12-2002

A Test of Factor Analysis as a Validation Procedure for Physician Certification Examinations

Jeffrey D. Greene
Western Michigan University

A TEST OF FACTOR ANALYSIS AS A VALIDATION PROCEDURE FOR PHYSICIAN CERTIFICATION EXAMINATIONS

by

Jeffrey D. Greene

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Educational Studies

Western Michigan University
Kalamazoo, Michigan
December 2002

A TEST OF FACTOR ANALYSIS AS A VALIDATION PROCEDURE FOR PHYSICIAN CERTIFICATION EXAMINATIONS

Jeffrey D. Greene, Ph.D.

Western Michigan University, 2002

Two physician certification examinations from different medical specialties were investigated. The purpose of the study was twofold: 1) to determine the similarities between the factor structure of the examinations and their respective tables of specifications; and 2) to demonstrate the relative efficacy of factor analysis in differentiating the structure between two related but dissimilar domains of information.

Specialty A is a homogeneous discipline focused on a relatively narrow concentration of organs, body systems and anatomy. This examination contained 309 items. There were 845 cases available for analysis. Specialty B is a heterogeneous area of specialty concerned with numerous areas of anatomy and physiology. The Specialty B examination contained 336 items and was completed by 1460 examinees.

The table of specifications for Specialty A called for six dimensions of content arrayed in relatively large areas ranging from 10% to 25% of the total examination length. The two largest areas in Specialty A accounted for half of the content. Conversely, Specialty B contained 22 areas ranging from 1% to 11% of the total content. Its two largest areas represented only 22% of the total examination content.

A principal components analysis with varimax rotation was conducted on both examinations. The results of the study showed that neither obtained structure revealed any dimensions that approximated the elements of the respective tables of specifications. In both examinations the number of viable obtained factors was less than the number of factors expected by the researcher and less than the number derived by the Minimum Average Partial procedure. The two viable factors derived from the homogeneous discipline (Specialty A), diagnostic skills and treatment choices, accounted for about 5% of the variance. Analysis of the heterogeneous specialty (Specialty B) returned three viable factors (diagnosis and treatment, internal medicine, and symptom recognition), explaining about 9% of the variance. Because the percentage of variance explained falls below a reasonable threshold, these results are considered to be inconclusive. It was not possible to make a definitive statement about the comparative structure of the two disciplines.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGEMENTS

Although I never had the honor of meeting him, I wish to acknowledge the influence of the late Dr. Samuel Messick for his remarkable thinking and writing on the topic of validity in measurement. His words have inspired and guided me in this dissertation and in my work.

Secondly, I am deeply indebted to the chairperson of my dissertation committee, Dr. Mary Anne Bunda. It is primarily through her urging and judicious use of a thundering velvet hand that I have been able to craft a cohesive paper that is worthy of defending. I also wish to thank the other members of my committee, Dr. James Sanders and Dr. Diane Hamilton, for the effort they expended in providing a thoughtful review of my work. David Overton, M.D., E. Dennis Lyne, M.D., and Joseph Chess, M.D. also provided enthusiastic and invaluable assistance that was much appreciated.

Lastly, I want to recognize the role that my family has played in my being able to complete this degree. Thanks and love to Madelyne and Mackenzie, my daughters, for their patience with my episodic availability to them over the past five years. Most importantly I would like to thank my wife, Penni, whose enduring support and tolerance allowed me to avoid the economic and emotional fate of many less fortunate graduate students.

Jeffrey D. Greene

TABLE OF CONTENTS

Table of Contents—Continued

CHAPTER

Table of Contents—Continued

Table of Contents—Continued

APPENDICES

LIST OF TABLES

List of Tables—Continued

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The Purpose of Competency Testing for Physicians

Competency testing holds a crucial status in the process and practice of medical education. Depending on the particular discipline, a physician has taken and passed at least eight national examinations of aptitude and achievement between application to medical school and certification by a specialty board. Presumptively, passing examinations of this nature provides the evidence that the examinee has mastered a particular domain of knowledge (Jaeger, 1989). In the case of physicians, such documentation is particularly important in that it provides a basis for identifying practitioners who possess the knowledge and skill to handle certain medical conditions and problems – a distinction that we must be able to make in terms of serving the public and societal good (Raymond, 1995).

Paramount in the sequence of examinations for physicians is the specialty certification examination. "Board exams," as they are known in the medical community, are developed and administered by each of the 23 respective boards of directors in the individual areas of medical practice recognized by the American Board of Medical Specialties (see Appendix A for complete listing of recognized specialties). The examinations are intended to assess a physician's knowledge of information that has been determined to be relevant to his/her field of specialization and essential for effective practice in the specialty discipline. Passing the examination implies that the candidate has at least the amount of knowledge and skill necessary to certify their competence. Although the subject matter of these examinations is concerned with knowledge of constructs that are parts of very complex and intricate domains, ultimately we must expect and anticipate that the instruments employed in the determination of specialty certification possess the psychometric properties that substantiate the implicit trust we have placed in the notion of a physician being "board certified."

From the physician's perspective, being board certified (or not) also has a number of implications. For example, in some hospitals physicians who are not certified in their specialty are not awarded admitting privileges, seriously curtailing their ability to practice. Additionally, some third-party payers will not reimburse physicians for services rendered or procedures that have been performed if they have not attained the endorsement of their professional discipline (Kane, 1986). In academic medicine, accrediting bodies can deny recognition to residency programs that have faculty members who are not board certified (ACGME, 1997). To be sure, the attainment of specialty certification is a high-stakes proposition that has important implications for both patients and physicians.

The Scope and Complexity of Validity

As with any worthy exam, the central property of interest for certification testing is validity – the extent to which empirical evidence and theoretical rationale support the inferences and actions that are based on a test score or other manner of assessment (Messick, 1989). This elegant statement has guided the thinking of many leading psychometricians for more than a decade, but the grace of the words belies the intricacy of the elements subsumed under the task of validation.

Although it is now thought of as a unitary concept, there is a matrix of considerations – psychometric, political, social, and theoretical – that represent the complexity of validity. From the realm of psychometrics, contemporary ideas about validity have been dominated by the "Trinitarian Doctrine" of content-, construct-, and criterion-related evidence (Shepard, 1992). While these aspects of validity can be discussed individually, it is only for expediency. Far from being mutually exclusive, the three components are interrelated logically and pragmatically and it takes all three types of evidence to make a cogent validity argument. Choices about the relative emphasis on any of the types are determined by the nature of the inference that is to be drawn from the examination.

In addition to the inherent complexity of having these three integrated and varying characteristics at its foundation, establishing validity for any particular interpretation is not a dichotomous proposition. That is to say, it is not all or nothing; rather, validity is a matter of degree. Moreover, the validity of a score interpretation is not an immutable condition. Evidence can be augmented or breached by new findings, making validity an evolving property.

The political dimension of validity derives from the fact that the uses of tests – especially those of a high stakes variety – may concern contending values that are susceptible to political vicissitudes. As a result, validity evidence can be viewed with a lens having variable focal points, depending on the end condition being advocated.

Validity is also a consequential issue. Whether or not the interpretation of a test score represents the meaning for which it was intended is of great concern, but of equal interest is whether the interpretation results in unintended consequences or outcomes. Intentional inquiry into the possible side effects of a particular interpretation, then, helps to answer the question of whether we *should* use a score for a particular purpose. Irrespective of how well the evidence supports the intended interpretation of a score, proper execution of the validation effort demands that we consider any consequences that lay outside of the traditional content-criterion-construct chain.

Another factor in the validity matrix relates to the fit between the proposed score interpretation and the theory of the domain being measured. True validation calls for a rational linkage between the empirical tenets and grounding of a construct and a particular interpretation that is made of its purported measure. For example, because it is possible to conduct correlational studies on any two constructs, one can assemble a great amount of quantitative evidence that does not necessarily yield any greater understanding about the validity of a score interpretation. Thus, authentic validity evidence has a clear relationship with the theory of the construct under study.

To sum up, the importance of establishing a compelling validity argument is overshadowed only by the complexity of the factors that must be considered in doing so. Validity evolves on a continuum; it is buffeted by political agendas and the value preferences of contending constituencies; it extends beyond the mere collection of empirical support to the consideration of the consequences of a particular interpretation; and it rests on the proper weighing of evidence in the realms of content, criterion, and construct validity.

It is this last element of the validity matrix that is the province of the current study. We know that for maximum clarity, strength, and influence, the argument for validity should provide evidence in a variety of forms in support of a given score interpretation (Messick, 1989; Shepard, 1992). Any of the three "types" of validity evidence – whether derived by correlational methods, through the rendering of expert judgment, or support demonstrated through experimental procedures – can be useful in strengthening a test score interpretation. As we have seen, the issue is not one of choosing between separate and distinct types of evidence; all are needed for a trustworthy score interpretation. Instead the focus is one of *emphasis* on one or more of these approaches.

Content Validity in Testing

For the present work, the concentration is on the domain of content validity. On tests of achievement within a relatively bounded domain, validation of content may be one of the more relevant factors to consider (Nunnally & Bernstein, 1994; Mehrens & Lehmann, 1991).

Most often, content validity is the primary consideration in establishing the representativeness of test content for a specified domain (Crocker & Algina, 1986; Messick, 1989). Inasmuch as achievement test behavior serves as a sample of one's attainment within a universe of subject matter, content validity is critical to insuring whether the test items collectively constitute a representative sample of the domain. LaDuca (1994) supported this perspective when he wrote that in a high stakes, domain-referenced examination that has important consequential characteristics, developing validity evidence in the realm of content is paramount. In addition, Kane (1982) drew similar conclusions when he stated that in the case of certification exams, the desired interpretation of scores is that they provide evidence of the examinee's competence on specific knowledge that is needed for effective practice in the occupation. Therefore, an interpretation of the results of the examination should be made in terms of the presence or absence of particular constructs of knowledge. This approach suggests the prominence of content as a point of inquiry for validation activities.

Similar to the approach of the present work, Kane's and LaDuca's content-based approach to validation is separate and distinct from other perspectives about validation of physician competency. While other points of view on the structure of competency examinations – such as investigating the relationship between examination score and effectiveness of clinical practice, studying the validity or methods of how content domain is defined, or researching the determination of cut scores – may be worthy of further inquiry, such issues are outside the parameters of the current study. Most certainly, these matters are critical in the testing of competence and deserve attention from researchers, but they approach topics and questions that transcend the focus here.

The Table of Specifications as Representation of the Content Domain

Understanding that content validity is a critical dimension of domain-referenced examinations, it is also important to specify the topics and processes that are to be inclusive of the domain. Most often, of course, the structure of a test is developed based on a table of specifications or "blueprint" (Dills, 1998; Mehrens & Lehmann, 1991). More specific to the present issue here, the content outline for a test of mastery is expected to provide a very specific description of the domain to be covered and represents the pool of judgments about what should have been learned (Thorndike, 1982). Usually, development of those judgments falls within the purview of subject matter experts in the discipline (Nelson, 1994). The test specifications and its structure are ultimately informed by the responses of a large representative sample of practitioners (Raymond, 1995). A blueprint can be thought of as providing the expert-sanctioned enumeration of the content, topics, and processes to be assessed on a domain-referenced examination and the content categories of the table of specifications for a competency examination may stand as the most definitive reflection of a profession's domains (Thorndike, 1982).

For physician certification exams, the blueprint and the items that are developed to reveal the details of its content are created by physicians themselves. Most often, these individuals are part of the discipline of academic medicine for their specialty and/or are long-time practitioners (American Board of Medical Specialties, 1999). These physicians serve as the authorities in their respective fields and their input into the examination blueprint and its items can be thought of as how the discipline is conceived by "experts."

Methods of Validating a Table of Specifications

A piece of a comprehensive validity argument is the development of evidence linking the specifications of the examination to the structure of the items that constitute it. More to the point, in the realm of minimum competency examinations the primary function of test specifications is to enhance the validity of test score inferences (Millman & Greene, 1989).

Among the methods that might be considered in the task of validation, factor analysis is intimately associated with questions of validity (Thompson & Daniel, 1996). Guilford (1946) described the relationship between the two when he wrote, "The factorial validity of a test is given by its loadings in meaningful, common reference factors. This is the kind of validity that is meant when the question is asked, 'Does this test measure what it is supposed to measure?'" (p. 429). For the purposes of the present work there is support for the notion that with an appropriate model, factor analysis can play a valuable role in the validation of the structure of a discipline (Schoenfeldt, 1984).

Given that the table of specifications for a test defines the domain of knowledge to be assessed, a properly derived factor solution can serve to help identify the structure of that domain as conceived by its expert practitioners. Together, the blueprint and the results of a factor study provide the opportunity to identify some of the evidence that can be used to support the validation of selected physician competency examinations.

Statement of the Problem

Questions about examination structure appeal to the technique of factor analysis. But given the wide-ranging content domains of medical specialty practice, what is the efficacy of factor analysis in identifying such structures? Analysis of medical specialty examinations provides a fertile opportunity to address this question. Some medical specialties are concerned with relatively limited scopes of anatomy and physiology. For example, disciplines such as dermatology, neurology, orthopedic surgery, obstetrics and gynecology, urology, ophthalmology, and psychiatry seem to deal with fairly limited body systems and functions. In terms of structure, such disciplines might be considered to be "homogeneous" in that they focus on a single organ or a small collection of related organs or tissues. On the other hand, specialties such as internal medicine, pediatrics, emergency medicine, general surgery, family practice, pathology, and radiology often deal with a wide range of anatomy and physiology. These types of specialties might be considered to be more "heterogeneous" in that work in these disciplines is most often simultaneously concerned with multiple organs and body systems.

The current study attempts to take advantage of this delineation between types of specialties to assess the efficacy of factor analysis as a method of providing evidence to validate a table of specifications. Two questions will guide the study. To what extent do the results of a factor analysis represent the table of specifications of a physician competency examination? And, given the great variety of physician specialties and their respective domains of knowledge, another question arises. To what degree does factor analysis provide different characteristics of information when the data come from the analysis of a broad or heterogeneous field of practice versus a more tightly bounded, homogeneous type of specialty? In an effort to answer these questions, the current study employs factor analysis on two physician specialty examinations – Specialty A, a homogeneous discipline, and Specialty B, a heterogeneous field of practice. The primary elements of data used in the analysis will be congruence between the factor structure of each examination and their respective tables of specifications and the percentage of variance explained by the analysis procedure.

The answers to these questions will provide information about the extent to which there is factorial evidence to support the interpretation we make of specialty certification scores. The findings will also reveal something about the power of factor analysis as a technique for identifying the structure of examinations when the subject matter under analysis is related (i.e., medical) but differential in scope.

CHAPTER II

REVIEW OF PERTINENT LITERATURE

The purpose of the current study is twofold: 1) to contrast the extent to which the results of a factor analysis represent the table of specifications for two different physician competency examinations; and 2) to investigate the efficacy of factor analysis as a method of identifying structure in two different physician certification examinations when one examination is homogeneous and the other is more heterogeneous. This chapter is intended to provide context and background for the study related to the nature of validity, along with a discussion of current practice in the development of validity evidence in the realm of medical education examinations. This section also provides a brief discussion of the method of factor analysis along with technical issues to be considered in employing this technique in studies of validity.

An Overview of Validity

Validity is the most important consideration of the measurement enterprise (AERA, APA, NCME, 1999). The entire premise for the utility of testing is based on the extent to which we can make meaning of the results obtained from an examination instrument. The process of meaning making occurs as the result of validation – the development of a sound argument resulting from scientific undertakings that integrate various strands of evidence – experimental, statistical, and philosophical – against a theory (or theories) of interpretation (Moss, 1992). An inference about the results of an examination without such a fund of evidence is untenable. Schoenfeldt (1984) put this issue even more succinctly by noting that a testing program that does not involve concerns about validity and validation research "is at best an unknown and at worst may be an outright fraud" (p. 61).

Messick (1989) is often credited with providing the contemporary perspective on validity. He wrote, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (p. 13, emphasis in original). Messick's work emphasizes that the attribute of validity is reserved for how we interpret test results – it is not applied to a testing instrument itself. Put another way, a testing instrument cannot be valid (or invalid, for that matter); only the meaning we ascribe to the results can carry the property of validity.

As with any other examination, the central issue for physician certification examinations is validity. That is, what can we say with good reason about candidates for certification based on the scores they obtain on the certification examination? Our basic assumption may be that those who pass are more likely to be safe and effective practitioners than those who fail (Kane, 1986). This interpretation, or any other, must be grounded in the validity evidence of the examination.

Contrary to our current conception of validity as encompassing a variety of elements or types of evidence, historical perspectives on the issue arose from mostly operational terms. In the middle part of the 20th century, validity was seen as simply being the correlation of scores on a test "with some other objective measure of that which the test is used to measure" (Angoff, 1988, p. 22). Guilford's (1946) writing was influential in this regard. He wrote that generally, "a test is valid for anything with which it correlates" (p. 428). Even Cureton's (1950) chapter in the inaugural edition of *Educational Measurement* defines validity as the correlation of observed scores on a test with the true scores on some criterion. Contemporary authors have also written about the practice of accepting a single correlation coefficient as evidence of validity in the early years of testing (Shepard, 1992; Schoenfeldt, 1984).

Of course, it is not now perceived to be effective measurement practice to allow a lone coefficient to stand as evidence of validity. Instead, as noted, validity arises from any number of sources of evidence. Most often, aspects of validity are categorized in terms of being associated with a criterion, a construct, or some domain of content (Messick, 1989; Shepard, 1992; Kane, 1986).

Validity Evidence Types in the Literature of Medical Education

In terms of the types of methods employed, the history of developing validity evidence in the realm of medical education examinations appears to be relatively narrow. The majority of studies reviewed for the present work related to validation in medical education seem to employ correlational techniques. Relatively few have used expert judgment, factor analysis, or other types of methods. Aside from a few ex post facto investigations, no studies of validity in medical education could be found that employed any type of experimental design. A review of some of the available studies will help the reader to get a sense of how medical academicians have approached the study of measurement in the field. Such a review will also reveal the types of things that are known about validity in examinations in medical education, as well as providing a perspective on the standing of validity studies in the profession of medicine.

