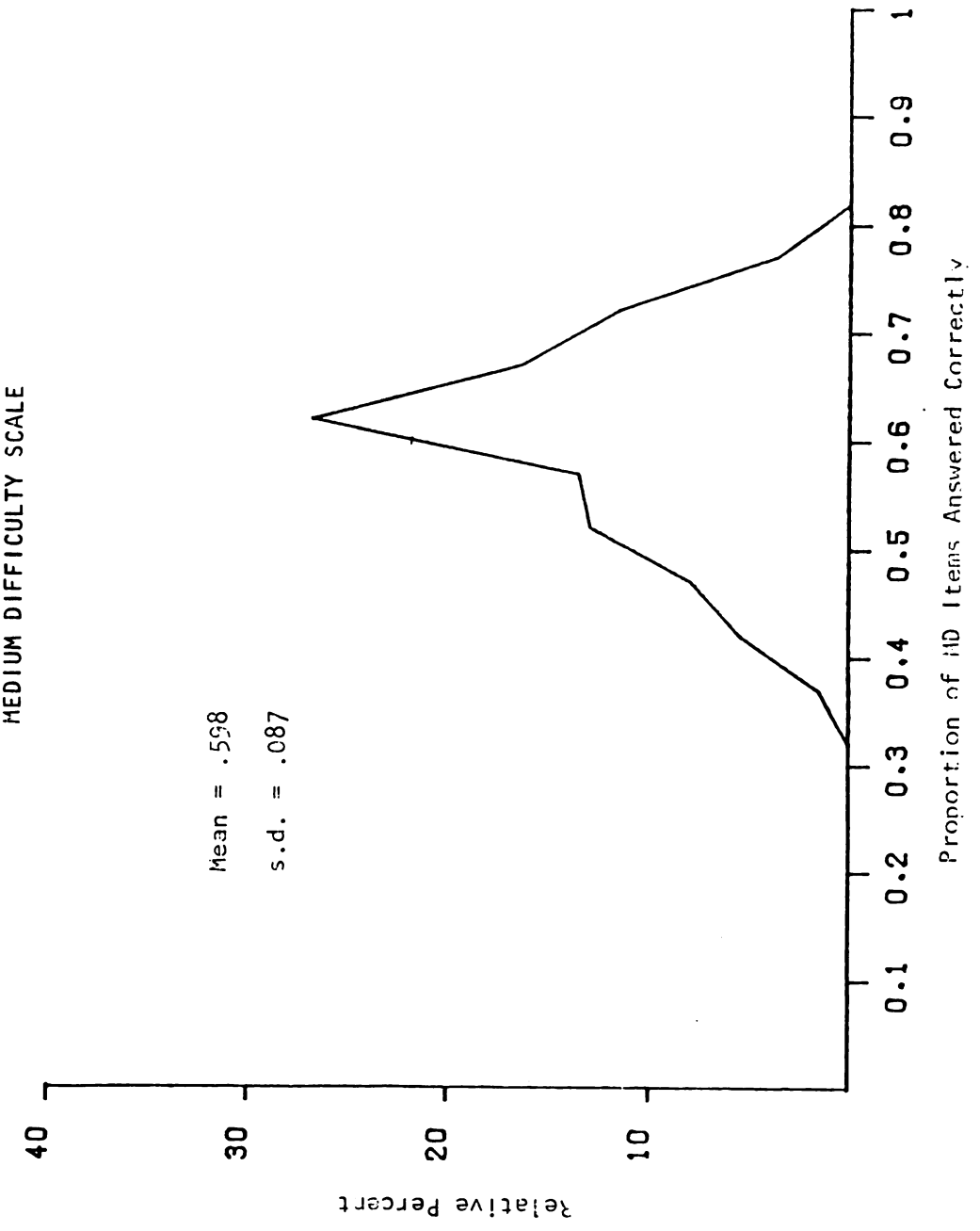

FIGURE 5

HIGH CLINICAL RELEVANCE SCALE

----- = Practice-Qualified Group

——— = Residency-Qualified Group

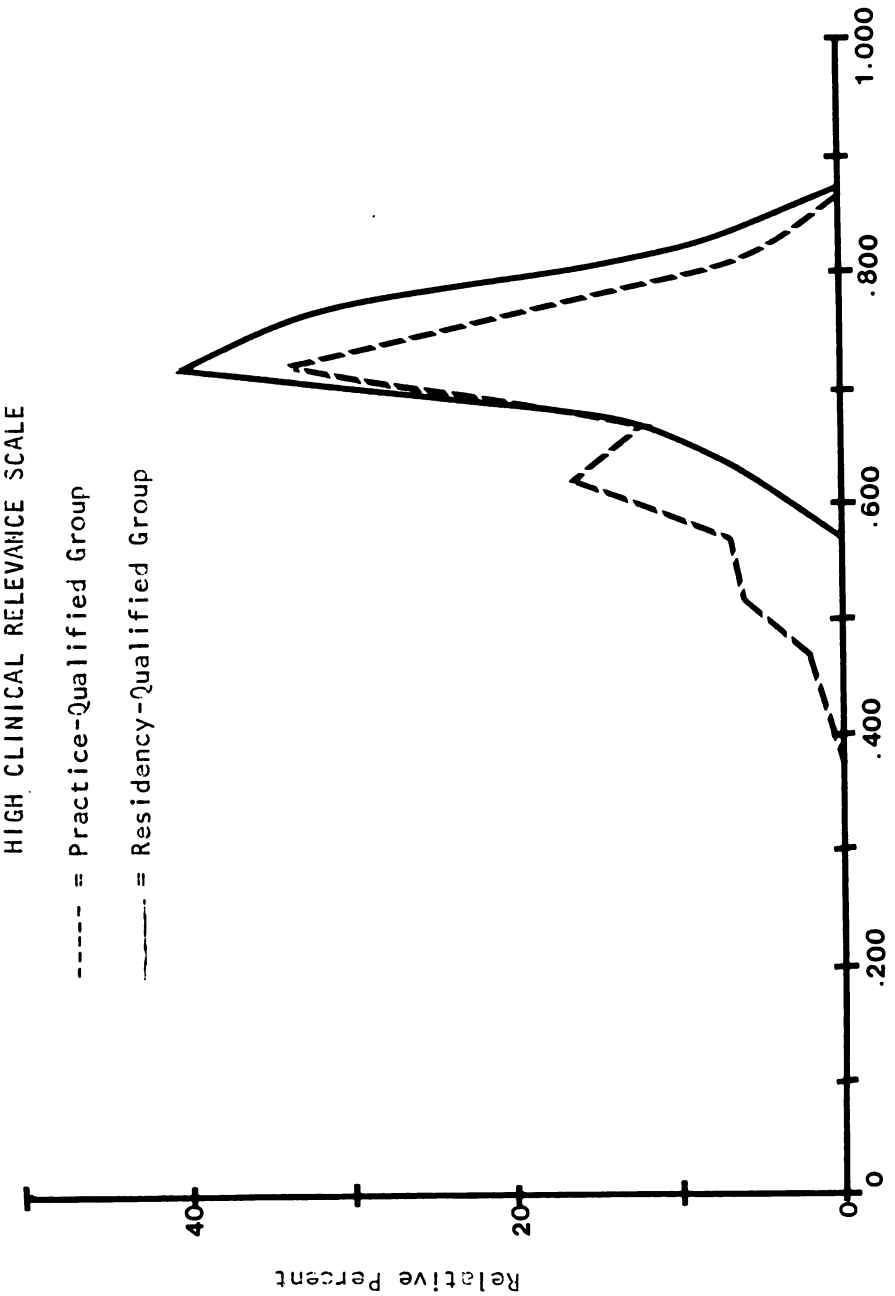

FIGURE 6

BASIC SCIENCE SCALE

----- = Practice-Qualified Group

——— = Residency-Qualified Group