It is also useful to recognize that most of the validity studies in medical education are concentrated in the area of undergraduate medical education (e.g., medical school). There is a smaller group of studies that were conducted during post-graduate education (residency). There is a dearth of studies concerning the assessment of validity for instruments administered to fully-licensed practitioners.

Correlational Designs

As noted, the validity literature in medical education appears to be dominated by correlational studies. Most often, these investigations are focused on predicting the academic or clinical performance of medical students and residents. The Medical College Admission Test (MCAT), admissions assessments, and grade point averages are often used as the predictor variables. Measures of clinical competency – such as clerkship examination scores, results from standardized examinations and faculty ratings – are typically used as criterion variables in these types of studies. Table 1 presents the listing correlational studies in medical education obtained from a search of Medline for the period of 1999 to 2002. The articles cited are not intended to illustrate exemplary or seminal work in the field. Rather, they are provided to suggest some impression about the nature and results of contemporary correlational studies that have been conducted.

Overall, it can be supported that the use of correlational methods has dominated the landscape of validity studies in medical education. Most frequently, GPA, MCAT scores, and data from various licensing examinations have served as the variables under study. Most studies of this nature show moderate correlations between variables.

Other Study Designs

There have also been other distinct methods of developing validity evidence discovered in the literature of medical education, but their frequency of use does not approach that of correlational designs. Mitchell & Molidor (1986) investigated the structure of the MCAT through the use of exploratory factor analysis. Analyzing 2,876 scores, their work revealed a two-factor solution (quantitative and communication) that accounted for 80% of the variance. In a similar study of 1,452 MCAT scores, Jones & Thomae-Forgues (1981) found a three-factor structure (science/quantitative, verbal ability, interpretation skills) that explained 78% of the variance. Employing linear regression, Blue, Gilbert, Elam and Basco (2000) studied 545 medical school applicants and found that the combination of MCAT scores,

Table 1

Studies in Current Literature of Medical Education
Using Correlational Designs to
Predict Performance

Predictor	Criterion	Size of Sample	*r*	Authors
Residency interview score	Faculty rating	660	.42	Char, et al. (2002)
Resident evaluation	In-training examination score	72	.35	Boudreaux, et al. (2002)
USMLE Step I	In-training examination score	56	.76	Ambroz & Chan (2002)
Year 3 GPA	USMLE Step I	306	.47	Fields, et al. (2000)
Application score	Faculty rating	222	.20	Basco, et al. (2000)
USMLE Step II	Core clerkship exam scores	2,158	.38	Callahan, et al. (2000)
MCAT science score	USMLE Step I	6,239	.34	Veloski, et al. (2000)
USMLE Step II	In-training examination score	52	.45	Bruno, et al. (1999)

medical school GPA, and selectivity of undergraduate institution contributed the most to the prediction of scores on medical licensing examinations ($r^2 = .38$).

More to the point of the current investigation, the author could find no studies in which clinical outcomes (the presumed dependent variable relative to physician certification) varied with the presence or absence of examination scores or physician certification. However, there was a single study that provided evidence that seems to support the existence of an association between examination scores and the frequency of competent practice activity. Tamblyn, et al. (1998) showed that increases in practice competency scores predicted increases in referral activity, dispensing of appropriate prescriptions, and use of preventative techniques.

To sum up, the preponderance of validity studies in the field of medical education are correlational in nature. Most of these studies were implemented in the population of medical students and were aimed at predicting academic or clinical success during undergraduate medical education. While there was one study that analyzed the structure of the MCAT, there no documented evidence of studies using factor analysis to explore the structure of physician certification examinations. An investigation in this area would appear to have a chance to make a contribution to the body of knowledge for the discipline.

An Overview of Factor Analysis

Factor analysis is the name given to a broad category of statistical procedures used for determining the existence of relationships among a group of measures (Nunnally & Bernstein, 1994). Mathematically speaking, using factor analysis it is possible to describe the total variance of a number of abilities in terms of a smaller number of independent components of variance (Jensen, 1998).

There are two broad applications of factor analysis – exploratory and confirmatory. In the former, the goal is to investigate the extent to which a large amount of information can be reduced to fewer main constructs or dimensions. In exploratory factor analysis (EFA), the researcher has no firm *a priori* theory or expectations about the composition of the derived factors that may evolve from the procedure. Instead, this type of analysis is employed in order to find evidence of hypothetical constructs that underlie the data. On the other hand, the purpose of confirmatory factor analysis is to test the hypotheses of whether the derived factor structure fits a theorized model (Kline, 1994). Inasmuch as the current study is concerned with the use of EFA, the discussion will not attend to the confirmatory aspect of factory analysis.

Exploratory factor analysis has often been used in the development and refinement of psychological and educational instruments and examinations. More specifically, the method has a long history of involvement with questions of validity. In fact, factor analysis is viewed primarily as a method for assessing the construct validity of measures, rather than as a means of data reduction (Thompson & Daniel

1996). There is support for construct validity if the obtained factor structure of a scale is consistent with the constructs that the instrument purportedly measures (Floyd & Widaman, 1995; Schoenfeldt, 1984).

The previous section presented some illustrations (e.g., Mitchell & Molidor, 1986; Jones and Thomae-Forgues, 1981) of the use of factor analysis in the validation of instruments from the field of medical education. In order to expand upon some important points about the technique, it is helpful to examine other studies to review the principles of how one interprets and uses data from a factor analytic study. Doing so will provide for easier understanding of the examples provided and the results of the present work.

Conventions in Reporting Factor Analytic Data

Exploratory factor analysis studies involving instruments having a variety of different purposes are presented in the subsequent section. Examinations and assessment instruments from a wide range of settings – clinical, educational, and psychosocial – are reviewed in order to illustrate several conventions related to reporting factor analytic data. A first common practice is to report the size of the factor loading for each item. This allows the reader to more clearly understand the magnitude of the correlation between the item and latent factor. Second, citing of the eigenvalues for each factor provides a sense of the weight the factor carries in the constellation of factors that emerge. Third, a description is given of the variance accounted for by each factor and by the total factor structure. This practice offers a sense of how much the observed structure explains about the phenomenon under study, and how much has yet to be explained by other, unknown, factors. Recognizing these conventions may help the reader interpret the findings presented here and in the results section of the present study.

Examples of the Use of Factor Analysis

There are numerous examples of the use of exploratory factor analysis in the literature related to the structure of affective and cognitive instruments. For example, a review of the 1999 index for *Educational and Psychological Measurement* reveals that of the 66 studies published, 16 of them were conducted using factor analysis as the method of inquiry. A few contemporary studies from educational, sociological, and clinical settings will be cited here to illustrate some applications of the procedure in the manner in which it will be used in the present work. Table 2 summarizes some characteristics from a sample of studies from the clinical, psychosocial research, and educational achievement settings.

In reviewing Table 2, there are some observations about factor analytic data that bear mention. First, the number of factors extracted is independent from the number of items in the instrument under study. In Instrument 3 (McCarthy & Archer, 1998), 478 items resolve to only two factors. On the other hand, Instrument 5 (Choi, Fuqua, & Griffin, 2001) consists of only 57 items, yet it produced seven factors.

Secondly, the amount of explained variance often corresponds to the congruence between the number of expected factors and the number of obtained

Table 2

Examples of Findings from Exploratory Factor Analysis Research

Instrument (α)	Number of Items	Number of Factors Expected/Obtained	Percentage of Variance	Obtained Factor Names
Clinical				
1, AMAS-E (,91)	46	3/3	85%	Worry, fear, physical
2, SASSI (,80)	29	3/2	53%	Alcohol, other drug
3, MMPI-A (,70)	478	2/2	55%	Maladjustment, externalizing
Psychosocial Research				
4, Work Group ID (NA)	17	3/2	69%	Mission, contribution
5, Self-Efficacy (,77)	57	9/7	59%	Learning efficacy, resistance to risk, support, social efficacy, physical efficacy, meeting others' expectations
6, Math Anxiety (,92)	10	2/2	61%	Worry, negative affect
Educational Achievement				
7, MCAT (NA)	221	2/2	80%	Quantitative, communication
8, MCAT (NA)	221	3/3	78%	Science/quantitative, verbal ability, interpretation skills

factors. For example, Instrument 1 (Lowe & Reynolds, 2000) was expected to reveal three factors, and, in fact, did, explaining 85% of the variance in the data. The data for Instrument 7 (Mitchell & Molidor, 1986) was similar, two factors were expected and two resolved, accounting for 80% of the variance. Conversely, a previous analysis of Instrument 7 (Jones & Thomae-Forgues, 1981) showed three factors that explained 78% of the variance. Analysis of Instrument 2 (Gray, 2001) resulted in only two factors when three were anticipated. In this case, only 53% of the variance was explained. Correspondingly, only 59% of the variance was explained for Instrument 7 (Choi, Fuqua, & Griffin, 2001) when seven of nine expected factors resolved during the analysis.

To summarize this section, researchers have been able to use exploratory factor analysis to identify the structure of instruments designed for the clinical, psychosocial, educational settings. These findings provided a contribution to the validity evidence for the respective instruments, although there was no documentation as to whether or not the findings supported the respective tables of specifications from the instruments. Additionally, review of these instruments illustrated some important points about some characteristics of factor analytic findings.

Introduction to Technical Issues in Factor Analysis

There are a variety of issues to which one must attend when considering the use of factor analysis. Which factor analytic procedure to use, how large the data sample must for a stable analysis, determining the number of factors to retain, how (or if) to rotate the derived structure, and interpreting the results are some of the primary concerns with which a factor analyst must be concerned.

Classes of Factor Analysis

Merenda (1997) identifies three essential classes or "types" of factor analysis procedures. These categories – common factor analysis, principal components analysis, and confirmatory factor analysis – are not substitutes for each other. Each serves a different purpose and differs in its underlying assumptions and mathematical procedures. For example, only confirmatory factor analysis tests a hypothesis. The other two classes are used for data reduction.

More to the point of the present work, principal components analysis (PCA) is often selected as the procedure of choice for work that is purely exploratory (Merenda, 1997). Thompson and Daniel (1996) cite two reasons for this practice. First, PCA produces factor scores that are equivalent to the correlation coefficients of the rotated factors. Thus, there is greater ease of interpretation of the derived factor solution. Second, principal components analysis has an advantage in that it does not unduly rely on sampling error as the price for estimating measurement error.

Technical advantages aside, there is a more pragmatic reason for using PCA in exploratory factor analysis. A number of researchers have discovered that principal components analysis and common factor analysis produce essentially equivalent results (Zwick & Velicer, 1986; Guadagnoli & Velicer, 1988). Thompson and Daniel (1996) found two conditions under which the two procedures yield similar findings. One such situation occurs when there are 30 or more variables or examination items under study. That is, the more variables, the more parallel the results between the two procedures. The second circumstance is related to the reliability of the factored variables; the greater the reliability, the more comparable the results.

Sample Size

A major issue in the literature relates to the size of the sample that is required to appropriately conduct a factor analytic study (Aleamoni, 1973). The importance of this issue is that small sample sizes are likely to lead to low component loadings and weak identification of factors. This is especially true when the small n is accompanied by overextraction (Merenda, 1997).

Traditional rules of thumb have held that a particular ratio of observations-to-variables or items (i.e., *n:p*) is required in order to ensure the stability of the factor patterns when using PCA. A variety of opinions have emerged on the inviolability of these rules. Lindeman, Merenda & Gold (1980) suggested that an *n:p* ratio of 20:1 is required. Thorndike's (1978) work supports reducing the ratio to 10:1. Still other authors (Velicer & Fava, 1987) recommended that the minimum ratio could be 3:1. Each of these findings, of course, calls for a different sample size for the conduct of a factor analysis study. In an analysis of a 25-item instrument, for example, following Lindeman et al. (1980) would require a sample size of at least 500. The formulas of Thorndike (1978) and Velicer and Fava (1987) would involve samples of 250 and 75, respectively. Judging from these findings, one could conclude that the issue of sample size in factor analysis remains a murky proposition.

More recent studies have suggested that factor stability as a function of a fixed ratio of observations-to-variables is not as critical is once believed (Knight, 2000). There do not appear to be any sound theoretical or empirical bases for enforcing broad recommendations of *n:p*. In their oft-cited study, Guadagnoli and Velicer (1988) wrote, "The results obtained in this study provide little support for current sample size rules in factor analysis. The most popular rules involved an *n:p* ratio and were clearly not substantiated. The concept that more observations are needed as the number of variables increases is clearly incorrect" (p. 271). Sample size as a function of the number of variables was not an important factor in determining the stability of factor structures.

In general, it can be supported that a large sample size is required for whatever type of factor analytic work one is utilizing (Knight, 2000). However, there is probably little to be gained in most situations by exceeding 1,000 observations in the sample (Shah, 1985).

Determining the Number of Factors to Retain

Another key consideration in the use of factor analysis is making a determination as to the number of factors to retain for interpretation. There are several rules or guidelines from which the factor analyst can choose, but there appear to be qualitative differences between them. Merenda (1997) and Zwick and Velicer

(1986) each studied commonly used methods for selecting the number of factors to retain.

The Kaiser Rule (eigenvalues > 1) was determined to be one of the most ineffective methods of factor selection. Seral authors have discovered that Kaiser tends to systematically overestimate the number of factors to be retained (Knight, 2000; Cliff, 1988; Streiner, 1994; Zwick & Velicer, 1986; Merenda, 1997). Universally, these authors do not recommend the use of the Kaiser Rule because of the inconsistency of the results it provides.

The Scree method is a graphical depiction that derives from the size of the increments between the eigenvalues of adjacent factors that have been extracted. The process of this method is to terminate extraction at the point at which the "elbow" occurs in the plot of successive eigenvalues. While the Scree method does rely to some extent on subjective judgment, it has been shown to produce reasonably reliable results.

Velicer's Minimum Average Partial (MAP) employs a matrix of partial correlations in order to determine an exact stopping point in the extraction process. In this procedure, the components that are extracted are those that are below the point at which the squared partial correlations reach a minimum.

The Horn Parallel Analysis (PA) is a sample-based adaptation of the Kaiser Rule. In PA, eigenvalues of a correlation matrix of *p* random uncorrelated variables (in which *p* is the number of variables) are derived from the data set. These values are then compared with those of the entire data set under study. Factors with eigenvalues greater than those in the randomly generated matrix are retained.

Merenda (1997) concluded that the Scree method, MAP, and PA, respectively, are among the most effective methods of extraction, and they are the ones that should receive first consideration in a factor analysis study. Zwick and Velicer (1986) made similar distinctions, ranking the efficacy of the methods in this order: PA, MAP, and Scree. However, the latter authors noted that the method of extraction selected will have a relatively inconsequential impact on the outcome of a factor analysis study because of the demonstrated robustness of results across the alternative methods. Attention on the matter of extraction method might be best placed on efficiency. That is, the determination is based on which method can be used with the least expenditure of money, effort, and time.

Rotation

Rotation of factor solutions in Euclidean space can serve to make factors more interpretable (Kline, 1994). Factors manipulated in this way are mathematically equivalent; they do not account for any more covariation among variables than the initial solution (Kim & Mueller, 1978a). At the same time, though, a rotated solution does change the relationship between the factors and the axes. This change results in different loadings, which may elicit simpler and more readily interpretable results.

There are two broad classes of rotation – orthogonal and oblique. Orthogonal rotation assumes that the derived factors are not correlated, while the oblique method presumes that the factors are related to each other (Nunnally & Bernstein, 1994). In the case of an exploratory analysis in which the objective is to reduce many variables to a smaller number of dimensions, orthogonal rotation is preferred (Kim & Mueller, 1978a). A primary reason for this is that imposing the property of independence on the factors makes the rotated solution easier to interpret. As the researcher's understanding of the phenomenon under study becomes clearer, it is possible to more meaningfully analyze the effects of other methods of rotation.

Interpretation

To be sure, factor analysis is an empirical procedure that requires knowledge about statistics and experimental analysis. However, irrespective of whether the research goal is exploratory or confirmatory, final decisions about the sufficiency of the derived factor structure is the responsibility of the researcher (Merenda, 1997). Thompson and Daniel (1996) note that factor analysis has great power to inform our judgment, but, ultimately, we are responsible for the judgement that is exercised. In other words, factor analysis has an important element of "art" integrated into its statistical methods. Interpretation of factor solutions must be based on underlying theoretical models, not simply on the outcome of statistical procedures.

It is also crucial to remember that the nomenclature of factor analysis may be misleading. The discovery of a "factor" means nothing more than certain relationships exist among the sampled responses. It does not mean that some entity exists outside of the tested behavior (Linn, 1979). It is important to recall these cautions as one considers using factor analytic data in the development of validity evidence.

To sum up this section, the practice of factor analysis calls for close attention to a variety of conventions and rules. Failure to attend to these practices can result in incorrect or misunderstood results. But at the same time the procedure is flexible, fairly robust, and offers a tremendous opportunity to conceptualize data from a useful and pragmatic perspective.

Summary

Validity is an exceptionally complex, but critical, aspect of any worthy examination. Historically, validity evidence has been presented in terms of its relationship to a construct, a criterion, or a domain of content. The realm of medical education has developed much of the validity evidence in its professional instruments by relying on correlational methods. Furthermore, many of the studies conducted in the field have been concerned with the aspect of undergraduate medical education. Comparatively few investigations have been conducted in the area of graduate medical education (residency training). Even fewer have been initiated in the domain of professional certification in medicine.

For its part, factor analysis has been a frequently used method in the development of validity evidence on examinations and other types of instruments. The prolific use of the procedure and its frequent appearance in the literature provides numerous opportunities for understanding the variety of technical rules and complementary processes that make factor analysis a powerful technique for helping to understand complex data.

The current study employs factor analysis on two physician specialty examinations – Specialty A, a homogeneous discipline, and Specialty B, a heterogeneous field of practice. The results of a factor analysis will be used to determine the congruence between the factor structure of each examination and their respective tables of specifications. Additionally, the data will also provide information about the power of factor analysis as a technique for identifying the structure of examinations when the subject matter under analysis is related (i.e., medical) but differential in nature (homogeneous discipline vs. heterogeneous discipline).

CHAPTER III

METHODS

The current chapter contains four sections intended to explain the methods used to conduct the present study on the efficacy of factor analysis. Included in the chapter are subsections on: (a) an introduction to the examinations under study; (b) the content of the examinations; (c) the characteristics of the examinations; and (d) a discussion of the analytic procedures used for the study.

Introduction to the Examinations

Two examinations, each from a different recognized medical specialty, were selected to undergo factor analysis. The specialties, identified only as Specialty A and Specialty B, were selected for two reasons. One basis for selection was based on the willingness of the respective boards to release item-level data for a recent cohort of examinees. The second basis was the researcher's intention to compare the factor structure of two medical specialties with markedly different (i.e., homogeneous and heterogeneous) content areas. Both of the examinations under study serve as the written test of knowledge for candidates who wish to obtain diplomat status (i.e., board certification) in their specialty area.

Specialty A

For Specialty A, the homogeneous discipline, the examination is the first of two elements of the certification process. The examination is designed to evaluate a candidate's knowledge of the specialty discipline, including basic science and his or her ability to use this information for problem solving in the diagnosis and treatment of patients. In order to qualify for the certification examination, candidates in Specialty A must supply verification of having completed residency education in a program accredited by the Accreditation Council for Graduate Medical Education (ACGME) prior to the date of the examination. In order to attempt the examination, candidates must complete an application form, pay a registration fee, and sign an agreement to be bound by the examination rules and procedures of the sanctioning board.

The examination contains 320 best answer multiple-choice items and is administered in one day. Seven hours of writing time are allowed to complete it, 3 1/2 hours in the morning and 3 1/2 hours in the afternoon. Testing for Specialty A is standardized and occurs in one location on one date each year. For the purposes of the present work, Specialty A is considered to be a homogeneous discipline.

Specialty B

The examination for Specialty B, the heterogeneous discipline, is the first of two segments that must be completed in order for a candidate to become board certified. The examination for Specialty B is comprehensive in nature and is intended to cover the breadth of the specialty area. Candidates for certification in Specialty B must have evidence of having completed an accredited residency education program prior to attempting the examination. Candidates complete a notarized application form and submit a fee in order to register for the examination.

The examination for Specialty B consists of 336 single best answer multiple-choice questions. Between 10% – 15% of the questions have a pictorial stimulus. Candidates are given six and one-half hours in which to complete the examination, three hours in the morning and three and one-half hours in the afternoon of a single day.

The written certification examination for Specialty B is administered in five different locations across the United States on one date each year. Candidates may elect to take the examination at any of those locations, but must name their choice at the time of application. The examination is given under standardized and proctored conditions in a large room at each site. Specialty B is presumed to be a heterogeneous discipline in the current study.

In sum, both examinations under study are of approximately the same length and are allotted about the same amount of time for completion. Each instrument also uses items in the multiple choice and pictorial format. Other than their subject matter, of course, the difference in the procedures of the examinations is that Specialty A has only one location for administration while Specialty B has five. Table 3 displays a summary of the examinations and the testing processes used in the present study.

Table 3

Summary of Examinations and Processes

Examination	Number of Items	Time Allotted for Completion	Questions Types	Examinati on Locations
Specialty A	320	7 Hours	Multiple Choice and Pictorial	One
Specialty B	336	6.5 Hours	Multiple Choice and Pictorial	Five

Review of Actual Examinations

In order to facilitate the identification and naming of the factors residing in the examinations, the researcher traveled to the headquarters of the respective specialty boards to personally review the actual test forms. The researcher analyzed the stems and the associated correct answer for each item on both examinations in order to determine the core topic of each item. An item's core topic was determined by making a subjective assessment about the main theme or idea of the question based on the language and terminology used in constructing the item stem and the correct answer. For purposes of examination security, the researcher was allowed to make notes concerning the general content focus of the item and the correct answer, but was not permitted to copy the items and answers verbatim. The notes allowed the researcher to identify and name the dimensions revealed by the results of the factor analysis.

Electronic Data File Procedure

The current study is conducted as a secondary analysis using data obtained from the respective sanctioning bodies of Specialty A and Specialty B. Prior to the researcher requesting the data from the respective medical specialty boards, the Human Subjects Institutional Review Board approved the study under the exempt category of review (approval letter, Appendix B). The researcher obtained the item-level data for each examination in the form of an electronic spreadsheet from the respective specialty boards.

The responses for each examination were binary in nature. When the examinee selected a correct answer, a "1" appeared in the row in the spreadsheet for that item. Incorrect responses for each item were represented as "0." Unanswered questions were blank in the data file. Inasmuch as an unanswered question was scored as an incorrect response, in those cases the researcher entered a "0" in the file. This concession was made in order to optimize the number of cases available for valid analysis. There were no such adjustments made for the Specialty B examination; all responses were present in the original file. For Specialty A, there were five items that had a substantial number of missing answers – Item 151 (313 missing), Item 152 (110), Item 153 (150), Item 154 (526), and Item 155 (49). Review of the instrument revealed that, other than being serial in the examination, these items had at least ten responses from which examinees were to choose. Virtually all of the other items on the examination were limited to the more "traditional" five possible responses. Other than having numerous item foils, it remains unclear as to why so many examinees did not provide responses to Items 151-155.

Only the examinees' responses and their composite scores were provided in the data sets. There was no identifying or demographic information in the files.

Instrument Content

This section provides a brief description of the examinations under study. Information is provided related to the how each examination was developed and presents the structure of the table of specifications.

The Tables of Specifications for the Examinations

The table of specifications for each examination was designed by a subcommittee appointed by the respective specialty boards. Identification of the item writers, decisions about correct answers and the nature of the foils, and deliberations about the range of content in each examination is privileged information that was unavailable to the present researcher. Anecdotally, it can be said that each of the subcommittees used data from job analyses and their own reflections about practice in their disciplines in order to determine the nature and magnitude of representation of content for the two tables. Volunteer writers selected by the subcommittees developed the individual items. Typically, the members of the subcommittees and the item writers are practitioners of some years of experience and/or are full-time medical academicians in the specialty. The tables of specifications for which these individuals write items are intended to represent the breadth and depth of knowledge, skills, and abilities minimally required for effective practice in each specialty as perceived by its expert practitioners.

Table 4 represents the table of specifications for Specialty A. The table describes the percentage distribution of content for the examination and the number of items under each topic.

Table 4

Content Blueprint for Specialty A

Content Area	Percentage of Items on Examination	Number of Items on Examination
Adult Disease	15%	48
Adult Trauma	25%	80
Pediatric Disease	15%	48
Pediatric Trauma	25%	80
Rehabilitation	10%	32
Basic/Applied Science	10%	32
Total	**100%**	**320**

Table 4 indicates that the topics of adult and pediatric trauma contain the preponderance of items on the examination (50%). About half of the examination is concerned with these areas. Nearly a third of the examination is focused on adult and pediatric disease processes that are associated with the specialty. Rehabilitation and basic/applied science make up the remainder of the content.

The examination for Specialty A is characterized by relatively narrow areas of content; there are only six dimensions identified in the table of specifications. Two dimensions make up half of the examination. Two others account for another 30% of its content. From this perspective, Specialty A could be described as a homogeneous discipline with a relatively focused scope.

Expected Number of Factors for Specialty A

The researcher used two points of reference in order to determine the expected number of factors for Specialty A. First, review of the table of specifications for the specialty suggested some broad parameters for the number that would be extracted. Second, an appraisal of the individual examination items themselves provided a clearer idea. Based on this, the researcher anticipated that three factors would manifest themselves in the analysis: trauma, disease, and treatment/procedures.

The examination for Specialty B consists of 22 distinct areas. The two largest dimensions of the examination each account for only 11% of its content. Most of the other areas covered consist of between 1% and 5%. For ease of review, two tables are used to display the content of Specialty B. Table 5 represents the areas of content that constitute 5% or more of the subject matter of the examination. Table 6 shows those areas of content that constitute 4% or less of the topics in the Specialty B examination.

Table 5 depicts that the larger content areas of the Specialty B examination account for nearly two-thirds of the instrument. The smaller content areas shown in

Table 6 consist of widely dispersed topics, ranging from 1% to 4% of the examination's content.

Table 5

Content Blueprint for Specialty B, Largest Content Areas

Content Area	Percentage of Items on Examination	Number of Items on Examination
Traumatic disorders	11%	37
Cardiovascular disorders	11%	37
Head, ear, eye, nose, & throat disorders	8%	27
Pediatric disorders	8%	27
Thoracic-respiratory disorders	7%	25
Abdominal & gastrointestinal disorders	7%	25
Endocrine, metabolic, & nutritional disorders	6%	20
Nervous system disorders	5%	17

Table 6 shows the distribution of content for the smaller content areas of the Specialty B examination.

Expected Number of Factors for Specialty B

In order to hypothesize the expected number of factors for Specialty B, the researcher again examined the table of specifications and the actual examination items. From this process, it was anticipated that six factors would result from the

Table 6

Content Blueprint for Specialty B, Smaller Content Areas

Content Area	Percentage of Items on Examination	Nurnber of Items on Examination
Procedures/skills	4%	13
Systemic infectious disorders	3%	11
Nontraumatic musculoskeletal disorders	3%	11
Psychobehavioral disorders	3%	11
Urogenital/gynecologic disorders	3%	10
Disaster medicine	3%	11
Obstetrics & disorders of pregnancy	2%	8
Hematologic disorders	2%	8
Environmental disorders	2%	8
Renal disorders	2%	8
Administrative medicine	2%	8
Clinical pharmacology	2%	8
Cutaneous disorders	1%	3
Immune system disorders	1%	3
Total	**100%**	**336**

analysis procedure – pharmacology, procedures, cardiology, pulmonology, trauma, and pediatrics.

To summarize this section, the respective tables of specifications for Specialties A and B show some noteworthy differences. Specialty A presents as a more homogeneous discipline, having only six major delineations of content. Two topics make up half of the content on the instrument; together, four topics comprise 80%. On the other hand, Specialty B shows a great deal more heterogeneity in its make-up, with 22 discrete areas of content in its table of specifications. No area of Specialty B accounts for more than 11% of the examination. Many areas are less than 5% of the total.

Based on a review of the respective tables of specifications and the actual examinations for each specialty, the researcher hypothesized the number of factors that would result from the factor analysis procedure. Specialty A was anticipated to contain three factors, while Specialty B was thought to contain six factors.

Characteristics of the Examinations

A brief examination of some of the key characteristics of the two examinations will help illustrate the context of the present study. Information about the length of the examinations, the number of examinees, some measures of central tendency, and other psychometric properties are presented.

Specialty A

Table 7 illustrates some of the key characteristics of the examination for Specialty A.

Table 7

Selected Characteristics of Examination Performance for Specialty A

Characteristic	Value
Number of items	309
Total number of cases	845
Total number of valid cases	845
Mean score	219.1
Range of scores	106 to 272
Standard deviation	23.7
Internal consistency coefficient (α)	.88

Table 7 shows that the examination for Specialty A consists of 309 items. The items were placed randomly in the examination booklet. A total of 845 people completed the examination. As described above, the data file was adjusted (i.e., the researcher filled in missing values with a "0") such that all 845 cases contained examinee responses in order to provide the most favorable opportunity for a valid number of cases for analysis. This adjustment accounted for less than 0.5% of all responses in the file. The internal consistency coefficient of .88 is within the range of anticipated values for an examination of this nature (Mehrens & Lehmann, 1991). Composite scores ranged from 106 to 272. The mean score was 219.1 and the standard deviation was 23.7. Figures 1, 2, and 3, respectively depict histograms of the composite scores, the difficulty indices, and the point biserial correlations for the

Figure 1

Histogram of Composite Scores for Specialty A

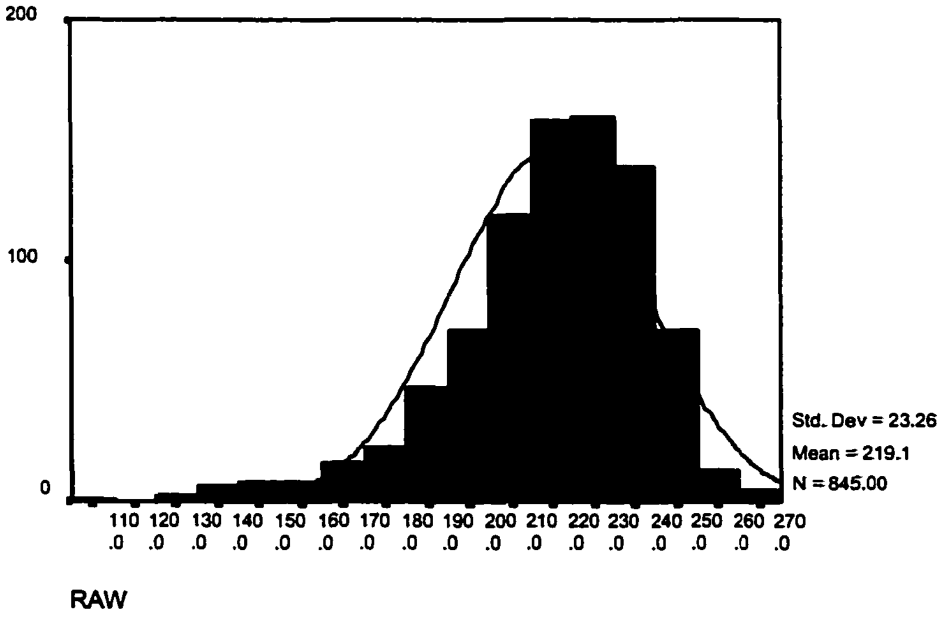


Figure 2

Histogram of Difficulty Indices for Specialty A

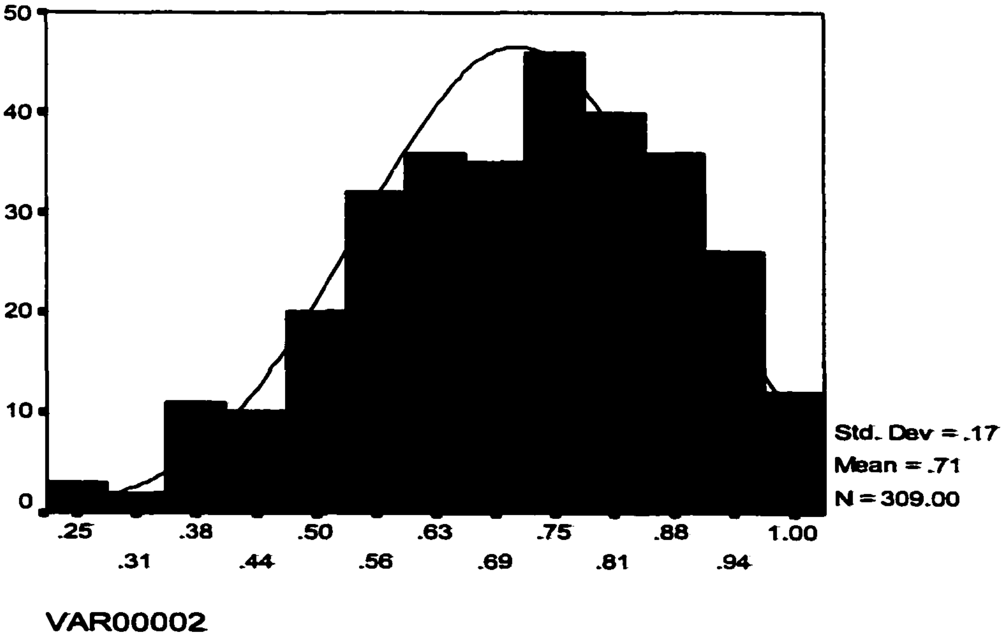
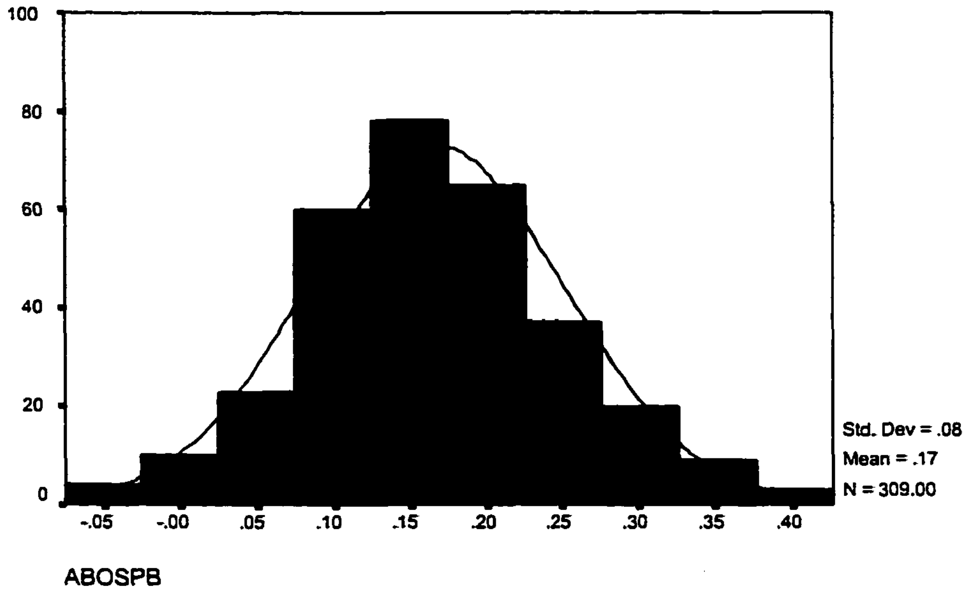


Figure 3

Histogram of Point Biserial Correlations for Specialty A



items on the Specialty A examination (a listing of these values is found in Appendix C).

An analysis of statistical and psychometric properties for Specialty A reveals that the distribution of composite scores shows a modest negative skew. The difficulty indices show moderate characteristics that suggest items that effectively create variance. (It should be noted here that no items manifested a difficulty index of 1.0. The limitations of the software used for the analysis resulted in this labeling abnormality.) Additionally, neither the difficulty indices nor the point biserial values for items placed at the end of the examination were notably lower than for other items. That is to say, there is no evidence that examinees ran out of time and resorted to guessing or randomly answering the questions in order to beat the clock. This provides support for the assertion that the item data for Specialty A is a reasonable presentation of an authentically-completed examination. The distribution of r_{bis} values is fairly normal. However, none of the 309 items for Specialty A met a criterion of $r_{\text{bis}} \geq .5$. Furthermore, six items for this specialty resulted in negative r_{bis} values.