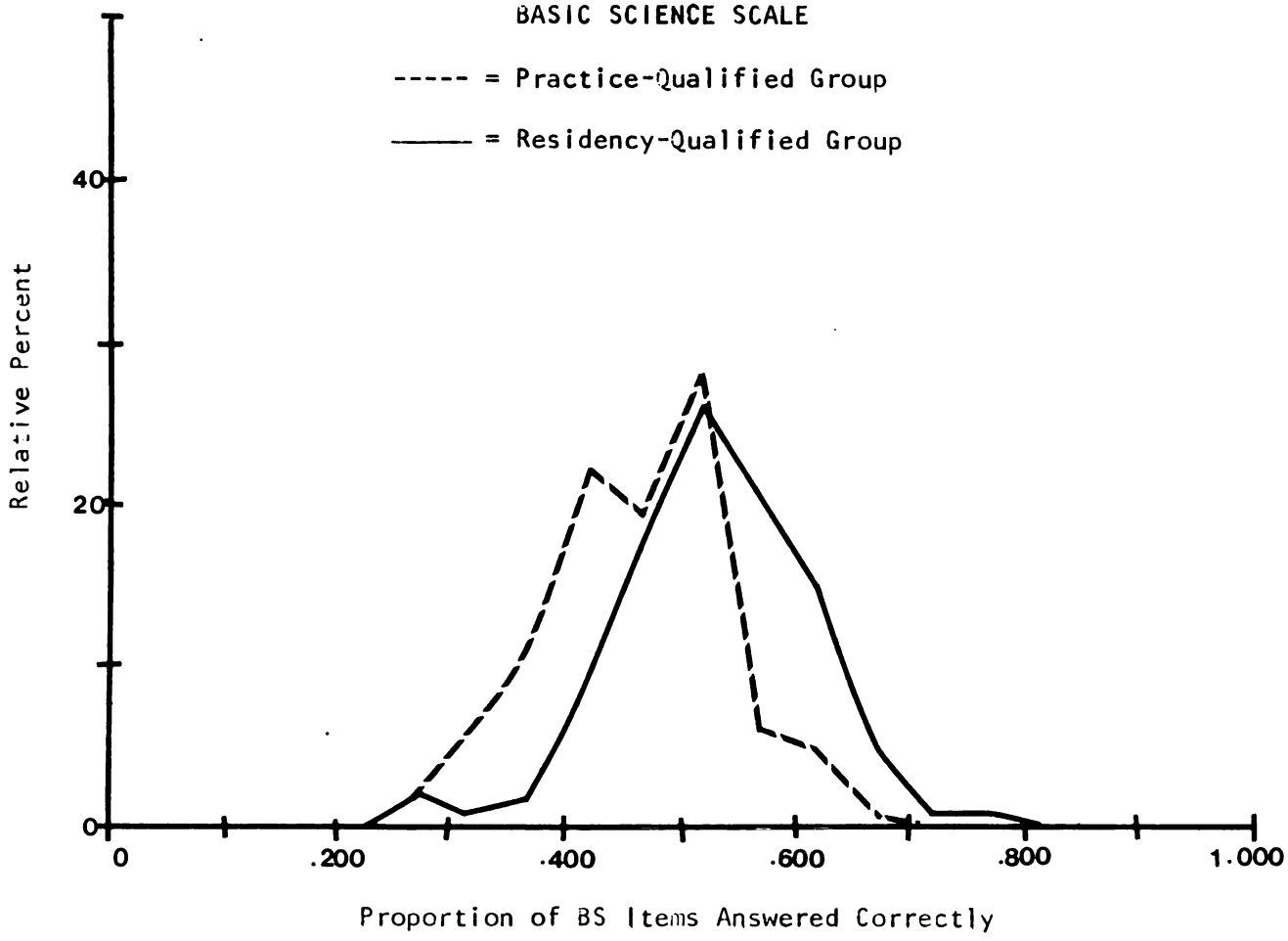

FIGURE 7

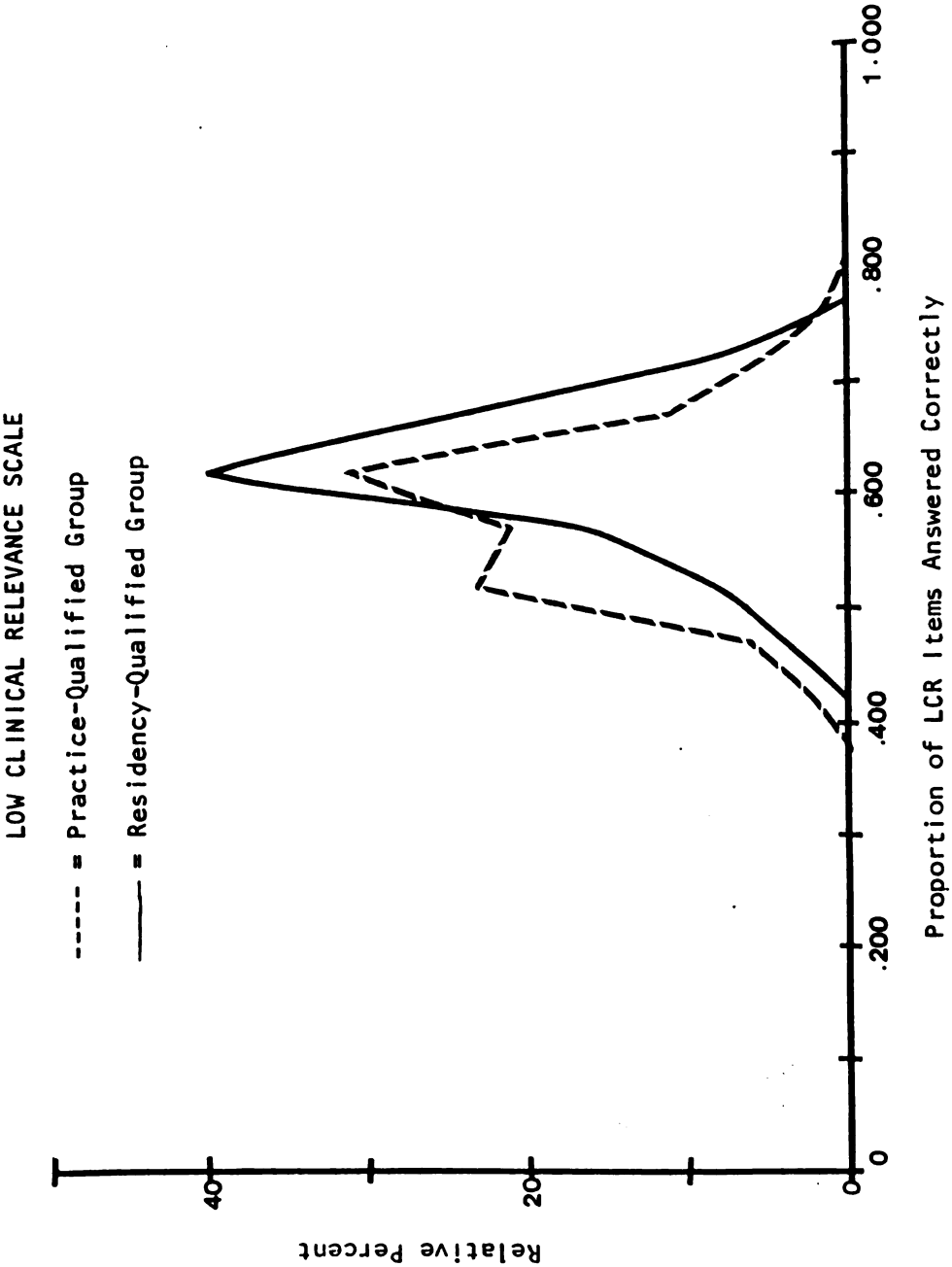
LOW CLINICAL RELEVANCE SCALE
----- = Practice-Qualified Group
——— = Residency-Qualified Group
Relative Percent
40
20
0
0
.200
.400
.600
.800
1.000
Proportion of LCR Items Answered Correctly