Specialty B

Table 8 shows that the examination for Specialty B consists of 336 items. The items were placed at random on the examination form. As described previously, all 1,460 cases contained item responses and, thus, were accepted as valid for analysis. The internal consistency coefficient of .95 is well within the range of anticipated values for an examination of this nature. Additionally, even though Specialty B is considered to be the heterogeneous discipline that tests more disparate subject matter, the internal consistency coefficient is higher than that for Specialty A. Composite scores for Specialty B ranged from 88 to 306. The mean score was 247.6 and the standard deviation was 32.7. Figure 4 depicts a histogram of the composite scores for Specialty B. (A table of the difficulty index and the point biserial correlation coefficients for the items on Specialty B is found in Appendix D).

For Specialty B, the distributions of composite scores and difficulty indices are leptokurtic with a moderate negative skew. (Again, note that the limitations of the software used for the analysis inappropriately labeled some difficulty indices as being 1.0). Additionally, neither the difficulty indices nor the point biserial values for items

Table 8

Selected Characteristics of Examination Performance for Specialty B

Characteristic	Value
Number of items	336
Total number of cases	1460
Total number of valid cases	1460
Mean score	247.6
Range of scores	88 to 306
Standard deviation	32.7
Internal consistency coefficient	.95

Figure 4

Histogram of Composite Scores for Specialty B

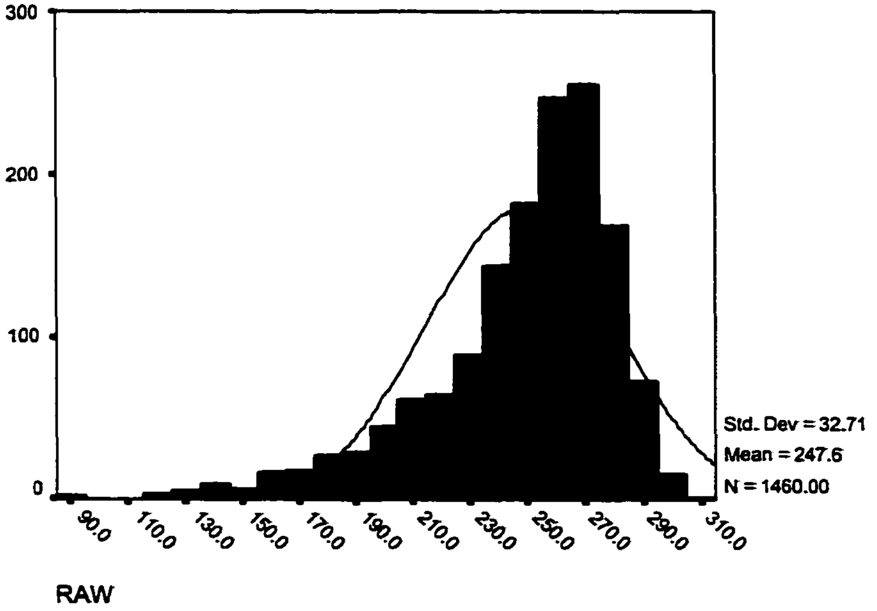


Figure 5

Histogram of Difficulty Indices for Specialty B

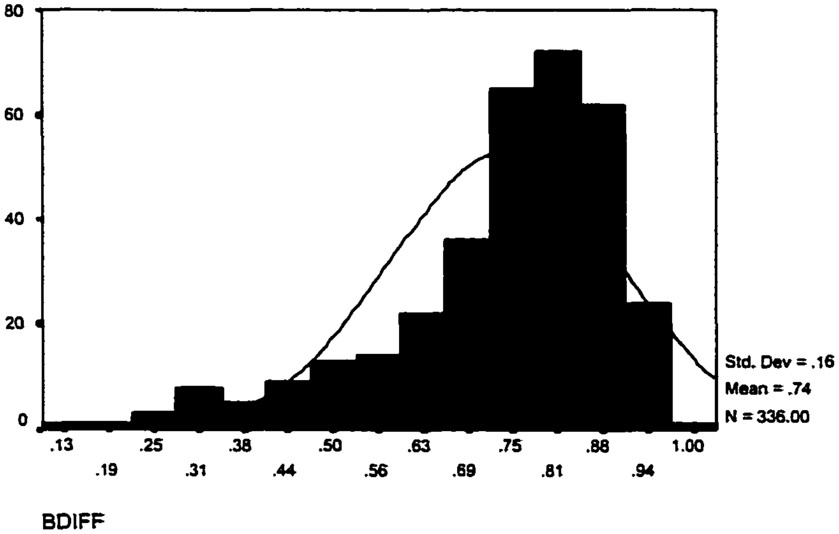
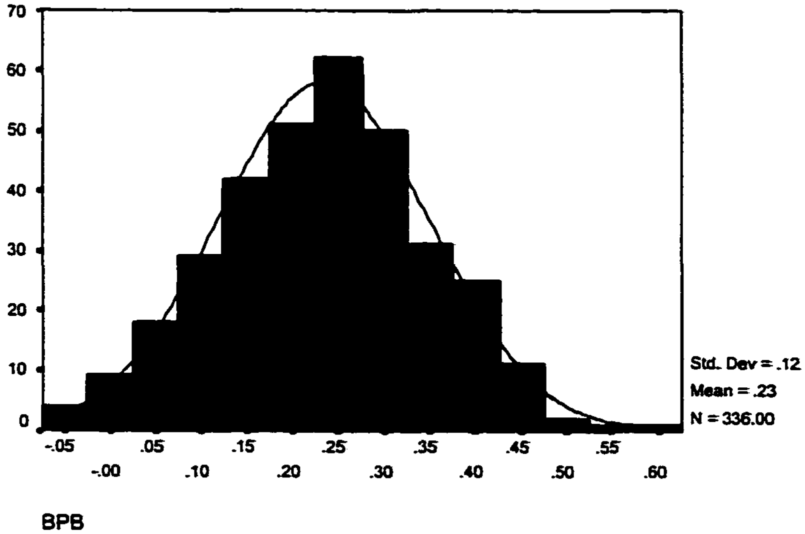


Figure 6

Histogram of Point Biserial Correlations for Specialty B



placed at the end of the examination were notably lower than for other items. That is to say, there is no evidence that examinees ran out of time and resorted to guessing or randomly answering the questions in order to beat the clock. This provides support for the assertion that the item data for Specialty B is a reasonable presentation of an authentically-completed examination. The distribution of r_{bis} is quite normal, but only two of the 336 items meet the criterion of $r_{\text{bis}} \geq .5$. In addition, six items exhibited a negative r_{bis}.

Inspecting the characteristics of the examinations side-by-side, we can see that they are of similar length and have comparable internal consistency coefficients. On the other hand, the number of examinees for the Specialty B examination is about 73% greater than the number that attempted the examination for Specialty A. Additionally, the examination for Specialty B exhibits a greater mean score, a greater range of scores, and a larger standard deviation than that of Specialty A. Table 9 shows these comparisons.

Analysis

Factor analysis was used in order to identify the structure for the examinations of Specialty A and Specialty B. Prior to analysis, the examination responses were inspected for missing or corrupted data. As reported earlier, inasmuch as an unanswered question was scored as an incorrect response, for the data in Specialty A the researcher entered a "0" in the database for those cases in order to have a

Table 9

Comparison of Selected Characteristics of Examinations for Specialty A and Specialty B

Characteristic	Specialty A (Homogeneous)	Specialty B (Heterogeneous)
Number of Items	309	336
Total Number of Cases	845	1460
Mean Composite Score	219.1	247.6
Range of Scores	106 to 272	88 to 306
Standard Deviation	23.7	32.7
Internal Consistency (α)	.88	.95

sufficient number of cases available for analysis. The data were analyzed using SPSS (v. 10.0).

Method of Extraction and Rotation

Because of the exploratory nature of the study, the researcher employed Principal Components Analysis as the data reduction procedure. The number of factors to extract was determined by conducting Velicer's Minimum Average Partial.

The initial factor scores were rotated orthogonally using the varimax procedure. This rotation procedure was selected because at this early stage of analysis for these examinations, a solution rotated orthogonally will be easier to interpret.

Interpretation of Factors

The framework for interpretation of the rotated factors was predicated on the structure of the table of specifications for each examination. As described earlier, the researcher postulated that the data for Specialty A would reveal three factors, and the data for Specialty B would resolve to six factors.

To begin, the magnitude of the factor scores for each item was considered. Those items that loaded on a particular factor at ≥ 0.30 were considered to be elements of that factor. The rotated factors were named using the researcher's notes that were obtained as described above. Titles for each cluster of items were developed by deducing the commonality evidenced among the core topics of the clusters of items. Lastly, the researcher compared the titled structures to the respective examination tables of specifications in order to determine the extent to which the derived structures reflected the table outlines.

Summary of Chapter

The preceding chapter describes the methodology for the present work. The two examinations under study, one representing homogeneous content and the other heterogeneous, are each of approximately the same length and have similar time constraints for their administration. Information was provided regarding the context, procedures, and format of each examination. The content of each instrument was also illustrated. Specialty A was hypothesized to contain three factors and Specialty B was anticipated to reveal six factors. Some psychometric properties for each examination were also presented. The internal consistency coefficients of the examinations were .88 and .95 for Specialties A and B, respectively. Neither examination showed any end-of-test effects; the difficulty indices and point biserial values were not appreciably different for items at the end of the examination when compared to those at the beginning.

The study will employ Velicer's MAP procedure to determine the number of factors to extract. Principal components analysis with varimax rotation was selected as the statistical procedure.

CHAPTER IV

RESULTS

The results of the current analysis will be presented sequentially. First, all findings for Specialty A will be presented, followed by the results obtained for Specialty B. In both cases, the findings will be represented by three key findings: the number of factors extracted, the percent of variance accounted for by the factors, and the magnitude of item loadings. Finally, we will compare and contrast the findings. Interpretation of the data in relation to the tables of specification will occur in the Discussion section.

Specialty A

An exploratory factor analysis using principal components analysis with varimax rotation was conducted on data from the examination for Specialty A. Prior to the analysis, the number of factors to be extracted for Specialty A was determined. Using the Minimum Average Partial (MAP) procedure, four factors were suggested as the number to be extracted. Recall that the researcher postulated that three factors would result from the principal components analysis. Table 10 shows these results.

Table 10

Results of MAP Procedure for Specialty A

Examination	Smallest Average Squared Correlation	MAP Factors	Expected Factors
Specialty A	.001377	4	3

Using the number of factors identified by the MAP procedure, a principal components factor analysis with varimax rotation was conducted. Table 11 depicts the percentage of variance explained for the four factors suggested for extraction from the data for Specialty A. (The explained variance for all factors for Specialty A can be found in Appendix E).

Table 11

Explained Variances for Factors 1 – 4, Specialty A

Factor	Percentage of Variance	Cumulative Percentage
1	4.21	4.21
2	1.24	5.45
3	.936	6.39
4	.866	7.25

Table 11 illustrates that the first three factors extracted for Specialty A account for slightly more than 6% of the variance in the data. Moreover, there are only two factors that account for ≥ 1 % of the variance. Every other factor extracted accounts for less than 1%. Seventy-seven factors were required to account for 50% of the variance in the examination data.

Notwithstanding that there is no meaningfully interpretable structure present relative to the table of specifications and the obtained factors, there are some things that can be observed in the data. The largest factor for Specialty A consisted of 22 items, loaded in a range from 0.30 to 0.56, and accounted for about 4% of the observed variance. The second largest factor for Specialty A (12 items) loaded in a range from 0.31 to 0.44 and explained slightly more than 1% of the variance.

The MAP procedure conducted suggests that there are two additional factors present in the data for Specialty A. However, these factors consist of only two items apiece and each accounts for only a tiny fraction of the explained variance. Thus, they do not appear to be viable factors and will not be considered in the analysis.

Tables 12 and 13 illustrate the items that clustered for Factors 1 and 2, respectively. Readers should recall that the information in the examinations was privileged. It was not possible to obtain item stems and answers verbatim. Notes from each item stem are presented. Notes from the correct answers are separated by a slash (/).

Factor 1 is named Diagnostic Skills. Table 12 presents the 22 items that load at ≤ 0.30 on the factor. Factor 1 accounts for about 4% of the variance in the data.

Table 12

Factor 1, Specialty A – *Diagnostic Skills*

No.	Item	Factor Loading
15	Short stature, frontal bossing, rhizomelia/fibroblast growth factor receptor 3	.56
9	Fibrous dysplasia/g-protein	.50
102	Pediatric clavicle mass, x-ray/No treatment	.41
191	Treatment for acute frostbite/Rapid warming at 102 degrees	.40
116	Confusion in postoperative patient, high volume of fluids had been given over 3 days/ Restriction of fluids	.39
41	Duchenne's Syndrome, scoliosis, x-ray/Posterior spinal instrumentation and fusion to sacrum	.38
14	Pain after nondisplaced subcapital fracture of femoral neck, x-ray/Osteotomy and angled blade plate	.37
185	Crush injury to foot, twitching and cramping while walking/Use of SACH orthosis	.35
118	Changes during marathon training/Increased mito-chondrial density	.34
260	Pediatric scoliosis, 1+ rotation of apical vertebrae/MRI of spine	.34
265	History of diabetes mellitus, pain and swelling in hindfoot, x-ray/open reduction and arthrodesis	.34
13	Loss of articular cartilage in osteoarthritis/Neutral proteases secreted by articular chondrocites	.32
157	Neurovascular bundle displacement/Pretendinous chord	.33
89	Infant with foot deformity, x-ray/Lateral radiograph of foot in plantar flexion	.32
104	Newborn, deformity of leg and foot, x-ray/residual limb length deformity	.32
107	C-5 quadriplegia; equinus deformity/Release of Achilles tendon and toe flexors	.32
187	Ulnar claw deformity/Metacarpophalangeal extension block splint	.32
166	Pediatric hip pain after jumping off chair, x-ray/Unicameral bone cyst	.31
192	Six month history of non-traumatic ankle pain,x-ray/Osteomyelitis	.31
266	Pediatric fall, swelling and tenderness in elbow, x-ray/Closed reduction and pin fixation	.31
72	Laceration of digital sheath/Flexion contracture of PIP joint	.30
232	Definition of epidemiological prevalence/.4% of population	.30

Although there is no clear collection of substance among the items in Factor 1, the majority of these items (except for 9, 13, 15, 72, 118, 157, and 232) are designed to measure examinees' ability to recognize and use information required in making proper diagnoses for a variety of medical conditions.

Factor 2 is named Treatment Choices. Table 13 portrays the 12 items that loaded at < 0.30 for this factor. Factor 2 accounted for a little over 1% of the variance in the data for Specialty A. Except for Item 302, Factor 2 appears to be composed of items that are designed to assess examinees' ability to recognize and select the appropriate treatment for a variety of medical conditions.

Neither of these constructs are described or delineated in a direct way within the table of specifications for Specialty A. That is to say, these factors do not manifest clustering or dimensionality of items related to particular organ systems or disease processes as described in the table of specifications. Instead, the items seem to be grouped around the way physicians approach or think about cases.

Table 13

Factor 2, Specialty A – *Treatment Choices*

No.	Item	Factor Loading
289	Flexor digitorum profundus repair, full passive ROM/Tenolysis	.44
97	Geriatric in MVA, fractured femur, x-ray/Open reduction and internal fixation using fixed angle device	.40
290	MVA, loss of consciousness, dyspnea, deformity of femur, intubation/ Auscultate chest	.37
57	Pediatric shoulder x-ray/closed reduction and percutaneous pinning	.37
154	Following pneumothorax, urine positive for occult blood/Crush injury to muscles	.34
153	Oxygen saturation decreased after 15 minutes of resuscitation/ Pneumothorax	.33
119	Hand x-ray/Fingertip ulceration and scarring	.32
306	Geriatric pain and swelling following total knee arthroplasty/ Component removal, irrigation and debridement, reimplant	.32
252	Geriatric thigh pain after fall, mass, x-ray/Metastatic carcinoma	.31
256	Ankle pain, mass, tenderness, fracture 2 days later/ Below knee amputation	.31
284	Twisted ankle, x-ray/Arthroscopic debridement	.31
302	Biomechanical properties of fixation plates/Greatest rigidity	.31

Neither of these constructs are described or delineated in a direct way within the table of specifications for Specialty A. That is to say, these factors do not manifest clustering or dimensionality of items related to particular organ systems or disease processes as described in the table of specifications. Instead, the items seem to be grouped around the way physicians approach or think about cases.

Specialty B

The data from the examination for Specialty B was subjected to the identical analysis as that of Specialty A. First, the MAP procedure was conducted. Prior to the analysis, the number of factors to be extracted for Specialty A was determined. Using the Minimum Average Partial (MAP) procedure, five factors were suggested as the number to be extracted. The researcher postulated that six factors would result from the principal components analysis. Table 14 shows these results.

Table 14

Results of MAP Procedure for Specialty B

Examination	Smallest Average Squared Correlation	MAP Factors	Expected Factors
Specialty B	.000848	5	6

The same component analysis procedures were conducted for the data on Specialty B. A principal components factor analysis with varimax rotation was conducted for the five factor solution suggested by the MAP.

Table 15 shows the percentage of variance explained for the first five factors extracted for Specialty B as prescribed by the MAP procedure. (The explained variance for all factors for Specialty B can be found in Appendix G).

Table 15

Explained Variances for Factors 1 – 5, Specialty B

Factor	Percentage of Variance	Cumulative Percentage
1	7.48	7.48
2	1.37	1.37
3	.879	9.73
4	.741	10.47
5	.719	11.19

Table 15 shows that only about 11% of the variance is explained by a five factor solution derived by the MAP procedure. Only two factors have the potency to account for ≥ 1 % of the variance. Every other factor extracted accounts for less than

1%. It required 89 factors to explain 50% of the variance in the examination data for Specialty B.

Factor 1 for Specialty B was comprised of 37 items that loaded in a range from 0.30 to 0.61. About 80% of the items in Factor 1 appear to be concerned with the measurement of examinee's skills in diagnosing and selecting a course of treatment for a variety of traumatic and medical circumstances. The exceptions to this tendency are Items 103, 131, 210, 285, 289, 309, and 319. This factor explained a little more than 7% of the variance and was named Diagnosis and Treatment Choices. Table 16 presents these 37 items.

Again it is noted that the information in the examinations was privileged. It was not possible to obtain item stems and answers verbatim. Notes from each item stem are presented. Notes from the correct answers are separated by a slash (/).

Factor 2 was made up of 26 items, loading from 0.30 to 0.47, and accounting for a little more than 1% of the variance. This factor seems to address issues that are primarily related to medical (as opposed to traumatic) conditions. Four items (86, 250, 269, and 275) are an exception to this interpretation. Factor 2 was named Internal Medicine Related Items.

Factor 3 held eight items, virtually all of which loaded in the low 0.30 range. Factor 3 is the most diffuse of the three viable factors extracted, accounting for a little less than 1% of the variance. It appears to assess a diffuse construct of physicians' ability to recognize signs and symptoms. Table 18 presents the items and loadings for Factor 3.

Table 16

Factor 1, Specialty B – *Diagnosis and Treatment Choices*

No.	Item notes	Factor Loadings
299	Hypertensive on new medication, lip and tongue swelling/Dilated ventricles	.61
167	ETOH cirrhosis, fever/Spontaneous bacterial peritonitis	.59
309	Adverse reaction to angioderma/Captopril	55
321	Pediatric, 6-day fever, oropharyngeal erythema, induration of hands and feet/ Cardiology consultation	.47
305	Cricoid pressure during intubation/Prevent passive regurgitation	.46
113	Irritability in infant, fever/Antibiotics and follow-up in 24 hours	.44
285	Herpes zoster on tip of nose/Corneal involvement	.44
324	Sudden onset of vertigo and vomiting, worsened by head position/Vestibular neuronitis	.44
30	Dyspnea, hypertropic cardiomyopathy/Propranolol	.43
103	Mechanism of beta-adregenic drugs/Convert ATP to AMP	.41
181	MVA, abdominal pain, ultrasound/intraperitoneal hemorrhage	.41
319	Action of calcium channel blocker in sub-arachnoid hemorrhage/prevents cerebral artery vasospasm	.41
327	Treatment for pepper spray/Bronchodialator	.41
322	Indicator of need for immediate delivery/Persistent late decelerations	.39
108	Ineffective thrombolytics, elevated ST segments/Rescue angioplasty	.38
211	Geriatric, vomiting, diarrhea, lab values/0.9 NS	.38
33	Bronchiolitis/Lab values provided	.37
323	Status epilepticus/Lorazepam	.37
289	Approval of hospital research protocol/HSIRB	.37
1	Obese, falls asleep easily, cognitive symptoms, polycythemia/Obstructive sleep apnea	.36
15	Wide complex tachycardia/Fusion beats	.36
20	LBP after lifting/Discharge on NSAID	.36
243	Factor in MI outcome/Early defibrillation	.36
177	Rhythm strip presented/Right-sided ECG tracing	.35
292	Worsening angina, ECG negative/Aspirin, heparin IV, nitroglycerin IV	.35
183	HIV, fever, headache, diplopia, weakness in arms/Toxoplasma gondii infection	.35
131	Vertigo with otologic origin/Nausea and vomiting	.35
227	Puncture wound in foot, osteomyelitis/Ciprofloxacin and tobramycin	.35
128	Low priority at disaster/Penetrating chest trauma with agonal respiration	.34
158	Infant, slow weight gain, small amounts of formula at each feeding/Echocardiogram	.34
210	Periorbital cellulites/Normal vision	.33
330	Digitalis toxicity, electrolyte abnormality/Hyperkalemia	.33
3	Screening for rupture of thoracic aorta/Upright P-A chest x-ray	.33
52	Renal failure in CHF, diabetic neuropathy/Low dose dopamine	.32
11	Contraindication for succinylcholine/Perforating eye injury	.31
260	Angioplasty as diagnostic and therapeutic/Crushed pelvis	.31
146	Falsely depressed oximetery reading/Shock with hypoperfusion	.30
298	Progressive loss of cognitive skills, gait disturbance, history of head trauma/Dilated ventricles	.30
315	New onset of grand mal seizure, recently stopped taking "nerve" pills/Diazepam is likely cause of seizures	.30

Table 17

Factor 2, Specialty B – *Internal Medicine Related Knowledge*

No.	Item notes	Factor Loadings
95	Jones criterion of rheumatic fever/Desquamination of skin	.47
249	Thickened bowel wall, longitudinalulcerations/Regional enteritis	.45
49	Osmolal gap, lab values presented/50	.44
229	LBP radiating down left leg, diminished sensation/Problem at L 4-5	.43
199	Photo of lesion/Delayed bleeding	.42
224	Pediatric, acute adrenal hemorrhage/Hypoglycemia	.41
256	Least useful in tricyclic antidepressant arrhythmia/Quinidine	.39
34	Confusion, polyuria, shortened QT interval/Hypercalcemia	.39
169	Contraindicated in hypertropic cardiomyopathy/Digitalis	.39
308	Von Willebrand's Disease versus hemophilia/Bleeding time	.36
188	Photo of ophthalmologic condition/Rehemorrhage	.36
205	Photo of eyes/Hypoesthesia in the distribution of the inferior orbital nerve	.36
247	Somnolence, prominent neck veins, cyanosis/Chest x-ray	.35
185	Photo of rash/Meningoencephalitis, facial nerve palsy	.34
250	Acute otitis media in adults/Streptococcus pneumoniae	.33
259	Asymptomatic chlamydia/Erythromycin base	.33
86	Lab values presented on comatose woman, serum osmolarity/350 osm/l	.32
269	ETOH, lab values/Acute pancreatitis	.32
214	Overdose, lethargy, respiratory depression, bradycardia/Clonidine	.31
275	View of zygomatic arch/Submental-vertex	.31
141	Chronic urinary retention relieved by catheter/Post-obstructive diuresis	.30
174	Chalazion/Chronic focal granulomatous inflammation of upper and lower eyelid	.30
12	Volume-type respirators versus pressure-type/Constant vent volume	.33
16	Pediatric, constipation, bloody mucus, cyanotic circumferential mucosal tissue/ Steady pressure to reduce	.33

The table of specifications for Specialty B does not prescribe any direct constructs that approximate factors of this nature. As with Specialty A, the factors derived from Specialty B do not manifest clustering or dimensionality of items related to particular organ systems or disease processes as described in the table of specifications. Instead, the items seem to be grouped around the way physicians approach or think about cases.

Table 18

Factor 3, Specialty B – *Recognition of Signs and Symptoms*

No.	Item notes	Factor Loadings
290	CO_2 monitor tube doesn't change color/Poor pulmonary perfusion	.38
225	Prospective medical control in EMS/Triage policies for prehospital patients	.33
165	Unstable angina, non-fatal MI within a week/ Transient, pain induced ST elevation in 2 leads	.32
157	Traumatic lumbar puncture/Blood clot in tube	.32
144	Isolated closed head injury/Concern in infants	.31
138	Nasal congestion caused by odors and temperature changes/Vasomotor rhinitis	.30

Comparison of Specialty A and Specialty B

The results of the principal components analysis reveal some commonalities and differences between the two examinations. In terms of commonalities, in both examinations the number of obtained factors was less than the number of factors expected by the researcher and less than the number prescribed by the MAP procedure. Secondly, Specialty A and Specialty B both returned factors concerned with the domains of diagnosis and treatment. Thirdly, the factors from both analyses did not explain a substantial part (i.e., > 50%) of the variance in the data.

Looking at the differences between the two examinations, Specialty A revealed only viable two factors, compared to the three factors that emerged from Specialty B. The two viable factors for Specialty A explained about 5% of the variance in the data, while the viable factors identified in Specialty B accounted for approximately 9%. The largest factor that resolved in Specialty A (explaining 4% of the variance) was only about half the size as the largest factor in Specialty B (7% of the variance). It required 77 factors to explain half of the variance (a benchmark for the utility of a factor analysis study) in Specialty A. For Specialty B, accounting for half of the variance required 89 factors. Lastly, Specialty B was the only examination that seemed to identify a domain of substantive medical discipline (Internal Medicine) in its data. The data from Specialty A revealed only domains related to skill in applying knowledge. Table 19 presents the comparisons between the two examinations.

Summary of the Chapter

A principal components analysis with varimax rotation was conducted on the data for Specialty A and Specialty B. The prescribed four factor MAP solution for Specialty A explained about 7% of the variance. However, only two of the factors were viable; thus, 5% is a more accurate portrayal of the obtained structure. The other factors in Specialty A contained only two items apiece. The five factor MAP solution for Specialty B accounted for about 11% of the variance. Again, though, a fewer number of factors (three) were actually viable, accounting for 9% of the explained variance. For both examinations, none of the derived factors appeared to approximate constructs prescribed from the respective tables of specifications. However, the examinations did have a domain in common concerned with diagnosis and treatment. The examinations were most clearly differentiated by the factor concerned with knowledge of internal medicine that was found in Specialty B.

Table 19

Comparison of Results for Specialty A and Specialty B

	Specialty A (homogeneous)	Specialty B (heterogeneous)
Number of cases	845	1460
Number of items	309	336
Range of scores	106 – 272	88 – 306
Mean composite score	219.1	247.6
Standard deviation	23.7	32.7
Internal consistency (α)	.88	.95
Number of expected factors	3	6
Factors prescribed by MAP procedure	4	5
Variance explained by prescribed MAP solution	7%	11%
Viable factors derived	2	3
Variance explained by viable factors	5%	9%
Factors required to explain 50% of variance	77	89
Names of derived factors	Diagnostic skills Treatment choices	Diagnosis and treatment Internal Medicine Recognizing signs

Irrespective of the outcome of the present study, the examinations for Specialty A and Specialty B stand as good instruments, psychometrically speaking. Difficulty indices, distributions of composite scores, internal consistency coefficients, and point biserial correlations generally show evidence of well-written examinations.

CHAPTER V

DISCUSSION

There were two purposes for the present study. The first was to investigate the extent to which the factor structure of two physician specialty certification examinations represented the examinations' table of specifications. The second purpose was to assess the degree to which factor analysis provides a different magnitude of information when the data come from a broad field of medical practice versus a more tightly bounded specialty. The rationale for pursuing issues of this nature was to test the potency of the factor analysis procedure and to contribute to the validity evidence file in the realm of physician certification testing.

Overview

In sum, the results of the study revealed that the structure derived from principal components analysis with varimax rotation for Specialty A and Specialty B did not match the content categories of the respective tables of specifications. The two viable factors extracted for Specialty A (1 – diagnosis, 2 – treatment) accounted for about 5% of the observed variance. The viable three factor solution for Specialty B (1 – diagnosis and treatment, 2 – medicine, 3 – recognition of signs) accounted for about 9%. In neither examination was there any similarity between the derived factors and the constructs identified in the respective tables of specifications.

Before proceeding in the discussion, it is worth noting that the researcher made efforts to reveal a greater magnitude of structure in the data. First, attempts were made to follow the item selection criteria established by Ebel (1965). These authors suggest that each item must meet the criterion of $0.25 \leq p \leq 0.75$ and $r_{bis} \geq 0.30$ in order to be included in the factor analysis. Using this method, only eight items from the examination for Specialty A and 24 items for Specialty B met the criteria for inclusion. As a result, the researcher employed a slightly liberalized version of the criterion ($0.25 \leq p \leq 0.80$ and $r_{bis} \geq 0.25$). However, this effort resulted only in a total of 17 items for Specialty A and 97 for Specialty B, which the researcher also deemed too small for a valid analysis.

In addition, a one factor analysis was conducted for each examination to determine if any structure could be identified. This analysis resulted in only 36 items achieving the criteria of factor loading $\geq |0.30|$ for Specialty A and 81 for Specialty B. (Appendices G and H show the listing of item loadings for the one factor solution for each examination.)

As a result of these findings, we can say that there appears to be no factorial evidence to support the existence of a relationship between the tables of specifications for these examinations and their derived factor structures. That is to say, there is no clear structure in which one could differentiate or identify the various content constructs prescribed in the tables of specifications.

In discussing these findings, it is first important to clarify an interpretation that might be made. The lack of a clearly observable relationship between the factor structure and the examinations' table of specifications should not be interpreted as indicating a lack of validity for purposes of physician certification. Instead, it is more appropriate to say that this particular quantitative method (i.e., principal components analysis) does not contribute to the validity evidence file. As noted in chapters one and two, validity evidence comes from a number of sources. Not having a clear result in the present study does not obviate the utility and strength of other realms of construct-, content-, or criterion-related types of evidence. In other words, the current findings do not indicate that validity is missing, just that this method of inquiry does not show it. Validity evidence must be obtained from other sources and procedures. At the same time, however, there are differential constructs that emerged from the analyses.

Considering the other purpose of the study – determining the efficacy of factor analysis in differentiating between the structure of two types specialties – the results were somewhat more revealing. There were fewer viable factors in the data for Specialty A (i.e., 2) than there were in Specialty B (3). Even though the difference is of a lesser magnitude than anticipated, this finding does provide some support for the efficacy of EFA in the present context. Specialty A was considered to be a more closely-knit field, meaning that it concerns a relatively limited number of organs, body systems, or anatomy. Specialty B, on the other hand, concerns the entire range of anatomy and physiology found in the human body. The assumption was that Specialty B would show a more diffuse structure (i.e., more factors and less explained variance) than Specialty A. This assumption was modestly supported by the results.

Specialty B did reveal a greater number of viable factors (3) than did Specialty A (2). Specialty B also required 89 factors to account for 50% of the variance in the data, while Specialty A needed only 77.

Interpretation

Several points can be made in trying to make meaning out of these results. First, it would appear that augmentation of validity evidence for physician certification examinations will most likely need to occur outside of that provided by factor-type studies. Content studies, naturalistic observation, and associations between certification and results such as patient satisfaction or clinical outcomes may be the most effectual alternatives for developing the evidence file. Of course, content study is the general approach taken in the current development of specialty examinations. The latter two suggestions are rife with pitfalls. For example, there is no way to control for the acuity of illness among patients. A physician with a reputation as a "good doctor" may be more likely to receive the most difficult cases. His/her patients may be more seriously ill, and, thus, are more apt to result in an adverse outcome. Judging clinical efficacy by this standard may render a misleading result.

Second, the absence of a meaningful relationship between the results of the factor study and the respective tables of specification raises other issues. If content is the key dimension in the development and validation of an examination of physician certification, is assessment of structure a meaningful inquiry? More to the point, factor studies do not provide any evidence concerning how well the content domain has been sampled. Having information about the examination structure does not provide us with any insight into how much certified physicians know about the domain of their specialty. The point to be reinforced here is that it is possible that factor analysis is simply not a practical method for identifying the structure of examinations of this nature.

A third observation is concerned with epistemology. With no profoundly clear factors revealed in the structure of the examination data, what does that indicate about the nature of medical knowledge in these two specialties? One impression is that medical knowledge might be structured as individual "nuggets" of information. In other words, each item on an examination might be considered as a unique and discrete instance of practice. From this perspective, there is little in terms of generalizable dimensions of knowledge that a physician can call upon in approaching a case. Instead, knowledge is almost dichotomous – either he/she knows the right answer or doesn't. This interpretation of the data is somewhat reminiscent of the early theory of intelligence referred to as "*g*" (Spearman, 1927; Jensen, 1998). Applying this theory, a fund of knowledge in a medical specialty would consist of *g* – a general medical knowledge factor, and *s* – a factor specific to each case. For the current study, the viable factors identified in the instruments would be *s*, while the remainder of the examination would represent *g*.

The present findings also suggest another perspective concerning the cognitive dynamics of medical practice. Commonly, medical specialties are differentiated by organ system (e.g., pulmonology, nephrology, cardiology) and/or patient population (e.g., pediatrics, geriatrics). As such, this taxonomy likely served as a significant variable in structuring the examination instruments. On the other hand, the largest factors extracted from each specialty seemed to reside in the activities of diagnosis and treatment. Perhaps it is the case that the mental model physicians use for practice is based less on organ system or population than it is predicated on constructs such as "things to look for in a patient" and "decisions I need to make." Put a different way, this interpretation suggests that medical education and practice are more than the learning and reciting of facts that are to be applied under certain conditions, cognitively speaking. Instead, medical practice is more a way of thinking and problem solving. In a sense, it is the physician using what he/she does know and can see to figure out what he/she doesn't know and can't see.

Limitations of the Present Work

Before one can put any stock in these interpretations, there are some issues about the current study that need to be considered. First, the present work investigated only two examinations out of the 24 recognized medical specialties. It is possible that data from other disciplines may reveal greater structure and adherence to their respective tables of specification.

Secondly, only one examination year was utilized in the current analysis. For both specialties, the examination changes from year to year to a greater or lesser degree. Using data from a different year could possibly show different results.

Thirdly, this research applied only one factor analysis procedure – principal components analysis using varimax rotation. There are other procedures in the factor analysis family that could have been used and provided a different outcome. This consideration is less likely than the others cited because, as noted earlier, the various factor analysis techniques often provide very similar results.

Suggestions for Further Research

For those interested in continuing research in this vein, there are some ways to improve upon and complement the work completed here. One suggestion would be to attempt to replicate this study by selecting two different medical specialties. It is plausible that analysis of specialties that have even more widely disparate subject matter may show a different result.

A second idea would be to conduct the analysis using data from more than one year of testing. Because the examinations change to some degree each year, this proposal could prove to be difficult to execute. Perhaps if a researcher selected only the items that were in common to various versions of the examination there would be a sufficient volume of items to carry out an analysis.

A final suggestion would be to employ a different factor analysis procedure in the study. Although common methods of factor analysis have been demonstrated to reveal similar outcomes, it is possible that an entirely different technique from the family of factor analysis would render a different outcome.