FIGURE 8

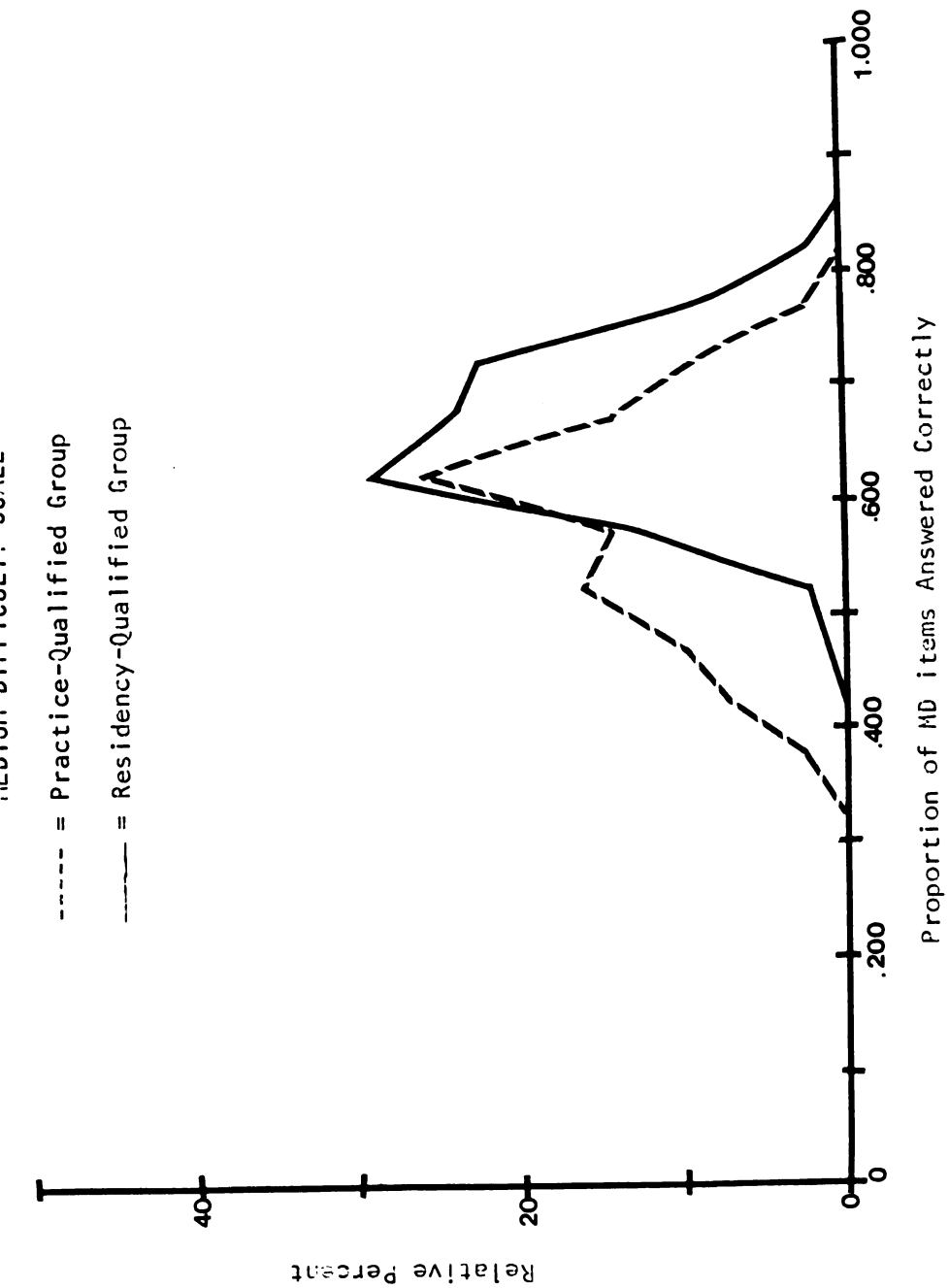
MEDIUM DIFFICULTY SCALE
----- = Practice-Qualified Group
——— = Residency-Qualified Group
Relative Percent
40
20
0
0
.200
.400
.600
.800
1.000
Proportion of MD items Answered Correctly

TABLE 10

GROUP MEAN PROPORTIONS OF CORRECT RESPONSES BY SCALE

(Parenthetical entries are standard deviations)

(k = 50 for each scale)

	HCR	BS	LCR	MD
Residency (n = 82)	.737 (.044)	.527 (.089)	.618 (.058)	.663 (.067)
Practice (n = 274)	.678 (.087)	.463 (.082)	.582 (.071)	.578 (.094)
Totals	.691 (.077)	.478 (.084)	.590 (.068)	.598 (.087)
Alpha	.737	.501	.346	.639

TABLE 11

DISCRIMINANT ANALYSES

Scale	Wilk's Lambda	$F_{1, 354}$
MD	.8594	57.90*
BS	.9075	36.06*
HCR	.9124	33.99*
LCR	.9545	16.88*

*p < .0001

TABLE 12

RESULTS OF STEPWISE SELECTION DISCRIMINANT ANALYSIS

Step	Entered	Wilk's Lambda	p	Standardized Discriminant Function Coefficients	Canonical r
1	MD	.8594	.0001	.746	
2	BS	.8300	.0001	.472	.41

Table 11, rank-ordered by Lambda value which suggests the relative importance of each scale. That is, each of the four scales discriminates ($p < .0001$) but the MD scale is the single best discriminating scale, the LCR scale the poorest of the four scales. Accordingly, null hypothesis number two which stated the HCR is not the most discriminating scale cannot be rejected.

To test in yet another way null hypothesis number two, a stepwise scale selection process was performed. This procedure loads the scales according to their independent and combined ability to maximize discrimination. Those results are shown in Table 12. The first scale selected was the medium difficulty which had the highest univariate F-ratio of the four scales (see Table 11). As an aside, this scale preference confirms the previous decision not to reject null hypothesis number two since the HCR scale is apparently not the most discriminating scale.

Having selected the first scale, F-ratios are then recomputed for each of the remaining three scales. On the basis of the relative values of these partial F-ratios the basic science was next selected. Since the partial F value for the Basic Science scale was significant ($p < .001$), its contribution to the first selected MD scale is statistically significant. As a practical matter, however, its addition to the MD scale was slight and served to reduce Lambda by less than 3/100 (see Table 12).

Partial F's were then again computed on the remaining HCR and LCR scales. Each F was less than one which resulted in the values (scales) not being selected. The computer program required an F to be greater than one to be considered for selection. An F of one is equivalent to $p = .5$ for large samples (SPSS^X, 1986, p. 698).

The best discrimination results from the MD/BS combination and to add to the HCR or LCR would only add noise to the prediction. Table 12 also

shows the MD/BS scales had a canonical correlation of .41 which is the relationship between the discriminant function and the prediction. This correlation squared equals 1 - Lambda and also equals the proportion of explained variance.

The discriminant function coefficients shown in Table 12 can be interpreted much like beta weights in a regression equation. Dividing the MD coefficient by the BS coefficient reveals that the MD scale is weighted by a ratio of 1.58: 1 relative to the BS scale.

Null hypothesis number three stated there is no difference among the reliabilities of the four scales. Cronbach alphas, measures of internal consistency, were computed for each of the four scales as estimates of the scale's reliability. Those values are presented in Table 13 which also shows the Fisher Z transformation value for each of the four reliabilities. Tests of differences between each pair of reliabilities were performed using the method of Glass and Stanley (1970, p. 308). These z tests, or critical ratios (Guilford, 1956, p. 194; McNemar, 1962, p. 50) yielded values which are presented in Table 14 which shows each reliability differed from each of the other three ($p < .001$, two-tailed tests). In other words, no two scale reliabilities are statistically equivalent. Null hypothesis number three is therefore rejected. Not only is there some difference among the scale's reliabilities, they are all different from one another.

To determine the relationships between the scales an intercorrelation matrix of Pearson coefficients was calculated. Those results are shown in Table 15. The reason these correlations are only of moderate strength was thought to be a function of the scales' reliabilities which are shown in Table 13 and are based on 50 item tests. When the intercorrelations are corrected for attenuation on both scales, the values increase considerably. See Table 16.