Summary and Conclusions

The present study found no evidence of a structure that approximated the tables of specifications in the examination data obtained from two differentiated medical specialties. As a result, there is no clear evidence linking the structure to the respective tables of specifications. Conversely, there was some support for the presumption that a more heterogeneous specialty has more elements of structure than does a more homogeneous specialty. This finding provides some support for the efficacy of factor analysis in studies of this nature. There are some implications for these findings. An important one is that factor studies may not be a practical tool for developing validity evidence in content-based instruments such as physician certification examinations. It may also be the case that physicians have an entirely different mental model for practice than that manifested by a specialty examination. Replication and/or expansion of the current study may help illuminate these possibilities.

These findings notwithstanding, it remains incumbent upon developers of high stakes examinations such as these to find and employ methods of validation that provide appropriate support for the interpretation of their results. As with any examination, failing to develop such evidence undermines the veracity of the inferences the test developers and the examinees themselves want us to make.

Appendix A

Medical Specialties Recognized by the American Board of Medical Specialties

American Board of:

Allergy and Immunology

Anesthesiology

Colon and Rectal Surgery

Dermatology

Emergency Medicine

Family Practice

Internal Medicine

Medical Genetics

Neurological Surgery

Nuclear Medicine

Obstetrics and Gynecology

Ophthalmology

Orthopedic Surgery

Otolaryngology

Pathology

Pediatrics

Physical Medicine and Rehabilitation

Plastic Surgery

Preventive Medicine

Psychiatry and Neurology

Radiology

Surgery

Thoracic Surgery

Urology

Appendix B

Protocol Clearance from the Human Subjects Institutional Review Board

WESTERN MICHIGAN UNIVERSITY

Date: June 14. 2001

To: Mary Anne Bunda, Principal Investigator
Jeffrey Green, Student Investigator for dissertation

From: Michael S. Pritchard, Interim Chair

Re: HSIRB Project Number 01-06-07

This letter will serve as confirmation that your research project entitled "A Factor Analytic Investigation of Two Medical Specialty Certification Examinations" has been **approved** under the **exempt** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note that you may **only** conduct this research exactly in the form it was approved. You must seek specific board approval for any changes in this project. You must also seek reapproval if the project extends beyond the termination date noted below. In addition if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: June 14. 2002

Appendix C

Difficulty Indices and Point Biserial Correlations for Examination of Specialty A

Item	Difficulty	Point Biserial
1	.73	.16
2	.73	.24
3	.80	.10
4	.83	.24
5	.56	-.03
6	.56	.07
7	.24	.04
8	.50	.15
9	.46	.27
10	.44	.06
11	.73	.13
12	.89	.11
13	.59	.26
14	.61	.37
15	.83	.35
16	.62	.21
17	.72	.19
18	.74	.12
19	.65	.12
20	.46	.15
21	.75	.03
22	.63	.13
23	.85	.11
24	.54	.15
25	.93	.21
26	.80	-.03
27	.58	.18
28	.88	.18
29	.77	.16
30	.55	.14

Appendix C – Continued

Item	Difficulty	Point Biserial
31	.86	.22
32	.62	.24
33	.82	.19
34	.60	.17
35	.51	.21
36	.69	.14
37	.61	.16
38	.66	.14
39	.87	.24
40	.49	.06
41	.81	.34
42	.64	.08
43	.72	.19
44	.90	.16
45	.42	.09
46	.86	.30
47	.80	.18
48	.87	.18
49	.37	.12
50	.28	.00
51	.59	.10
52	.58	.13
53	.70	.10
54	.59	.21
55	.85	.10
56	.62	.16
57	.64	.19
58	.43	.10
59	.96	.13
60	.94	.18

Appendix C – Continued

Item	Difficulty	Point Biserial
61	.38	.15
62	.95	.02
63	.71	.12
64	.76	.10
65	.91	.23
66	.98	.12
67	.72	.22
68	.71	.05
69	.81	.17
70	.51	.00
71	.77	.24
72	.89	.29
73	.60	.20
74	.64	.11
75	.76	.11
76	.97	.09
77	.91	.19
78	.49	-.04
79	.61	.16
80	.83	.21
81	.52	.19
82	.81	.07
83	.75	.26
84	.61	.10
85	.71	.13
86	.98	.17
87	.62	.10
88	.57	.13
89	.78	.10
90	.69	.19

Appendix C – Continued

Item	Difficulty	Point Biserial
91	.54	.10
92	.93	.08
93	.89	.09
94	.88	.21
95	.59	.23
96	.45	.12
97	.78	.36
98	.68	.20
99	.73	.15
100	.94	.21
101	.49	.10
102	.75	.30
103	.79	.19
104	.86	.25
105	.70	.23
106	.62	.22
107	.78	.15
108	.67	.16
109	.59	.33
110	.95	.18
111	.96	.28
112	.56	.15
113	.55	.09
114	.60	.15
115	.68	.16
116	.88	.40
117	.67	.08
118	.62	.28
119	.76	.28
120	.67	.17

Appendix C – Continued

Item	Difficulty	Point Biserial
121	.23	.11
122	.83	.22
123	.86	.15
124	.58	.06
125	.51	.22
126	.62	.09
127	.71	.23
128	.94	.22
129	.77	.19
130	.88	.17
131	.89	.20
132	.92	.06
133	.44	.15
134	.81	.12
135	.96	.15
136	.83	.04
137	.95	.13
138	.75	.17
139	.66	.15
140	.43	.04
141	.80	.02
142	.49	.04
143	.36	.12
144	.72	.09
145	.93	.27
146	.96	.26
147	.54	.07
148	.50	.26
149	.69	.20
150	.37	.01

Appendix C – Continued

Item	Difficulty	Point Biserial
151	.89	.23
152	.53	.22
153	.75	.36
154	.58	.25
155	.61	.17
156	.62	.20
157	.63	.41
158	.58	-.07
159	.78	.23
160	.82	.31
161	.88	.00
162	.82	.10
163	.83	.15
164	.79	.19
165	.49	.05
166	.89	.27
167	.51	.16
168	.67	.12
169	.71	.04
170	.81	.31
171	.40	.09
172	.97	.16
173	.41	.13
174	.60	.21
175	.85	.24
176	.62	.10
177	.70	.06
178	.88	.14
179	.68	.13
180	.56	.23

Appendix C – Continued

Item	Difficulty	Point Biserial
181	.34	.04
182	.97	.11
183	.50	.15
184	.65	.12
185	.88	.34
186	.98	.18
187	.73	.24
188	.68	.29
189	.80	.18
190	.56	.08
191	.67	.37
192	.79	.32
193	.79	.13
194	.79	.13
195	.86	.16
196	.67	.28
197	.62	.03
198	.57	.13
199	.33	.10
200	.62	.18
201	.88	.21
202	.71	.20
203	.81	.25
204	.74	.21
205	.49	.22
206	.96	.13
207	.83	.34
208	.91	.15
209	.97	.22
210	.99	.06

Appendix C – Continued

Item	Difficulty	Point Biserial
211	.81	.22
212	.84	.29
213	.54	.15
214	.72	.16
215	.90	.17
216	.68	.15
217	.86	.10
218	.73	.17
219	.56	.15
220	.67	.15
221	.93	.31
222	.58	.15
223	.85	.13
224	.49	.11
225	.85	.05
226	.98	.08
227	.76	.21
228	.97	.08
229	.58	.09
230	.53	.12
231	.65	.11
232	.67	.38
233	.74	.16
234	.92	.27
235	.77	.12
236	.71	.18
237	.70	.22
238	.79	.24
239	.57	.32
240	.65	.15

Appendix C – Continued

Item	Difficulty	Point Biserial
241	.73	.06
242	.51	.17
243	.91	.22
244	.66	.00
245	.87	.05
246	.83	.21
247	.59	.17
248	.62	.06
249	.63	-.02
250	.38	.09
251	.38	.17
252	.69	.27
253	.72	.17
254	.58	.18
255	.62	.17
256	.96	.22
257	.87	.17
258	.36	-.02
259	.40	.19
260	.86	.25
261	.59	.17
262	.75	.16
263	.91	.23
264	.98	.19
265	.84	.28
266	.73	.19
267	.80	.18
268	.35	.21
269	.98	.25
270	.79	.22

Appendix C – Continued

Item	Difficulty	Point Biserial
271	.77	.19
272	.73	.22
273	.72	.14
274	.89	.16
275	.47	.18
276	.78	.11
277	.81	.16
278	.93	.29
279	.84	.25
280	.76	.19
281	.93	.26
282	.80	.11
283	.86	.24
284	.61	.28
285	.40	.11
286	.75	.08
287	.70	.12
288	.76	.02
289	.78	.32
290	.83	.28
291	.64	.26
292	.90	.12
293	.60	.26
294	.91	.17
295	.78	.11
296	.71	.15
297	.60	.12
298	.68	.18
299	.77	.21
300	.51	.12

Appendix C – Continued

Item	Difficulty	Point Biserial
301	.89	.13
302	.55	.25
303	.45	.08
304	.67	.20
305	.54	.15
306	.83	.32
307	.84	.25
308	.82	.16
309	.78	.13

Appendix D

Difficulty Indices and Point Biserial Correlations for Examination of Specialty B

Item	Difficulty	Point Biserial
1	.85	.35
2	.85	.27
3	.82	.36
4	.37	.03
5	.85	.17
6	.48	.09
7	.79	.17
8	.79	.28
9	.83	.18
10	.58	.21
11	.60	.44
12	.67	.23
13	.71	.15
14	.75	.31
15	.58	.37
16	.79	.33
17	.75	.25
18	.85	.25
19	.64	.14
20	.86	.41
21	.62	.10
22	.76	.41
23	.76	.11
24	.79	.14
25	.92	.12
26	.86	.22
27	.89	.13
28	.53	.12
29	.84	.26
30	.75	.33

Appendix D – Continued

Item	Difficulty	Point Biserial
41	.30	.02
42	.83	.29
43	.77	.34
44	.65	-.02
45	.87	.22
46	.70	.27
47	.82	.11
48	.45	.27
49	.77	.44
50	.80	.09
51	.72	.34
52	.82	.32
53	.91	.20
54	.76	.25
55	.93	.28
56	.82	.45
57	.66	.27
58	.46	.20
59	.88	.13
60	.89	.17
61	.89	.11
62	.71	.20
63	.84	.33
64	.82	.27
65	.51	.15
66	.87	.23
67	.79	.29
68	.78	.26
69	.91	.42
70	.68	.26

Appendix D – Continued

Item	Difficulty	Point Biserial
71	.68	.12
72	.84	.30
73	.41	-.16
74	.76	.20
75	.76	.16
76	.47	.27
77	.54	.18
78	.73	.46
79	.92	.48
80	.80	.39
81	.79	.27
82	.76	.30
83	.85	.39
84	.79	.19
85	.48	.04
86	.81	.31
87	.65	.08
88	.67	.12
89	.62	.13
90	.72	.32
91	.78	.34
92	.74	.17
93	.91	.40
94	.79	.28
95	.74	.27
96	.78	.27
97	.75	.39
98	.85	.16
99	.81	.01
100	.84	.39

Appendix D – Continued

Item	Difficulty	Point Biserial
101	.67	.25
102	.75	.25
103	.90	.41
104	.60	.20
105	.51	.18
106	.75	.22
107	.76	.12
108	.91	.31
109	.85	.19
110	.77	.25
111	.80	.14
112	.24	.06
113	.78	.39
114	.84	.23
115	.73	.11
116	.71	.29
117	.86	.25
118	.44	.22
119	.85	.29
120	.81	.33
121	.83	.16
122	.73	.19
123	.52	.10
124	.73	.28
125	.85	.21
126	.81	.23
127	.31	.14
128	.85	.28
129	.80	.26
130	.90	.32

Appendix D – Continued

Item	Difficulty	Point Biserial
131	.87	.42
132	.83	.23
133	.80	.25
134	.86	.15
135	.68	.07
136	.85	.24
137	.56	.07
138	.78	.18
139	.72	.32
140	.62	.03
141	.78	.39
142	.89	.34
143	.63	.05
144	.57	.29
145	.84	.13
146	.86	.38
147	.92	.23
148	.87	.21
149	.93	.26
150	.81	.17
151	.82	.18
152	.81	.29
153	.69	.07
154	.86	.19
155	.29	.19
156	.92	.15
157	.71	.31
158	.76	.22
159	.81	.06
160	.46	.00

Appendix D – Continued

Item	Difficulty	Point Biserial
161	.89	.16
162	.96	.25
163	.92	.28
164	.30	.09
165	.62	.21
166	.52	.23
167	.87	.61
168	.63	-.04
169	.68	.38
170	.74	.10
171	.25	.03
172	.94	.15
173	.89	.18
174	.63	.23
175	.58	.33
176	.78	.29
177	.90	.39
178	.89	.11
179	.45	-.06
180	.76	.20
181	.67	.49
182	.70	.22
183	.81	.35
184	.50	.15
185	.74	.35
186	.81	.14
187	.81	.23
188	.85	.34
189	.77	.14
190	.69	.16

Appendix D – Continued

Item	Difficulty	Point Biserial
191	.88	.23
192	.90	.21
193	.78	.21
194	.36	.09
195	.68	.27
196	.80	.28
197	.77	.29
198	.76	.00
199	.76	.30
200	.90	.27
201	.31	.04
202	.75	.20
203	.62	.13
204	.18	.01
205	.88	.42
206	.90	.30
207	.80	.10
208	.87	.14
209	.56	.00
210	.87	.37
211	.79	.40
212	.65	.17
213	.92	.22
214	.63	.27
215	.71	.12
216	.86	.35
217	.84	.19
218	.57	.21
219	.66	.07
220	.92	.21

Appendix D – Continued

Item	Difficulty	Point Biserial
221	.87	.25
222	.83	.14
223	.49	.27
224	.67	.45
225	.67	.23
226	.89	.25
227	.79	.40
228	.85	.21
229	.72	.26
230	.93	.22
231	.98	.26
232	.81	.21
233	.57	.29
234	.86	.09
235	.78	.08
236	.84	.35
237	.83	.06
238	.88	.22
239	.35	.17
240	.80	.34
241	.80	.32
242	.83	.17
243	.88	.34
244	.82	.17
245	.70	.25
246	.69	.27
247	.62	.32
248	.45	.21
249	.68	.27
250	.82	.33

Appendix D – Continued

Item	Difficulty	Point Biserial
251	.60	.28
252	.92	.34
253	.57	.22
254	.78	.37
255	.76	.30
256	.70	.34
257	.65	.34
258	.78	.21
259	.70	.28
260	.80	.42
261	.15	.00
262	.71	.28
263	.75	.30
264	.73	.10
265	.84	.21
266	.86	.22
267	.71	.29
268	.70	.26
269	.84	.38
270	.86	.18
271	.77	.35
272	.32	.08
273	.78	.15
274	.81	.26
275	.65	.25
276	.69	.24
277	.83	.15
278	.72	.31
279	.65	.39
280	.83	.22

Appendix D – Continued

Item	Difficulty	Point Biserial
281	.88	.09
282	.74	.20
283	.84	.28
284	.89	.31
285	.95	.37
286	.86	.19
287	.78	.24
288	.83	.24
289	.70	.39
290	.78	.35
291	.60	.12
292	.76	.40
293	.46	-.01
294	.84	.24
295	.56	.14
296	.78	.18
297	.77	.28
298	.74	.43
299	.93	.54
300	.83	.20
301	.66	.17
302	.52	.25
303	.85	.28
304	.44	.10
305	.77	.47
306	.79	.29
307	.87	.25
308	.59	.18
309	.96	.46
310	.79	.31

Appendix D – Continued

Item	Difficulty	Point Biserial
311	.82	.28
312	.78	.36
313	.30	.07
314	.83	.23
315	.76	.32
316	.77	.28
317	.66	.05
318	.87	.16
319	.81	.46
320	.84	.27
321	.86	.47
322	.94	.35
323	.94	.33
324	.94	.38
325	.50	.29
326	.22	.14
327	.93	.34
328	.39	-.03
329	.37	.19
330	.81	.41
331	.64	.29
332	.58	.27
333	.69	.05
334	.86	.23
335	.85	.06
336	.85	.06

Appendix E

Explained Variance for Specialty A,
All Factors

Factor	Percentage of Variance	Cumulative Percentage
1	4.21	4.21
2	1.24	5.45
3	.936	6.39
4	.866	7.25
5	.836	8.08
6	.821	8.90
7	.808	9.71
8	.789	10.50
9	.769	11.27
10	.761	12.03
11	.753	12.78
12	.740	13.52
13	.733	14.26
14	.727	14.98
15	.718	15.70
16	.713	16.42
17	.701	17.12
18	.698	17.89
19	.693	18.51
20	.681	19.19
21	.673	19.86
22	.670	20.53
23	.664	21.20
24	.659	21.85
25	.654	22.51
26	.652	23.26
27	.639	23.80
28	.632	24.43
29	.627	25.06
30	.622	25.68

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
31	.617	26.30
32	.616	26.91
33	.611	27.52
34	.605	28.13
35	.596	28.73
36	.595	29.32
37	.593	29.91
38	.584	30.50
39	.580	31.08
40	.577	31.66
41	.574	32.23
42	.569	32.80
43	.563	33.36
44	.559	33.92
45	.554	34.47
46	.550	35.02
47	.546	35.57
48	.540	36.11
49	.535	36.64
50	.529	37.17
51	.528	37.70
52	.524	38.22
53	.520	38.74
54	.517	39.26
55	.514	39.78
56	.509	40.29
57	.507	40.79
58	.502	41.29
59	.502	41.80
60	.498	42.29

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
61	.497	42.79
62	.491	43.28
63	.487	43.77
64	.484	44.25
65	.482	44.73
66	.477	45.21
67	.476	45.69
68	.472	46.16
69	.470	46.63
70	.468	47.10
71	.465	47.56
72	.463	48.03
73	.458	48.48
74	.455	48.94
75	.451	49.39
76	.447	49.84
77	.446	50.28
78	.445	50.73
79	.439	51.17
80	.437	51.60
81	.436	52.04
82	.430	52.47
83	.426	52.90
84	.424	53.32
85	.422	53.74
86	.417	54.16
87	.417	54.58
88	.413	54.99
89	.411	55.40
90	.408	55.81

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
91	.402	56.21
92	.399	56.61
93	.397	57.01
94	.396	57.40
95	.393	57.79
96	.392	58.19
97	.388	58.58
98	.386	58.96
99	.384	59.35
100	.383	59.73
101	.380	60.11
102	.378	60.49
103	.374	60.86
104	.373	61.23
105	.368	61.60
106	.366	61.97
107	.365	62.33
108	.363	62.70
109	.361	63.06
110	.360	63.42
111	.359	63.77
112	.356	64.13
113	.353	64.48
114	.351	64.84
115	.348	65.18
116	.344	65.53
117	.343	65.87
118	.340	66.21
119	.337	66.55
120	.332	66.88

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
121	.332	67.21
122	.330	67.54
123	.329	67.87
124	.327	68.20
125	.325	68.52
126	.322	68.84
127	.320	69.16
128	.317	69.48
129	.314	69.79
130	.313	70.11
131	.310	70.42
132	.310	70.73
133	.306	71.03
134	.305	71.34
135	.304	71.64
136	.303	71.94
137	.301	72.25
138	.299	72.54
139	.296	72.84
140	.295	73.14
141	.293	73.43
142	.288	73.72
143	.288	74.00
144	.286	74.29
145	.285	74.57
146	.283	74.86
147	.282	75.14
148	.279	75.42
149	.277	75.69
150	.276	75.97

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
151	.274	76.25
152	.272	76.52
153	.271	76.79
154	.267	77.05
155	.264	77.32
156	.263	77.58
157	.261	77.84
158	.259	78.10
159	.258	78.36
160	.256	78.61
161	.255	78.87
162	.254	79.12
163	.251	79.37
164	.250	79.62
165	.248	79.87
166	.246	80.12
167	.244	80.36
168	.243	80.61
169	.241	80.85
170	.239	81.08
171	.235	81.32
172	.234	81.55
173	.234	81.79
174	.231	82.02
175	.229	82.25
176	.227	82.48
177	.227	82.70
178	.224	82.93
179	.222	83.15
180	.222	83.37

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
181	.220	83.59
182	.219	83.81
183	.216	84.02
184	.214	84.24
185	.213	84.45
186	.212	84.66
187	.209	84.87
188	.208	85.08
189	.206	85.29
190	.204	85.49
191	.203	85.69
192	.201	85.89
193	.201	86.09
194	.200	86.29
195	.197	86.49
196	.195	86.69
197	.193	86.88
198	.191	87.07
199	.190	87.26
200	.188	87.45
201	.187	87.63
202	.186	87.82
203	.184	88.00
204	.183	88.19
205	.182	88.37
206	.181	88.55
207	.179	88.73
208	.177	88.91
209	.176	89.08
210	.175	89.26

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
211	.173	89.43
212	.172	89.60
213	.171	89.77
214	.168	89.94
215	.166	90.11
216	.165	90.27
217	.164	90.43
218	.163	90.60
219	.162	90.76
220	.159	90.92
221	.157	91.08
222	.156	91.23
223	.154	91.39
224	.153	91.54
225	.152	91.69
226	.151	91.84
227	.150	91.99
228	.148	92.14
229	.147	92.29
230	.146	92.44
231	.145	92.58
232	.144	92.72
233	.140	92.86
234	.139	93.09
235	.139	93.14
236	.137	93.28
237	.136	93.42
238	.135	93.55
239	.132	93.68
240	.131	93.81

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
241	.131	93.95
242	.130	94.08
243	.129	94.20
244	.127	94.33
245	.125	94.46
246	.124	94.58
247	.123	94.70
248	.120	94.82
249	.119	94.94
250	.119	95.06
251	.118	95.18
252	.118	95.30
253	.115	95.41
254	.115	95.64
255	.112	95.64
256	.110	95.75
257	.109	95.86
258	.109	96.08
259	.108	96.18
260	.107	96.29
261	.107	96.39
262	.105	96.39
263	.104	96.50
264	.102	96.60
265	.101	96.70
266	.101	96.80
267	<.10	96.90
268	<.10	97.00
269	<.10	97.10
270	<.10	97.19

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
271	<.10	97.29
272	<.10	97.38
273	<.10	97.48
274	<.10	97.57
275	<.10	97.66
276	<.10	97.75
277	<.10	97.84
278	<.10	97.92
279	<.10	97.01
280	<.10	97.09
281	<.10	98.17
282	<.10	98.25
283	<.10	98.33
284	<.10	98.41
285	<.10	98.49
286	<.10	98.57
287	<.10	98.65
288	<.10	98.72
289	<.10	98.79
290	<.10	98.87
291	<.10	98.94
292	<.10	99.01
293	<.10	99.08
294	<.10	99.15
295	<.10	99.21
296	<.10	99.28
297	<.10	99.34
298	<.10	99.41
299	<.10	99.47
300	<.10	99.53

Appendix E – Continued

Factor	Percentage of Variance	Cumulative Percentage
301	<.10	99.59
302	<.10	99.64
303	<.10	99.70
304	<.10	99.75
305	<.10	99.80
306	<.10	99.85
307	<.10	99.90
308	<.10	99.95
309	<.10	100

Appendix F

Factor Loadings for Specialty A, One Factor Solution

Item Number	Factor Loading
1	.18
2	.26
3	.01
4	.28
5	.00
6	.01
7	.01
8	.16
9	***.31***
10	.01
11	.13
12	.13
13	***.30***
14	***.42***
15	***.42***
16	.23
17	.20
18	.12
19	.12
20	.13
21	.00
22	.15
23	.12
24	.15
25	.26
26	.00
27	.20
28	.20
29	.18
30	.16

Appendix F – Continued

Item Number	Factor Loading
31	.24
32	.26
33	.21
34	.17
35	.22
36	.14
37	.17
38	.17
39	.27
40	.01
41	***.38***
42	.01
43	.20
44	.18
45	.01
46	***.35***
47	.20
48	.19
49	.12
50	.00
51	.11
52	.14
53	.01
54	.22
55	.11
56	.20
57	.18
58	.11
59	.14
60	.20

Appendix F – Continued

Item Number	Factor Loading
61	.17
62	.00
63	.14
64	.12
65	.26
66	.13
67	.25
68	.01
69	.19
70	.00
71	.25
72	***.34***
73	.20
74	.10
75	.12
76	.01
77	.22
78	.00
79	.15
80	.23
81	.21
82	.01
83	.28
84	.10
85	.14
86	.18
87	.10
88	.12
89	.14
90	.19

Appendix F – Continued

Item Number	Factor Loading
91	.12
92	.01
93	.01
94	.24
95	.23
96	.13
97	***.39***
98	.22
99	.14
100	.22
101	.11
102	***.34***
103	.21
104	.29
105	.27
106	.24
107	.18
108	.18
109	***.34***
110	.18
111	***.30***
112	.15
113	.01
114	.17
115	.16
116	***.45***
117	.01
118	***.32***
119	.29
120	.20

Appendix F – Continued

Item Number	Factor Loading
121	.12
122	.25
123	.17
124	.01
125	.24
126	.01
127	.28
128	.25
129	.21
130	.17
131	.22
132	.01
133	.16
134	.12
135	.15
136	.00
137	.15
138	.17
139	.16
140	.00
141	.00
142	.00
143	.13
144	.10
145	***.31***
146	.29
147	.01
148	.26
149	.21
150	.00

Appendix F – Continued

Item Number	Factor Loading
151	.25
152	.25
153	***.39***
154	.27
155	.19
156	.22
157	***.44***
158	.-.01
159	.26
160	***.34***
161	.00
162	.01
163	.18
164	.19
165	.01
166	***.31***
167	.17
168	.15
169	.00
170	***.34***
171	.11
172	.18
173	.12
174	.22
175	.25
176	.01
177	.01
178	.16
179	.12
180	.26

Appendix F – Continued

Item Number	Factor Loading
181	.00
182	.13
183	.15
184	.12
185	***.38***
186	.21
187	.27
188	***.33***
189	.20
190	.01
191	***.43***
192	***.36***
193	.16
194	.13
195	.17
196	***.31***
197	.00
198	.15
199	.01
200	.19
201	.24
202	.24
203	.29
204	.23
205	.23
206	.13
207	***.36***
208	.17
209	.24
210	.01

Appendix F – Continued

Item Number	Factor Loading
211	.24
212	***.32***
213	.18
214	.18
215	.18
216	.17
217	.10
218	.19
219	.16
220	.16
221	***.35***
222	.17
223	.14
224	.11
225	.01
226	.01
227	.24
228	.01
229	.01
230	.01
231	.12
232	***.42***
233	.18
234	.31
235	.13
236	.22
237	.26
238	.28
239	***.36***
240	.16

Appendix F – Continued

Item Number	Factor Loading
241	.01
242	.18
243	.24
244	.00
245	.01
246	.22
247	.20
248	.01
249	.00
250	.01
251	.17
252	.29
253	.20
254	.20
255	.17
256	.25
257	.20
258	.00
259	.20
260	***.31***
261	.18
262	.18
263	.26
264	.21
265	***.32***
266	.22
267	.20
268	.21
269	.28
270	.25

Appendix F – Continued

Item Number	Factor Loading
271	.21
272	.22
273	.15
274	.17
275	.18
276	.12
277	.18
278	***.32***
279	.27
280	.22
281	.27
282	.13
283	.28
284	.29
285	.13
286	.01
287	.15
288	.00
289	***.34***
290	***.30***
291	.28
292	.13
293	.29
294	.19
295	.12
296	.18
297	.14
298	.19
299	.23
300	.13

Appendix F – Continued

Item Number	Factor Loading
301	.15
302	.27
303	.01
304	.23
305	.15
306	***.34***
307	.26
308	.17
309	.14

Appendix G

Explained Variance for Specialty B, All Factors

Factor	Percentage of Variance	Cumulative Percentage
1	7.48	7.48
2	1.37	1.37
3	.879	9.73
4	.741	10.47
5	.719	11.19
6	.695	11.89
7	.646	12.53
8	.640	13.17
9	.621	13.79
10	.612	14.40
11	.584	14.99
12	.583	15.57
13	.580	16.15
14	.576	16.72
15	.569	17.30
16	.562	17.86
17	.561	18.41
18	.546	18.97
19	.544	19.51
20	.540	20.05
21	.538	20.59
22	.537	21.12
23	.532	21.66
24	.527	22.18
25	.523	22.71
26	.516	23.22
27	.513	23.74
28	.511	24.25
29	.509	24.75
30	.506	25.26

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
31	.502	25.76
32	.497	26.26
33	.495	26.76
34	.489	27.24
35	.487	27.73
36	.484	28.22
37	.482	28.70
38	.477	29.17
39	.475	29.65
40	.473	30.12
41	.471	30.59
42	.465	31.06
43	.464	31.52
44	.462	31.99
45	.460	32.45
46	.458	32.90
47	.452	33.35
48	.450	33.80
49	.447	34.25
50	.445	34.70
51	.444	35.14
52	.440	35.58
53	.438	36.02
54	.433	36.45
55	.432	36.88
56	.431	37.31
57	.430	37.74
58	.427	38.17
59	.423	38.60
60	.420	39.02

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
61	.418	39.43
62	.417	39.85
63	.415	40.27
64	.414	40.68
65	.412	41.09
66	.409	41.50
67	.407	41.91
68	.404	42.31
69	.401	42.71
70	.399	43.11
71	.396	43.51
72	.394	43.90
73	.392	44.29
74	.390	44.68
75	.389	45.07
76	.388	45.46
77	.386	45.85
78	.384	46.23
79	.383	46.61
80	.381	47.00
81	.378	47.37
82	.376	47.75
83	.375	48.12
84	.373	48.50
85	.372	48.87
86	.370	49.24
87	.366	49.60
88	.365	49.97
89	.364	50.33
90	.363	50.70

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
91	.361	51.06
92	.356	51.41
93	.355	51.77
94	.353	52.12
95	.351	52.47
96	.350	52.82
97	.349	53.17
98	.347	53.52
99	.345	53.86
100	.343	54.21
101	.342	54.55
102	.340	54.89
103	.339	55.23
104	.338	55.57
105	.336	55.90
106	.334	56.24
107	.332	56.57
108	.330	56.90
109	.329	57.23
110	.327	57.55
111	.325	57.88
112	.324	58.20
113	.322	58.52
114	.320	58.84
115	.319	59.16
116	.317	59.48
117	.316	59.79
118	.315	60.11
119	.314	60.42
120	.312	60.74

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
121	.311	61.05
122	.311	61.36
123	.309	61.67
124	.308	61.98
125	.308	62.28
126	.304	62.59
127	.302	62.89
128	.300	63.19
129	.298	63.49
130	.297	63.79
131	.295	64.08
132	.294	64.37
133	.292	64.67
134	.291	64.96
135	.289	65.25
136	.287	65.53
137	.286	65.82
138	.285	66.10
139	.283	66.39
140	.282	66.67
141	.281	66.95
142	.280	67.23
143	.278	67.51
144	.277	67.79
145	.276	68.06
146	.276	68.34
147	.274	68.61
148	.272	68.88
149	.270	69.15
150	.269	69.42

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
151	.268	69.69
152	.267	69.96
153	.266	70.22
154	.264	70.49
155	.263	70.75
156	.261	71.01
157	.260	71.27
158	.258	71.53
159	.257	71.79
160	.255	72.04
161	.254	72.30
162	.253	72.55
163	.252	72.80
164	.251	73.05
165	.249	73.30
166	.248	73.55
167	.247	73.80
168	.245	74.04
169	.244	74.29
170	.243	74.53
171	.242	74.77
172	.240	75.01
173	.240	75.25
174	.238	75.49
175	.235	75.72
176	.235	75.96
177	.233	76.19
178	.233	76.43
179	.233	76.66
180	.230	76.89

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
181	.229	77.12
182	.229	77.35
183	.228	77.57
184	.228	77.80
185	.227	78.03
186	.225	78.25
187	.223	78.48
188	.222	78.70
189	.220	78.92
190	.220	79.14
191	.218	79.36
192	.218	79.57
193	.217	79.79
194	.215	80.00
195	.214	80.22
196	.213	80.43
197	.212	80.64
198	.210	80.85
199	.209	81.06
200	.208	81.27
201	.207	81.48
202	.206	81.68
203	.205	81.89
204	.203	82.09
205	.203	82.30
206	.202	82.50
207	.199	82.70
208	.199	82.90
209	.198	83.09
210	.198	83.29

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
211	.196	83.49
212	.194	83.68
213	.193	83.87
214	.193	86.07
215	.192	84.26
216	.189	84.45
217	.189	84.64
218	.188	84.83
219	.187	85.01
220	.185	85.20
221	.185	85.38
222	.184	85.57
223	.182	85.75
224	.181	85.93
225	.181	86.11
226	.180	86.29
227	.178	86.47
228	.177	86.65
229	.176	86.82
230	.175	87.00
231	.174	87.17
232	.173	87.34
233	.173	87.52
234	.171	87.69
235	.171	87.86
236	.171	88.03
237	.170	88.20
238	.169	88.37
239	.167	88.54
240	.167	88.70

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
241	.164	88.87
242	.163	89.03
243	.163	89.19
244	.162	89.36
245	.161	89.52
246	.161	89.68
247	.160	89.84
248	.158	90.00
249	.156	90.15
250	.156	90.31
251	.155	90.46
252	.154	90.62
253	.153	90.77
254	.152	90.92
255	.150	91.01
256	.150	91.22
257	.149	91.37
258	.148	91.52
259	.147	91.66
260	.146	91.81
261	.144	91.95
262	.143	92.01
263	.143	92.24
264	.141	92.38
265	.140	92.52
266	.139	92.66
267	.138	92.80
268	.137	92.93
269	.137	93.01
270	.136	93.21

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
271	.135	93.34
272	.133	93.48
273	.133	93.61
274	.132	93.74
275	.130	93.87
276	.130	94.00
277	.129	94.13
278	.129	94.26
279	.128	94.39
280	.127	94.51
281	.125	94.64
282	.125	95.76
283	.123	94.89
284	.123	95.01
285	.123	95.13
286	.121	95.25
287	.121	95.37
288	.120	95.49
289	.119	95.61
290	.118	95.73
291	.116	95.84
292	.115	95.96
293	.115	96.07
294	.114	96.19
295	.113	96.30
296	.112	96.41
297	.112	96.53
298	.110	96.64
299	.109	96.74
300	.108	96.85

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
301	.108	96.96
302	.107	97.07
303	.106	97.17
304	.105	97.28
305	.104	97.38
306	.103	97.49
307	.102	97.59
308	.102	97.70
309	.100	97.79
310	.100	97.89
311	<.10	97.99
312	<.10	98.08
313	<.10	98.18
314	<.10	98.27
315	<.10	98.37
316	<.10	98.46
317	<.10	98.55
318	<.10	98.64
319	<.10	98.73
320	<.10	98.82
321	<.10	98.91
322	<.10	99.00
323	<.10	99.01
324	<.10	99.17
325	<.10	99.25
326	<.10	99.33
327	<.10	99.41
328	<.10	99.49
329	>.10	99.57
330	<.10	99.65

Appendix G – Continued

Factor	Percentage of Variance	Cumulative Percentage
331	<.10	99.72
332	<.10	99.79
333	<.10	99.87
334	<.10	99.94
335	<.10	100
336	>.10	100

Appendix H

Factor Loadings for Specialty B, One Factor Solution

Item Number	Factor Loading
1	.36
2	.28
3	.39
4	.00
5	.17
6	.01
7	.16
8	.29
9	.19
10	.22
11	.46
12	.24
13	.15
14	.32
15	.38
16	.35
17	.27
18	.27
19	.15
20	.43
21	.10
22	.43
23	.11
24	.15
25	.13
26	.22
27	.13
28	.14
29	.27
30	.35

Appendix H – Continued

Item Number	Factor Loading
31	.13
32	.23
33	.46
34	.32
35	.12
36	.01
37	.01
38	.18
39	.28
40	.34
41	.01
42	.30
43	.35
44	.00
45	.23
46	.28
47	.11
48	.27
49	.47
50	.01
51	.35
52	.34
53	.22
54	.25
55	.29
56	.48
57	.28
58	.20
59	.12
60	.18