TABLE 13

CRONBACH ALPHA RELIABILITIES

Scale	Alpha	Fisher Z
HCR	.737	.944
MD	.639	.757
BS	.501	.550
LCR	.346	.367

TABLE 14

Z VALUES OF DIFFERENCES BETWEEN EACH PAIR OF SCALE RELIABILITIES

	HCR	MD	BS	LCR
HCR	---	3.515*	7.406*	10.846*
MD	---	---	3.891*	7.331*
BS	---	---	---	3.439*
LCR	---	---	---	---

*p <.001, two-tailed tests

TABLE 15

INTERCORRELATION MATRIX OF SCALES

(k = 50 for each scale)

	BS	HCR	LCR	MD
BS	—	.427	.327	.390
HCR	—	---	.538	.688
LCR	—	---	---	.460
MD	—	---	---	---

TABLE 16

INTERCORRELATION MATRIX OF SCALES

WHEN BOTH SIDES ARE CORRECTED FOR ATTENUATION

	BS	HCR	LCR	MD
BS	—	.703	.785	.689
HCR	—	---	1.065	1.002
LCR	—	---	---	.978
MD	—	---	---	---

An intercorrelation between two scales, for example, is thought to be less than its true value because of the unreliability of the scales being correlated To correct (adjust) for attenuation the following formula is used:

$$\bar{r}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}\ r_{yy}}}$$

where the corrected correlation equals the uncorrected correlation divided by the square root of the product of each measure's (scale's) estimated reliability.

The practice of correcting for attenuation is not without controversy. For example, corrected values can be greater than 1.0 in cases where the reliability(s) might have been underestimated (Nunnally, 1978, p. 237). Since two of the corrected evaluations shown in Table 16 are greater than 1.0, it may be that some of the reliabilities shown in Table 13 are underestimated.

Null hypothesis number four stated that the correlation between the HCR scale and the MD scale (.688, see Table 15) is no greater than any of the other correlations shown in Table 15. To test this hypothesis the values shown in the table were transformed to Fisher Z's. Tests of the differences of all pairs were then computed. The results, shown in Table 17, indicate that the correlation between the HCR scale and the MD scale is significantly greater than all of the other correlations shown in Table 15. Null hypothesis number four is thereby rejected. The HCR/MD correlation is greater than all the other correlations shown in Table 15. The only other difference among the correlations in Table 15 is the difference between the HCR/LCR correlation (.538, the second highest value in Table 15) and the BS/LCR correlation (.327 the lowest value in the Table).

Hypothesis number five dealt with the relative difficulty of the BS scale. The null hypothesis stated that the BS scale was not the most difficult scale; the alternate, or research hypothesis, posited that it was. The proportion of

TABLE 17

TESTS OF DIFFERENCES BETWEEN THE

SCALE INTERCORRELATIONS SHOWN IN TABLE 15

r	=	.427	.327	.390	.538	.688
Z	=	.456	.339	.412	.601	.844

r	Z					
.427	.456	----				
.327	.339	1.140	---			
.390	.412	.429	.711	---		
.538	.601	1.413	2.550*	1.842	---	
.688	.844	3.782*	4.922*	4.211*	2.368*	---
.460	.497	.399	1.539	.838	1.103	3.382*

* $p < .05$, two tailed test

correct responses of all candidates for each scale are shown in Table 10 above to be .691, .598, .590, and .478 for HCR, MD, LCR, and BS, respectively. T-tests for related groups were calculated to test the significance of the difference between the means of the BS scale and each of the other three scales. The t-values were 16.719, 9.269, and 8.876 for the HCR, LCR, and MD scales, respectively. Each t-value is significant ($p < .01$, two-tailed tests) even after dividing alpha by the number of tests performed (3) to correct for what would otherwise be an inflated likelihood of a Type 1 error (Hays, 1981, pp. 299, 425; Kirk, 1968, p. 78). Null hypothesis number five is rejected in favor of the alternate hypothesis which states the BS scale is the most difficult scale.

Basic Science Subscales

The eighty items which comprised the Basic Science scale represented six different subject matter subscales. However, since the anatomy subscale consisted of only three items and the histology subscale six items, those two subscales were not included in the subscale analysis. Group means and standard deviations for the four remaining subscales are shown in Table 18.

Null hypothesis six states that not all of these four subscales discriminate between the two physician groups. Shown in Table 18 are the t-values which were computed to test for differences between the subject group means on each of the four subscales. After dividing alpha by four to correct for the likelihood of a Type 1 error (Hays, 1981, pp. 299, 425; Kirk, 1968, p.78), each of the t's is significant ($p < .01$ in each case). Null hypothesis number six is thus rejected in favor of the alternate hypothesis which holds that each of the four BS subscales do discriminate.

TABLE 18

GROUP MEANS OF CORRECT RESPONSES BY BS SUBSCALE

(Parenthetical entries are standard deviations)

	Biochemistry (k = 22)	Microbiology (k = 13)	Pharmacology (k = 15)	Physiology (k = 21)
Residency (n = 82)	.405 (.120)	.590 (.150)	.665 (.125)	.553 (.114)
Practice (n = 274)	.340 (.106)	.510 (.144)	.564 (.156)	.471 (.125)
t-values:	6.06*	5.69*	6.90*	6.78*

* p < .01

Cronbach alpha estimates of reliability were computed for each of the four subscales. Those values are shown in Table 19 as are their respective Fisher Z values. Tests of differences between each pair of reliabilities were performed (Glass and Stanley, 1970, p. 308). Those results are presented in Table 20 which shows the four reliabilities differed from one another with but one exception.

Item-criterion point biserial correlation coefficients were computed between each BS item and the grand mean of Part II performance. Data for these computations were the responses of those 372 of the 616 who sat for Part II in the spring of 1980. The correlations between each of the BS items and Part II were arranged in intervals and are presented in Table 21. The median correlation was .02. Twenty-nine of the eighty BS items correlated negatively with Part II performance.

Null hypothesis number seven stated that the median correlation between the BS items and Part II (.02) and the equivalent value Downing found between his LCR items and Part II (.05) would not be statistically different. The differences between these two studys' correlations are shown in Table 22. To test this hypothesis a median test for independent groups was performed (Hays, 1973, 765-768). The test showed the correlations did not differ significantly (p=.5) nor do either of the two values differ from zero (Glass and Stanley, 1970, p. 536). Null hypothesis number seven of no difference cannot be rejected. Both values are held to be statistical equivalents.

Null hypothesis number eight states that the median item correlations between each of the BS subscales and Part II are zero. The median item-criterion correlations for each of the four BS subscales are shown in Table 23. None of the values in the table are significantly different than zero (Glass and Stanley, 1970, p.536), nor is there any significant difference between any two

TABLE 19

CRONBACH ALPHA RELIABILITIES OF BS SUBSCALES

Subscale	Alpha	Fisher Z
Biochemistry (k = 22)	.398	.421
Microbiology (k = 13)	.354	.370
Pharmacology (k = 15)	.156	.157
Physiology (k = 21)	.521	.577

TABLE 20

TESTS OF DIFFERENCES BETWEEN BS SUBSCALE RELIABILITIES

	Biochemistry	Microbiology	Pharmacology	Physiology
Biochemistry	----	1.262	6.535*	3.861*
Microbiology		----	5.272*	5.124*
Pharmacology			----	10.396*
Physiology				----

*p <.001, two-tailed tests

TABLE 21

FREQUENCY BY INTERVALS OF BS ITEMS - PART II CORRELATIONS

r	f
.15-.19	1
.10-.14	11
.05-.09	16
0-.04	23
-.05--.01	21
-.10--.06	7
-.15--.11	0
-.20--.16	1

TABLE 22

COMPARISON OF THE VALIDITY OF BS AND LCR ITEMS

	BS Items	Downing's LCR Items
Number of items comprising this scale	80	91
number of negative discriminators	29	28
proportion of negative discriminators	.36	.30
median correlation with Part II	.02	.05
modal correlation with Part II	-.01	.11

TABLE 23

BS SUBSCALES MEDIAN ITEM CORRELATIONS WITH PART II

Subscale	Median Item Correlation with Part II
Biochemistry	.025
Microbiology	.013
Pharmacology	.022
Physiology	.024

of the four tabled values (Glass and Stanley, 1970, p. 308; p.> .2 in all cases). Null hypothesis eight cannot be rejected; the conclusion is that the correlations shown in Table 23 do not differ from zero.

SUMMARY OF RESULTS

The results of the statistical tests of the hypothesis are summarized as follows:

	H_0.	Decisions and Conclusions
1.	Not all of the four scales (HCR, LCR, MD, BS) discriminate between the two physician groups.	Reject H_0. Each of the four scales discriminate between the two groups.
2.	The HCR is not the most discriminating scale.	Do not reject H_0. MD is the most discriminating scale. HCR is the third most discriminating scale. MD and BS combined form the most efficient combination scales.
3.	There is no difference among the estimated reliabilities of the four scales.	Reject H_0. The scale reliabilities are all different from one another.

4.	The correlations between the HCR scale and the MD scale is no greater than the correlation between any other two scales.	Do not reject H_0. The HCR/MD correlation is higher than the correlation between any other pair of scales.
5.	The BS scale is not the most difficult.	Reject H_0. BS items are the most difficult with the typical candidate answering correctly less than one-half of the BS items.
6.	Not all four of the BS subscales discriminate between the two physician groups.	Reject H_0. All four BS subscales do discriminate even though the subscales had as few as 13 items.
7.	The median correlation between the BS items and Part II does not differ from what Downing found the median correlation to be between the LCR items and Part II.	Do not reject H_0. Although the BS items correlated less with Part II (.02) than did Downing's LCR items (.05), the difference is not significant.

8.	The median item correlations between each of the BS subscales and Part II are zero.	Do not reject H_0. None of the four values differ from zero, nor do they differ among one another, ranging from .013 to .025.

These results will be discussed in the following chapter along with implications and recommendations which can be drawn from them.

CHAPTER V

SUMMARY AND CONCLUSIONS

DISCUSSION

This study found that each of four scales (HCR, MD, LCR, BS) discriminated between two groups of physicians which earlier research had shown differed in performance on test library items. The findings are similar to Downing's whose HCR and MD scales discriminated between three groups (residency trained, residents, and students) and whose LCR scale distinguished the residency trained without detecting a significant difference between residents and students.

On the other hand, Downing found the HCR scale to be the most discriminating scale, the MD second most, and the LCR scale the least discriminating. These findings contrast with the present results, which found the scale's ranked MD, BS, HCR, and LCR from high to low in relative discriminating ability. The most effective combination was created by adding the BS scale to the MD scale. Adding either the HCR or LCR to the BS/MD combination didn't improve that combination in a statistically significant way. The major agreement between the two studies is that the LCR scale has the least relative discriminating power. Beyond that, the HCR, which Downing found superior to all others, finished a distant third in the present study.

The reliability estimates of Downing's scales were considerably higher than those attained in the present study. Whereas Downing's reliabilities

ranged from .95 to .58, the values found in this study ranged from .73 to .34. It should be noted that Downing's values were based on a scale length of .91 items (50 in the present study). Moreover, Downing's subjects represented a much wider range of competence than the population that sat for the actual examination. Downing studied residency trained physicians, physicians who met the five year practice requirement, residents, and medical students. Those subjects who met the practice requirement performed very much like residents and were not included in most of his data analyses. Accordingly, Downing had represented in his study a wide range of competence: from students to residency trained specialists. Since reliability is a function of the variance of the test scores, which presumably should mirror the variability in competence represented by the sample being tested, it is not at all surprising that the reliabilities obtained by Downing were, in all cases, higher than those here obtained. Nonetheless, the breadth of range of values represented in the two studies is quite similar.

Downing found the HCR scale reliability was significantly higher than the MD scale reliability. That finding is in accord with the results of this study, which found the reliability of the HCR to be the significantly highest of all the scales. Both studies found the LCR scale reliability to be the lowest of all the scales.

In Downing, and in the present study, measures of internal consistency were used to estimate the scale's reliability. These values represent the scale's unidimensionality from which reliability is inferred. That may explain how HCR items, which both studies found to be the easiest type of item, nonetheless are the most reliable. By way of contrast, LCR items, the least reliable, by not being clinically relevant are lacking this feature of unidimensionality. Perhaps being least reliable and being least related to the

criterion are features that necessarily co-exist. This notion is given credence by examining the BS scale, which, like the LCR, was statistically unrelated to physician competence (see Table 22, p. 79) and, in addition, was significantly less reliable than the HCR and MD scale.

Just as the relatively truncated range of competence represented among the candidates in the present study probably accounts for the reliabilities which are less than those of Downing, it is likely a major cause of the smaller scale intercorrelations in the present study being smaller than those reported by Downing. Also, Downing's inter-scale correlation coefficients are inflated by auto-correlation, since some of his items comprised more than one scale. For example, a high (or low) relevance item might also be a medium (or low) difficulty item. No such item auto-correlation effect occurred in the present experiment, since no item served on more than one scale. (See page 53 above for a description of the steps taken to control for item redundancy across scale.) Even though the values in this study are lower, the intercorrelations in this study cover almost as wide a range as Downing's. In addition, the intercorrelations from the two studies appear to be rank-ordered similarly. If the values of the intercorrelation matrices from the two studies are rank ordered by value, an interesting pattern emerges:

	Downing	This Study
HCR/MD	.922	.688
LD/HCR	.879	
LD/MD	.772	
LCR/MD	.571	.460
LD/LCR	.465	
LCR/HCR	.458	.538
BS/HCR		.427
BS/MD		.390
BS/LCR		.327

Downing employed a low difficulty scale (LD) which has no equivalent in the present study, and if intercorrelations which deal with Downing's LD scale are disregarded, a near perfect parallelism is evident. The two studies are in agreement as to which pair of scales yields the highest intercorrelation (HCR/MD), Downing's second highest (.571) is this study's third highest and vice versa. This study's intercorrelations between the BS scale and each of the other three scales are the three smallest values in the above listing.

Several conclusions can be drawn from these results. The agreement that the HCR/MD intercorrelation is greater than all others may suggest that the underlying factor substructure may be linking those two substructures more so than any of the others. If this is so, it follows that the BS scale, having much lower correlations with the other scales, may be somewhat of an aberration.

THE PARADOX

To say that clinically relevant items discriminate between groups is not to say that items which discriminate between groups are clinically relevant. The basic science items proved to be very good discriminators -- even outperforming the HCR items. To now argue that the BS items aren't as unrelated to clinical practice as was earlier assumed seems inappropriate. It requires that we disregard the fact that the items were written primarily by non-physicians to be administered to students having no, or very little, clinical experience. It also requires that we disregard the correlational findings which showed the items were uncorrelated with the criterion which is a widely accepted proxy for the quality of clinical practice. And most disturbing, it would require that we disregard the feedback of hundreds of candidates who protested the inclusion of the BS items in the examination. They saw the items as irrelevant to their professional practice, void of any validity, and an insult to the integrity of serious professionals assembled for the sole purpose of demonstrating their capacity to deliver high quality health care.

Moreover, if the BS items have some inherent relevance, then they should correlate with the other scales, to the extent those other scales are intercorrelated. However, the scale intercorrelations which includes the BS scale are lower than the other intercorrelations (see Table 15, p. 72).

Herein lies the paradox. Items which discriminate between the subject groups, according to Downing, should be clinically relevant -- and the more clinically relevant they are, the better they should discriminate. The extent to which items are sensitive discriminators is directly related to the degree of clinical relevance inherent in those items.

Hence, the BS items which not only discriminate, but were significantly superior to the LCR, and even the HCR items, must be clinically relevant -- even more clinically relevant than the high clinical relevance items.