Appendix H – Continued

Item Number	Factor Loading
61	.10
62	.20
63	.35
64	.28
65	.17
66	.24
67	.29
68	.27
69	.45
70	.26
71	.12
72	.31
73	-.17
74	.21
75	.17
76	.29
77	.18
78	.48
79	.52
80	.42
81	.28
82	.33
83	.41
84	.20
85	.00
86	.33
87	.01
88	.11
89	.14
90	.34

Appendix H – Continued

Item Number	Factor Loading
91	.36
92	.19
93	.43
94	.29
95	.28
96	.29
97	.40
98	.17
99	.00
100	.40
101	.27
102	.24
103	.45
104	.21
105	.18
106	.23
107	.11
108	.34
109	.21
110	.25
111	.14
112	.01
113	.42
114	.23
115	.13
116	.29
117	.27
118	.22
119	.30
120	.36

Appendix H – Continued

Item Number	Factor Loading
121	.17
122	.20
123	.01
124	.29
125	.21
126	.23
127	.14
128	.30
129	.27
130	.34
131	.45
132	.25
133	.26
134	.16
135	.01
136	.25
137	.01
138	.17
139	.34
140	.00
141	.41
142	.36
143	.00
144	.29
145	.13
146	.39
147	.24
148	.21
149	.28
150	.17

Appendix H – Continued

Item Number	Factor Loading
151	.20
152	.30
153	.01
154	.20
155	.19
156	.14
157	.32
158	.25
159	.01
160	.00
161	.16
162	.27
163	.29
164	.11
165	.21
166	.24
167	.65
168	.00
169	.39
170	.10
171	.00
172	.15
173	.19
174	.23
175	.35
176	.31
177	.42
178	.11
179	.-.01
180	.22

Appendix H – Continued

Item Number	Factor Loading
181	.52
182	.23
183	.38
184	.15
185	.36
186	.14
187	.24
188	.37
189	.15
190	.16
191	.25
192	.22
193	.21
194	.01
195	.29
196	.29
197	.29
198	.00
199	.33
200	.28
201	.00
202	.22
203	.13
204	.00
205	.43
206	.32
207	.00
208	.14
209	.00
210	.39

Appendix H – Continued

Item Number	Factor Loading
211	.42
212	.17
213	.23
214	.27
215	.13
216	.37
217	.20
218	.22
219	.01
220	.21
221	.26
222	.14
223	.26
224	.46
225	.24
226	.26
227	.42
228	.23
229	.27
230	.23
231	.28
232	.22
233	.30
234	.01
235	.01
236	.38
237	.01
238	.22
239	.18
240	.37

Appendix H – Continued

Item Number	Factor Loading
241	.33
242	.19
243	.36
244	.18
245	.26
246	.28
247	.34
248	.22
249	.28
250	.35
251	.27
252	.36
253	.24
254	.38
255	.32
256	.36
257	.35
258	.22
259	.29
260	***.44***
261	.00
262	.29
263	.31
264	.10
265	.23
266	.23
267	.30
268	.27
269	.39
270	.19

Appendix H – Continued

Item Number	Factor Loading
271	.36
272	.01
273	.16
274	.27
275	.26
276	.26
277	.15
278	***.32***
279	.41
280	.22
281	.01
282	.20
283	.29
284	.33
285	.40
286	.21
287	.25
288	.25
289	***.42***
290	***.35***
291	.11
292	.42
293	.00
294	.24
295	.14
296	.20
297	.29
298	.45
299	.59
300	.21

Appendix H – Continued

Item Number	Factor Loading
301	.18
302	.26
303	.29
304	.01
305	.50
306	.29
307	.25
308	.19
309	.51
310	.33
311	.29
312	.38
313	.01
314	.24
315	.33
316	.29
317	.01
318	.16
319	.49
320	.29
321	.51
322	.38
323	.35
324	.41
325	.30
326	.14
327	.37
328	.00
329	.20
330	.43

Appendix H – Continued

Item Number	Factor Loading
331	.28
332	.28
332	.28
333	.00
334	.23
335	.01

BIBLIOGRAPHY

Abroz, K. G., & Chan, S. B. (2002). Correlation of the USMLE with the emergency residents' in-service exam. *Academic Emergency Medicine, 9*(5), 480.

Accreditation Council for Graduate Medical Education (ACGME). (1997). Residency Review Committee section, retrieved from http://www.acgme.org.

Aleamoni, L. M. (1973). Effects of the size of sample eigenvalues, observed communalities, and factor loadings. *Journal of Applied Psychology*, *58*(2), 266–269.

Alexander, G. W., Davis, W. K., Yan, A. C., & Fantone, J. C. (2000). Following medical school graduates into practice: Residency directors' assessments after the first year of residency. *Academic Medicine*, *75*(10), S15-S17.

American Board of Medical Specialties. Available at www.ABMS.org.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1988). Validity: an evolving concept. In R. Wainer & H.I. Brauns (Eds.), *Test validity* (pp. 19–32). Hillsdale, NJ: Lawrence Erlbaum, Associates, Inc.

Basco, W. T., Gilbert, G. E., Chessman, A. W., & Blue, A. V. (2000). The ability of a medical school admission process to predict clinical performance and patient's satisfaction. *Academic Medicine, 75*, 743–747.

Berk, R. A. (1986). Minimum competency testing: Status and potential. In James V. Mitchell, Jr. (Ed.), *The Future of Testing, Vol. 2.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Berner, E. S., Brooks, C. M,, & Erdmann, J. B. (1993). Use of USMLE to select residents. *Academic Medicine*, *68*(10), 753–759.

Blue, A. V., Gilbert, G. E., Elam, C. L., & Basco, W. T. (2000). Does institutional selectivity aid in the prediction of medical school performance? *Academic Medicine*, *75*(10), S31–S33.

Bonder, B. R. (1989). Planning the initial version. *Physical and Occupational Therapy in Pediatrics* *9*(1), 15–42.

Boudreaux, E. D., Perret, N., Mandry, C. V., Broussard, J., & Wood, K. (2002). The relation between standardized test scores and evaluation of medical competence among emergency medicine residents. *Academic Emergency Medicine*, *9*(5), 495.

Bruno, G. R., Su, M., & Lucchesi, M. (1999). Predictors of EM residency applicants' success early in residency. *Academic Emergency Medicine, 6*(5), 412.

Bryman A., & Cramer, D. (1990). *Quantitative data analysis for social scientists.* New York: Routledge

Callahan, C. A., Erdmann, J. B., Hojat, M., Veloski, J. J., Rattner, S., Nasca, T. J., & Gonnella, J. S. (2000). Validity of faculty ratings of students' clinical competence in core clerkships in relation to scores on licensing examinations and supervisors' ratings in residency. *Academic Medicine, 75*, 71S–73S.

Carrothers, R. M., Gregory, Jr., S. W., & Gallagher T. J. (2000). Measuring emotional intelligence of medical school applicants. *Academic Medicine*, *75*, 456–463.

Cattell, R. B. (1978). *The scientific use of factor analysis.* New York: Plenum Press.

Char, D. M., Chapman, D. M., & Benenson, R. (2002). Enhancing resident recruitment by better predicting who will be the best residents through blinded, structured interviews with critical incident scenarios. *Academic Emergency Medicine, 9*(5), 538–539.

Choi, N., Fuqua, D. R., & Griffin, B. W. (2001). Exploratory analysis of the structure of scores from the multidimensional scales of perceived self-efficacy. *Educational and Psychological Measurement*, *61*(3), 475–489.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*(2), 276–279.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In R. Wainer & H. I. Brauns (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum, Associates, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Cureton, E. E. (1950). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council in Education.

Custers, E. J., Stuyt, P. M., & DeVries-Robbe, P. F. (2000). Clinical problem analysis (CPA): A systematic approach to teaching complex medical problem solving. *Academic Medicine*, *75*(3), 291–297.

Dear, G. E., & Roberts, C. M. (2000). The Holyoake codependency index: Investigation of the factor structure and psychometric properties. *Psychological Reports*, *87*, 991–1002.

Dills, C. R. (1998). The table of specifications: A tool for instructional design and development. *Educational Technology*, *38*(3), 44–51.

Dowaliby, F. J., & Andrew, B. J. (1976). Relationships between clinical competence ratings and examination performance. *Journal of Medical Education*, *51*, 181–188.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, NJ: Prentice Hall.

Fields, S. A., Morris, C., Toffler, W. L., & Keenan, E. J. (2000). Early identification of students at risk for poor academic performance in clinical clerkships. *Academic Medicine*, *75*(10), 78S–80S.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299.

Forsythe, G. B., McGaghie, W. C., & Friedman, C. P. (1986). Construct validity of Medical clinical competence measures: A multitrait-multimethod matrix study using confirmatory factor analysis. *American Educational Research Journal*, *23*, 315–336.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, *27*(2), 197–222.

Gonnella, J. S., & Hojat, M. (1983). Relationship between performance in medical school and postgraduate competence. *Journal of Medical Education*, *58*, 679–685.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gray, B. T. (2001). A factor analytic study of the substance abuse subtle screening Inventory (SASSI). *Educational and Psychological Measurement*, *61*(1), 102–118.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427–439.

Hall, F. R. & Bailey, B. A. (1992). Correlating students' undergraduate science gpa, their MCAT scores, and the academic caliber of their undergraduate college with their first-year academic performances across five classes at Dartmouth Medical School. *Academic Medicine*, *67*(2), 121–123.

Hall, M. L., & Stocks, M. T. (1995). Relationship between quantity of undergraduate science and preclinical performance in medical school. *Academic Medicine*, *70*, 230–235.

Hecht, K. A. (1979). Current status and methodological problems of validating Professional licensing and certification exams. In M. A. Bunda & J. L. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 16–27). Washington, DC: National Council on Measurement in Education.

Hojat, M., Erdmann, J. B., Veloski, J. J., Nasca, T. J., Callahan, C. A., Julian, E., & Peck, J. (2000). A validity study of the writing sample section of the medical college admission test. *Academic Medicine*, *75*(10), S25–S27.

Hojat, M., Veloski, J. J., & Zeleznik, C. (1985). Predictive validity of the MCAT for students with two sets of scores. *Academic Medicine*, *60*, 911–918.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement*, (p. 486). Phoenix, AZ: Oryx Press.

Jensen, A. R. (1998). *The g factor*. Westport, CT: Praeger Publishers.

Johnson, J. H., Null, C., Butcher, J. N., & Johnson, K. N. (1984). Replicated item level Factor analysis of the full MMPI. *Journal of Personality and Social Psychology, 47*(1), 105–114.

Jones, R. F., & Thomae-Forgues, M. (1981). A factor comparison of old and new MCAT scales. *Journal of Medical Education, 56*(3), 161–166.

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37*(8), 911–918.

Kane, M. T. (1986). The future of testing for licensure and certification examinations. In James V. Mitchell, Jr. (Ed.), *The Future of Testing, Vol. 2,* (p. 146). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, J. O., & Mueller, C. W. (1978a). *Introduction to factor analysis: What it is and how to do it.* Beverly Hills: Sage Publications.

Kim, J. O., & Mueller, C. W. (1978b). *Factor analysis: Statistical methods and practical procedures.* Beverly Hills: Sage Publications.

Kline, P. (1994). *An easy guide to factor analysis.* London: Routledge.

Knight, J. L. (2000, November). Toward reflective judgment in exploratory factor analysis decisions: Determining the extraction method and number of factors to retain. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY. (ERIC Document Reproduction Service No. ED 449 224).

Kulinna, P. H., & Silverman, S. (1999). The development and validation of scores on a measure of teachers' attitudes towards teaching physical activity and fitness. *Educational and Psychological Measurement, 59*(3), 507–517.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation and the Health Professions, 17*(2), 178–197.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Glenview, IL: Scott-Foresman.

Linn, R. L. (1979). Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and Problems in Competency-Based Education,* (p. 111). Washington, DC: National Council on Measurement in Education.

Lowe, P. A., & Reynolds, C. R. (2000). Exploratory analyses of the latent structure of anxiety among older adults. *Educational and Psychological Measurement, 60*(1), 100–116.

Lunz, M. E., Stahl, J. A., & James, K. (1989). Content validity revisited: Transforming job analysis data into test specifications. *Evaluation and the Health Professions, 12*(2), 192–206.

McCarthy, L., & Archer, R. P. (1998). Factor structure of the mmpi-a content scales: Item-level and scale-level findings. *Journal of Personality Assessment, 71*(1), 84–97.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and Evaluation in Education and Psychology*. Fort Worth: Harcourt Brace College Publishers.

Merenda, P. F. (1997). A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid. *Measurement and Evaluation in Counseling and Development, 30*, 156–164.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 13–103). New York: Educational Council on Education and Macmillan.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning And consequences of measurement. In R. Wainer & H. I. Brauns (Eds.), *Test validity* (p. 33–45). Hillsdale, NJ: Lawrence Erlbaum, Associates, Inc.

Messick, S. (1980). Test validity and the ethic of assessment. *American Psychologist, 35*(11), 1012–1027.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement (2nd ed.)*, (pp. 335–366). Phoenix, AZ: Oryx Press.

Mitchell, K. J., & Molidor, J. B. (1986). Factor structure of the MCAT and pilot essay. *Educational and Psychological Measurement, 46*, 1019–1027.

Moran, J. R., Fleming, C. M., Somervell, P., & Manson, S. M. (1999). Measuring bicultural identity among American Indian adolescents: A factor analytic study. *Journal of Adolescent Research, 14*(4), 405–426.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: implications or performance assessment. *Review of Education Research, 62*(3), 229–258.

Murdock, M., Kressin, N., Fortier, L, Giuffre, P. A., & Oswald, L. (2001). Evaluating the psychometric properties of a scale to measure medical students' career-related values. *Academic Medicine*, *76*(2), 157–165.

Nelson, D. S. (1994). Job analysis for licensure and certification examinations: science or politics? *Educational Measurement: Issues and Practice, 13*(3), 29–35.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill, Inc.

Osman, A., Gutierrez, P. M., Downs, W. R., Kopper, B. A., Barrios, F. X., & Haraburda, C. M. (2001). Development and psychometric properties of the student worry questionnaire-30. *Psychological Reports*, *88*, 277–290.

Pajares, F., & Urdan, T. (1996). Exploratory factor analysis of the mathematics anxiety scale. *Measurement and Evaluation in Counseling and Development*, *29*, 35–47.

Pottinger, P. S. (1979). Competence testing as a basis for licensing: Problems and prospects. In M. A. Bunda & J. L. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 28–47). Washington, DC: National Council on Measurement in Education.

Ramsey, P. G., Carline, J. D., Inui, T. S., Larson, E. B., LoGerfo, J. P., & Wenrich, M. D. (1989). Predictive validity of certification by the American Board of Internal Medicine. *Annals of Internal Medicine*, *110*, 719–726.

Raymond, M. R. (1995, April). Job analysis and the development of test specifications for licensing and certification. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA. (ERIC Document Reproduction Service No. ED 430 095).

Riordan, C. M., & Weatherly, E. W. (1999). Defining and measuring employees' Identification with their work groups. *Educational and Psychological Measurement*, *59*(2), 310–324.

Ronai, A. K., Golmon, M. E., Shanks, C. A., Shafer, M. F., & Brunner, E. A. (1984). Relationship between past academic performance and results of specialty in-training examinations. *Journal of Medical Education*, *59*(4), 341–344.

Schoenfeldt, L. F. (1984). The status of test validation research. In S. N. Elliott & J. V. Mitchell (Eds.), *Social and Technical Issues in Testing: Implications for Test Construction and Usage*, (p. 68). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shah, A. (1985). Interpreting factor analysis: Some issues and illustrations. *Indian Journal of Social Work, 46*(3), 369–387.

Shepard, L. A. (1992). Evaluating test validity. *Review of Research in Education, 19*, 405–451.

Smith, S. R. (1993). Correlations between graduates' performance as first-year residents and their performance as medical students. *Academic Medicine, 68*, 633–634.

Spearman, C. E. (1927). *The Abilities of Man: Their Nature and Measurement.* London: The Macmillan Company.

Stapleton, C. D. (1997, January). Basic concepts in exploratory factor analysis (EFA) as a tool to evaluate score validity: A right-brained approach. Paper presented at the Annual Meeting of the Southeast Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED 407 419).

Streiner, D. L. (1994). Figuring out factors: The use and misuse of factor analysis. *Canadian Journal of Psychiatry, 39*(3), 135–140.

Strong, G. (1995). A survey of issues in item writing in language testing. *Thought Currents in English Literature, 68*, 281–312.

Tamblyn, R., Abrahamowicz, M., Brailovsky, C., Grand'Maison, P., Lescop, J., Norcini, J., Girard, N., & Haggerty, J. (1998). Association between licensing examination scores and resource use and quality of care in primary care practice. *Journal of the American Medical Association, 280*(11), 989–996.

Tamblyn, R. (1994). Is the public being protected? Prevention of suboptimal medical practice through training programs and credentialing examinations. *Evaluation and the Health Professions, 17*, 198–221.

Thompson, B., & Daniel, G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*(2), 197–208.

Thorndike, R. L. (1982). *Applied Psychometrics.* Boston: Houghton-Mifflin.

Thorndike, R. M. (1978). *Correlational procedures for research.* New York: Gradner.

Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling Psychology research. *Journal of Counseling Psychology, 34*, 414–424.

Velicer, W. F., & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research, 22*, 193–210.

Veloski, J. J., Callahan, C. A., Xu, G., Hojat, M., & Nash, D. B. (2000). Prediction of students' performances on licensing examinations using age, race, sex, undergraduate GPAs, and MCAT scores. *Academic Medicine, 75*(10), S28–S30.

Vosti, K. L., Bloch, D. A., & Jacobs, C. D. (1997). Relationship of clinical knowledge to months of clinical training among medical students. *Academic Medicine, 72*, 305–307.

Wolf, F. M. (2000). Lessons to be learned from evidence-based medicine: Practice and promise of evidence-based medicine and evidence-based education. *Medical Teacher, 3*, 251–259.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*(8), 10–14.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432–442.