Not only is such a conclusion counter intuitive, it does not fit with the candidates' reaction to the BS items, nor with the fact the items were written by basic scientists. Probably the most positive evidence that the BS items are not clinically relevant lies in the scale's intercorrelation matrix. The BS items simply do not correlate with the other scales very well.

How then can items which are not clinically relevant discriminate between the subject groups so well? This writer suggests that the relevance dimension introduced by Downing may be of negligible influence in any setting other than his.

Relevance is not the sole determinant of how well an item performs. Rather, the items performed much like classical test theory expects them to perform. The BS items proved to be powerful discriminators because they approached the ideal level of difficulty, having a mean p value of .501 (see Table 10, p. 67). The issue of relevance, if it is an influence, simply in the face of the more potent factor of difficulty which, in turn, influences favorably dispersion, reliability, and, ultimately, discriminating capacity.

This is not to say that inferior or irrelevant items might be acceptable. Neither this study nor Downing's utilized low quality items or items conceptually unrelated to emergency medical practice. All items were carefully drafted, edited, revised, and eventually approved by subject-matter specialists. Items considered LCR were not intended to be lacking in clinical relevance but were identified after the fact based on statistical analyses. Even the LCR items were thought to be relevant, and are still considered

relevant by exemplary practitioners of emergency medicine. The results of this study indicate little can be gained from quibbling about whether items are statistically relevant, irrelevant, partially relevant, etc. Good items are those having sound, widely recognized and accepted, psychometric properties.

DIFFICULTY OF BASIC SCIENCE

Finding that physicians don't score very high on basic science tests is not without precedent. In an effort to curb the growing numbers of cultists practicing health care and to protect the public against quackery, medical societies began lobbying states to adopt basic science acts. The first such statute was enacted in Wisconsin in 1925. Alabama in 1960 became the twenty-fourth and last state (District of Columbia included) to adopt a basic science act. Michigan's basic science statute was enacted and approved May 27, 1937 and became effective October 29 of that year. Like the other states, it was fashioned after Wisconsin's and required those "desiring to practice healing" to be of good moral character, a high school graduate or equivalent, and earn a grade of at least 75% in each of seven basic science subjects (anatomy, physiology, pathology, bacteriology, hygiene and public health, and chemistry).

Although the states which had a basic science law generally felt the acts were effective in controlling the problem of the incompetents, the acts were not without problems and controversy. Most often heard criticisms included:

1. There was no consensus among the twenty four states as to what constituted basic science. Each state board defined basic science operationally to include whatever disciplines it considered appropriate, and then tested over those disciplines.

2. No uniformity or inherent fairness existed regarding who should be exempted from the act's requirements and how grandfather provisions should be applied.
3. It was not always clear what constituted passing performance by the individual states. For example, in several states a candidate might, in one sitting, pass some subjects and fail others. Upon retaking the exam he may fail a subject(s) earlier passed. Over successive re-takes the candidate may have earned a passing grade in each subject.
4. There were no established reciprocity provisions among the states. This tended to thwart the travel and relocation of persons who would be qualified and licensed only in their home state.
5. The acts weren't fully effective; many non-traditionalists were able to pass the examination though often requiring repeated attempts.
6. Physicians had a high failure rate on the basic science examinations.

It is this last point that is most germane to the present results. Unfortunately, no definitive data regarding this sixth point exist, since state basic science boards made very little public and never revealed the identity of individuals or groups. Overall failure rates varied from state to state but ranged as high as 75%. In the only attempt to determine MD failure rate, one researcher, through a laborious effort, reported that the MD failure rate in New Mexico from 1961 through 1966 was 64.9%.

In response to this high failure rate, and the other criticisms listed above, states began repealing their basic science statutes. Florida, in 1967, was the first to do so. Michigan's basic science act was finally repealed in June of 1972 (P.A. 172, 1972), an earlier effort to do so having been successfully defeated by the state medical society lobby which argued "high

standards must be maintained" (Derbyshire, 1969, p. 132). Nation-wide the basic science acts were replaced much quicker than they were enacted. In the eight years following Florida's repeal, seventeen states replaced their basic science law. The remaining are believed to have since been repealed. Although present licensing boards can, and do, test over basic science, there are no longer any independent statutory requirements to do so.

IMPLICATIONS

The major finding of this study is that relevance is not the key attribute that determines how well multiple-choice examination questions function. These findings differ irreconcilably from Downing's in that these data clearly indicate that reasonably well written test items having p values around .5 will best discriminate among the examinees on the trait being tested. Classical test theory has long shown that items of mid-level difficulty are superior to all others. Finding the MD scale to be the superior discriminator is therefore predictable from this long line of writers. It then follows that the BS scale, which was made up of items having mid-level difficulty values would also be a more discriminating scale. The BS scale, though much more difficult than the HCR scales, was superior to it because the mean p value of the BS scale item was much closer to .5 than was the mean of value of the HCR item. It was the relative difficulty, not the absence of relevance, of the BS items that resulted in their superior ability to make discriminations.

This implication is consistent with other findings that were learned in developing and refining the Certification Examination. For example, earlier data made quite clear that it was not necessary to painstakingly determine the individual domains which comprised the phenomenon being examined and then

construct the test so that each domain was represented in direct proportion to its real-world relative importance.

Further work established that testing formats yield no differential effect. This finding which follows from the "no domain effect" finding is likely to be of great interest to test constructors and credentialling authorities. Quite simply, test developers can simplify their procedures and be less concerned with attending to details once thought important. Domains can be disregarded and items need only represent the general area, not the presumed components, being tested over. Expensive, cumbersome and difficult to produce formats are apparently not necessary and can be avoided in those situations where cognitive skills are being measured -- even where these measures are to be used to make inferences about a real-world correlate. Simple paper and pencil multiple-choice examinations have been shown to outperform more exotic multiple-choice items and testing formats. The present results extend this line of findings by indicating that the concept of item relevance is not decisive in constructing an examination or determining the quality of a developed examination. For non-psychometric reasons, however, the test should generally be content valid and appear to the examinees and other consumers to be facially valid.

Although the findings of this research are thought to be applicable to most subject-matters and levels thereof, to better know the limits of such generalizability we must await the results of future research.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abraham Flexner is dead at 92; Revolutionized medical schools. The New York Times, September 22, 1959, pp. 1,35.

Adams, Numa P.G. Sources of supply of Negro health personnel. Section A: Physicians. Journal of Negro Education, 1937, 6, 468-476.

Banta, H. David. Medical Education: Abraham Flexner - A reappraisal. Social Science and Medicine, 1971, 5, 655-661.

Berliner, Howard S. A larger perspective on the Flexner Report. International Journal of Health Services, 1975, 5, 573-592.

Berliner, Howard S. New light on the Flexner Report: Notes on the AMA-Carnegie Foundation background. Bulletin of the History of Medicine, 1977, 51, 603-609.

Bevan, Arthur A. Cooperation in medical education and medical service. Journal of the American Medical Association, 1928, 90, 1173-1177.

Bloom, Benjamin S. (Ed.). Taxonomy of Educational Objectives. New York, NY: David McKay Co., Inc., 1956.

Bordley, James, III, and Harvey, A. McGehee. Two Centuries of American Medicine, 1776-1976. Philadelphia, PA: W.B. Saunders Co., 1976.

Bowers, John Z. The influence of Edinburgh on American medicine. In Gordon McLachlan (Ed.), Medical Care and Medical Education. London: Oxford University Press, 1977.

Buros, Oscar K. (Ed.) Personality Tests and Reviews. Highland Park, NJ: Gryphon Press, 1970.

Cangi, Ellen Corwin. Abraham Flexner's philanthropy: The full-time system in the department of surgery at the University of Cincinnati College of Medicine, 1910-1930. Bulletin of the History of Medicine, 1982, 56. 160-174.

Chapman, Carleton B. The Flexner Report. Daedalus, 1974, 103, 105-113.

Comroe, Julius H., Drabkin, David L., Ehrich, William E., Flick, John A., and Kety, Seymour, S. Integrated teaching in the basic sciences: Report on five years experience. Journal of the American Medical Association, 1951, 147, 1221-1223.

Cook, Desmond L. Relevance categories and item statistics. In The Sixteenth Yearbook of the National Council on Measurements Used in Education. New York, NY: National Council on Measurement Used in Education, 1959.

Cook, Desmond L. A note on relevance categories and item statistics. Educational and Psychological Measurement, 1960, 20, 321-331.

Cronbach, Lee J. Essentials of Psychological Testing (3rd ed.). New York, NY: Harper and Row Publishers, 1970.

Cronbach, Lee J. and Gleser, Goldine C. Psychological Tests and Personnel Decisions (2nd ed.). Urbana, Ill: University of Illinois Press, 1965.

Cullen, Thomas J., Dohner, Charles W., and Schwarz, Roy. Evaluating decentralized basic science medical education: A model. Evaluation and the Health Professions, 1981, 4, 407-417.

Cullen, Thomas J., Dohner, Charles W., Schwarz, Roy, and Striker, Gary E. The development and use of a comprehensive test for evaluating decentralized medical education---The WAMI experience. In Proceedings of the Sixteenth Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1977.

Cullen, Thomas J., Dohner, Charles W., Striker, Gary E., and Schwarz, Roy. Evaluating student performance in a decentralized basic science program. Journal of Medical Education, 1976, 51, 473-477.

Cureton, Edward E. Validity. In E.F. Lindquist (Ed.), Educational Measurement. Washington, DC: American Council on Education, 1951.

Curti, Merle. The historical scholarship of Richard H. Shryock. Journal of the History of Medicine, 1974, 29, 7-14.

Derbyshire, Robert C. Medical Licensure and Discipline in the United States. Baltimore, MD: Johns Hopkins Press, 1969.

Dickman, Robert L., Sarnacki, Randolph E., Schimpfhauser, Frank T., and Katz, Leonard A. Medical students from natural science and nonscience undergraduate backgrounds. Journal of the American Medical Association, 1980, 243, 2506-2509.

Dobbs, Gideon S. Adventures in medical education. Journal of Medical Education, 1957, 32, 781-794.

Downing, Steven M. An analysis of the effects of different multiple-choice item selection strategies on the reliability and validity of measures on physician competence in specialty certification. Unpublished doctoral dissertation, Michigan State University, 1979.

Downing, Steven M. The assessment of clinical competence on the Emergency Medicine Specialty Certification Examination: The validity of clinically relevant multiple-choice items. Annals of Emergency Medicine, 1980, 9, 554-556.

DuBois, Arthur B., Nemir, Paul, Jr., Schumacher, Charles F., and Hubbard, John P. Graduate medical education in basic sciences. Journal of Medical Education, 1969, 44, 1035-1043.

Ebel, Robert L. How an examination service helps college teachers to give better tests. In Proceedings of the 1953 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1954.

Ebel, Robert L. Essentials of Educational Measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979.

Engel, George L. Biomedicine's failure to achieve Flexnerian standards of education. Journal of Medical Education, 1978, 53, 387-392.

Essex, Diane L, and Sorlie, W.E. Effectiveness of instructional computers in teaching basic medical sciences. Medical Education, 1979, 13, 189-193.

Essex, Diane L., and Sorlie, William E. Curriculum bias: A study of a college-certifying system for students at two basic medical science schools. Evaluation and the Health Professions, 1982, 5, 229-238.

Flexner, Abraham. Medical Education in the United States and Canada. New York, NY: The Carnegie Foundation, 1910.

Flexner, Abraham. The usefulness of useless knowledge. Harper's, October, 1939, pp. 544-550.

Floden, Robert E. Flexner, accreditation, and evaluation. Educational Evaluation and Policy Analysis, 1980, 2, 35-46.

Fox, Daniel M. Abraham Flexner's unpublished report: Foundations and medical education, 1909-1928. Bulletin of the History of Medicine, 1980, 54, 475-496.

Frishauf, Peter. The AAMC story: The rise to power. The New Physician, 1974, 23, 36-39.

Garrard, Judith, and Weber, Richard G. Comparison of three- and four-year medical school graduates. Journal of Medical Education, 1974, 49, 547-553.

Girdwood, Ronald H. Edinburgh in the history of medicine. In Gordon McMachlan (Ed.), Medical Care and Medical Education. London: Oxford University Press, 1977.

Glass, Gene V., and Stanley, Julian C. Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice Hall, Inc., 1970.

Goldberg, Jacob A. Jews in the medical profession-A national survey. Jewish Social Studies, 1939, 1, 327-336.

Goslings, W.R.O. Leiden and Edinburgh: The seed, the soil, and the climate. In The Early Years of the Edinburgh Medical School. Edinburgh, Scotland: Royal Scottish Museum Publication, 1976.

Gough, Harrison G. Some predictive implications of premedical scientific competence and preferences. Journal of Medical Education, 1978, 53, 291-300.

Guilford, J.P. Fundamental Statistics in Psychology and Education (3rd ed.). New York, NY: McGraw-Hill Book Co., 1956.

Guilford, J.P. Fundamental Statistics in Psychology and Education. (4th ed.). New York, NY: McGraw-Hill Book Co., 1965.

Gulliksen, Harold O. Theory of Mental Tests. New York, NY: John Wiley and Sons, Inc., 1950.

Guthrie, Douglas. The influence of the Leyden school upon Scottish medicine. Medical History, 1959, 3, 108-122.

Guyer, Kenneth E., Jr., Poland, James L., and Seibel, Hugo R. Trends in anatomy, biochemistry, and physiology lab programs in medical education. In Proceedings of the Thirteenth Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1974.

Harrington, Thomas F. The Harvard Medical School: A history, narrative, and documentary. (Vols. 1-3), Chicago, Ill: Lewis Publishing Co., 1905.

Hays, William L. Statistics for the Social Sciences (2nd ed.). New York, NY: Holt, Rinehart and Winston, Inc., 1973.

Hays, William L. Statistics (3rd ed.). New York, NY: Holt, Rinehart and Winston, Inc., 1981.

Hubbard, John P. Measuring Medical Education (2nd ed.). Philadelphia, PA: Lea and Febiger, 1978.

Jarcho, Saul. Medical education in the United States--1910-1956. Journal of the Mount Sinai Hospital, 1959, 26, 339-385.

Jason, Hilliard. The relevance of medical education to medical practice. Journal of the American Medical Association, 1970, 212, 2092-2095.

Kendall, Patricia L. Clinical teachers' views of the basic science curriculum. Journal of Medical Education, 1960, 35, 148-157.

Kennedy, William B., Kelley, Paul R., Jr., and Saffran, Murray. Use of NBME examinations to assess retention of basic science knowledge. Journal of Medical Education, 1981, 56, 167-173.

Kessel, Reuben A. The A.M.A. and the supply of physicians. Law and Contemporary Porblems, 1970, 35, 267-283.

King, Lester S. The Flexner Report of 1910. Journal of the American Medical Association, 1984, 251, 1079-1086.

Kirk, Roger E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Wadsworth Publishing Co., 1968.

Levy, Maurice, Bresnick, Edward, and Williams, W. Loren, Jr. A student feedback model designed to elicit data for effective curricular modification in the basic sciences. In Proceedings of the Eleventh Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1972.

Lord, Frederic M., and Novick, Melvin R. Statistical Theory of Mental Test Scores. Reading, Mass.: Addison-Wesley Publishing Co., 1968.

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

Maatsch, Jack L., 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, DC: Association of American Medical Colleges, 1978.

Maatsch, Jack L., and Elstein, Arthur S. Model for a Criterion-Referenced Medical Specialty Test: A Progress Report on Grant No. HS 02038. East Lansing, MI: Office of Medical Education Research and Development, Michigan State University, February 15, 1979.

Maatsch, Jack L., et al. Predictive Validity of Medical Specialty Examinations: A Final Report of Grant Number HS 02038-04. East Lansing, MI: Office of Medical Education Research and Development, Michigan State University, 1983.

Machotka, Pavel, Ott, John E., Moon, John B., and Silver, Henry K. Competence of child health associates: Comparison of basic science and clinical pediatric knowledge with that of medical students and pediatric residents. In Proceedings of the Tenth Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1971.

Magnusson, David. Test Theory. Reading, Mass.: Addison-Wesley Publishing Co., 1967.

Markowitz, Gerald E., and Rosner, David Karl. Doctors in crisis: A study of the use of medical education reform to establish modern professional elitism in medicine. American Quarterly, 1973, 25, 83-107.

McNemar, Quinn. Psychological Statistics (3rd ed.). New York, NY: John Wiley and Sons, Inc., 1962.

Medical Education in the United States. Journal of the American Medical Association. 1912, 59, 650-654.

Munger, Donna Bingham. Robert Brookings and the Flexner Report: A case study of the reorganization of medical education. Journal of the History of Medicine, 1968, 23, 356-371.

Nevins, Allan. The State Universities and Democracy. Urbana, Ill: University of Illinois Press, 1962.

Nie, Norman H., et al. SPSS: Statistical Package for the Social Sciences (2nd ed.). New York, NY: McGraw-Hill Book Co., 1970.

Nunnally, Jum C. Psychometric Theory. New York, NY: McGraw-Hill Book Co., 1967.

Nunnally, Jum C. Psychometric Theory (2nd ed.). New York, NY: McGraw-Hill Book Co., 1978.

Packard, Francis R. History of Medicine in the United States. New York, NY: P.B. Harber, Inc., 1932.

Pellegrino, Edmund D. Pruning an old root: Premedical science and medical school. Journal of the American Medical Association, 1980, 243, 2518-2519.

Remmers, Hermann H., and Gage, Nathaniel L. Educational Measurement and Evaluation. New York, NY: Harper and Row, Publishers, 1955.

Schudson, Michael. The Flexner Report and the Reed Report: Notes on the history of professional education in the United States. Social Science Quarterly, 1974, 55, 347-361.

Shultz, Douglas G. The relationship between scores on the science test of the medical college admission test and amount of training in biology, chemistry, and physics. Educational and Psychological Measurement, 1951, 11, 138-150.

Shryock, Richard H. American indifference to basic science during the nineteenth century. Archives Internationales d'Histoire des Sciences, 1948, 5, 50-65.

Shryock, Richard H. Women in American medicine. Journal of the American Medical Women's Association, 1950, 5, 371-379.

Shryock, Richard H. The influence of the Johns Hopkins University on American medical education. Journal of Medical Education, 1956, 31, 226-235.

Sorlie, William E., Anderson, John D., Gamble, Thomas E., and Bloomfield, Daniel K. A one-year program in basic medical sciences. Journal of Medical Education, 1973, 48, 371-373.

Sorlie, William E., Bloomfield, Daniel K., Anderson, John D., and Gamble, Thomas E. Medical basic sciences---A one year independent study experimental approach. In Proceedings of the Eleventh Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 1972.

SPSS* User's Guide. Chicago, Ill: SPSS Inc., 1986.

Stanley, Julian C., and Bolton, Dale L. Book reviews of Bloom's Taxonomy of Educational Objectives and Gerberich's Specimen Objective Test Items. Educational and Psychological Measurement, 1957, 17, 631-634.

Starr, Paul. The Social Transformation of American Medicine. New York, NY: Basic Books, Inc., 1982.

Thomae-Forgues, Maria, and Erdmann, James B. MCAT scores and academic records of natural science and humanities majors applying to medical school, 1978-79. Journal of Medical Education, 1980, 55, 971-972.

Webster, George D. What is the challenge? In Conference on Extending the Validity of Certification. Chicago, Ill: American Board of Medical Specialties, 1976.

Weiss, S.T., and Samet, J.M. An assessment of physicians' knowledge of epidemiology and biostatistics. Clinical Research, 1979, 27, 571A.

Whipple, George H. Mr. Abraham Flexner. Journal of Medical Education, 1960, 35, 451-453.

Williamson, John W. Valildation by performance measures. In Conference on Extending the Validity of Certification. Chicago, Ill: American Board of Medical Specialties, 1976.

Yens, David P., and Stimmel, Barry. Science versus nonscience undergraduate studies for medical school: A study of nine classes. Journal of Medical Education, 1982, 57, 429-435.

Zeleznik, Carter, Hojat, Mohammadreza, and Veloski, Jon. Baccalaureate preparation for medical school: Does type of degree make a difference? Journal of Medical Education, 1983, 58, 26-33.

Appendix A

Instructions which prefaced Booklets 1, 2, and 3.

EMERGENCY MEDICINE SPECIALTY CERTIFICATION EXAMINATION

MULTIPLE CHOICE

GENERAL DIRECTIONS

This part consists of approximately 90 Multiple-Choice Questions. Read each item carefully and choose the one most correct or best answer. It is to your advantage to answer each question, since your score will be the sum of the number of items corectly answered. Some experimental items, which may not count toward your score, are included in this part. You should, however, attempt to answer each question as best you can.

> Mark all of your answers on the separate answer sheet using the pencil you were given. Blacken completely the space on the answer sheet wich corresponds to the letter of the answer you have chosen. If you change an answer, please be sure to erase completely before marking your new answer. Marking more than one answer will cause the item to be scored as incorrect.

Before beginning this Examination, complete your answer sheet as follows:

1. Sign your name and your booklet number on the line marked "Signature" at the top of the answer sheet.

2. Complete the Name and Candidate Number grids. Be sure to blacken the spaces corresponding to each letter and number completely.

3. Print your name, location (Los Angeles, Chicago, or Philadelphia) and circle the Form (1...5). These spaces are found along the left margin of the answer sheet.

You will have 1 hour and 20 minutes for this part of the Multiple-Choice Examination.

EXAMPLE QUESTION

What is the most common cause of massive rectal bleeding in the adult?

A Polyps
B Diverticulosis
C Mesenteric Thrombosis
D Carcinoma of the right colon
E Carcinoma of the left colon

Since B is the best or most correct answer to this item, you would darken completely the space labeled B on the answer sheet for this question.

STOP: DO NOT BREAK THE SEAL UNTIL INSTRUCTED TO DO SO.

Appendix B

Instructions which prefaced Booklets 4 and 5.

EMERGENCY MEDICINE SPECIALTY CERTIFICATION EXAMINATION

PICTORIAL MULTIPLE-CHOICE

GENERAL DIRECTIONS

This part consists of Pictorial Multiple Choice Questions. Look at the visual material(s) that goes with each question, read each item carefully and choose the one most correct or best answer. It is to your advantage to answer each question, since your score will be the sum of the number of items correctly answered.

SPECIAL INSTRUCTIONS

Examination questions and the visual material(s) that go with these questions are contained in this booklet. The visual material(s) are presented on the left side of the page and the questions that go with these visual materials are on the right side of the page.

Mark all of your answers on the separate answer sheet using the pencil you were given. Blacken completely the space on the answer sheet which corresponds to the letter of the answer you have chosen. If you change an answer, please be sure to erase completely before marking your new answer. Marking more than one answer will cause the item to be scored as incorrect.

Before beginning this Examination, please write you name and address on the lines provided on the side of the answer sheet. Then, mark your Name and Candidate Identification Number in the grids at the top of the answer sheet. Be sure to blacken the spaces corresponding to the letters and the numbers completely.

You will have approximately 1 hour for this section of the Examination.

GO ON TO THE NEXT PAGE FOR THE EXAMPLE

EXAMPLE

Look at the x-ray on the opposite page. Now, read the following examination question, choose the best or most correct answer and record your answer on the separate answer sheet.

How many bones are fractured on this hand x-ray?

A one

B two

C three

D four

E five

Since the correct answer to the example question is A, you would darken completely the space labeled A on the answer sheet for this question.

STOP: DO NOT TURN THE PAGE UNTIL INSTRUCTED

Photograph of x-ray appeared on facing page.