A PSYCHOMETRIC EVALUATION OF SCRIPT CONCORDANCE TESTS FOR MEASURING CLINICAL REASONING

Adam Benjamin Wilson

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Anatomy and Cell Biology,
Indiana University

June 2013

Accepted by the faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.

Gary R. Pike, Ph.D., Chair

Aloysius J. Humbert, M.D.

Doctoral Committee

James J. Brokaw, Ph.D.

May 8, 2013

Mark F. Seifert, Ph.D.

© 2013

Adam Benjamin Wilson

ALL RIGHTS RESERVED.

DEDICATION

I dedicate this work to my wife Megan. Thank you for your unwavering support, encouragement, enduring love, and for making countless sacrifices so that I could see to fruition my life-long dream. I am glad to have had you by my side throughout this process and look forward to the forthcoming adventures we will embark on together. I also dedicate this dissertation to my family. I am thankful for your love and grateful that you believed in and nurtured the growth of my ambitions. I hope the completion of this degree has made you proud.

ACKNOWLEDGMENTS

With the sincerest of gratitude, I would like to thank several of the faculty of Indiana University for their guidance and support in helping me to achieve my academic goals.

I would first like to acknowledge my program director Dr. James Brokaw for conceptualizing and establishing the anatomy education track at IUSM. Had this program not come into existence, I likely would have never pursued a terminal degree. I feel Dr. Brokaw and others intimately involved have truly melded together the perfect heterogeneous curriculum that blends the rigors of biomedical science with the theories and practice of education. I hope this program will continue to flourish and produce outstanding graduates who will go forth to make meaning contributions to anatomy and medical education at large. I would also like to thank Dr. Brokaw for his willingness to share his perspective and unique insight and for his diligence in providing timely, thoughtful feedback.

I must extend a warm hearted thank you to my dissertation chair, Dr. Gary Pike. Not only was he an outstanding teacher that further developed my knowledge and interest in applied statistics, but he was a solid mentor and chair. His commitment to education and experience with educational research is impressive and I am fortunate to have studied under him. I only wish we had had more time together, because I know there are a number of statistical procedures I would still like to learn.

Without the work and involvement of Dr. Aloysius Humbert this dissertation on script concordance tests would not exist. I appreciate Dr. Humbert for allowing me to expand on his research and am grateful for his involvement in this project and throughout my graduate career. I also thank Dr. Humbert for being the sounding board for my research ideas, no matter how unrealistic they were.

I would also like to acknowledge Dr. Mark Seifert, whom I consider a noteworthy role model. Dr. Seifert's commitment to teaching excellence is inspiring. I hope to one day be as knowledgeable and purposeful in my teaching as he is on a daily basis.

Without realizing it, time and time again, he reminded me that as educators we must always be at our best and must strive to continue in our own development and education. The late author and teacher Howard Hendricks once wrote, “I would rather have my students drink from a running stream than a stagnant pool.” Dr. Seifert regularly modeled this philosophy, and for that I am grateful.

Lastly, I am pleased to have worked with Dr. Laura Torbeck during my graduate studies at IUSM. Although she was not directly involved in this project, she served as an influential mentor who opened my eyes to the world of surgical education and graduate medical education. I cannot thank her enough for steering me and advising me through the job search process. I appreciate all she has done to ensure I have a successful start to my academic career.

ABSTRACT

Adam Benjamin Wilson

A PSYCHOMETRIC EVALUATION OF SCRIPT CONCORDANCE TESTS

FOR MEASURING CLINICAL REASONING

Purpose: Script concordance tests (SCTs) are assessments purported to measure clinical data interpretation. The aims of this research were to (1) test the psychometric properties of SCT items, (2) directly examine the construct validity of SCTs, and (3) explore the concurrent validity of six SCT scoring methods while also considering validity at the item difficulty and item type levels.

Methods: SCT scores from a problem solving SCT (SCT-PS; n=522) and emergency medicine SCT (SCT-EM; n=1040) were used to investigate the aims of this research. An item analysis was conducted to optimize the SCT datasets, to categorize items into levels of difficulty and type, and to test for gender biases. A confirmatory factor analysis tested whether SCT scores conformed to a theorized unidimensional factor structure. Exploratory factor analyses examined the effects of six SCT scoring methods on construct validity. The concurrent validity of each scoring method was also tested via a one-way multivariate analysis of variance (MANOVA) and Pearson's product moment correlations. Repeated measures analysis of variance (ANOVA) and one-way ANOVA tested the discriminatory power of the SCTs according to item difficulty and type.

Results: Item analysis identified no gender biases. A combination of moderate model-fit indices and poor factor loadings from the confirmatory factor analysis suggested that the SCTs under investigation did not conform to a unidimensional factor structure. Exploratory factor analyses of six different scoring methods repeatedly revealed weak factor loadings, and extracted factors consistently explained only a small portion of the total variance. Results of the concurrent validity study showed that all six

scoring methods discriminated between medical training levels in spite of lower reliability coefficients on 3-point scoring methods. In addition, examinees as MS4s significantly ($p<0.001$) outperformed their MS2 SCT scores in all difficulty categories. Cross-sectional analysis of SCT-EM data reported significant differences ($p<0.001$) between experienced EM physicians, EM residents, and MS4s at each level of difficulty. When considering item type, diagnostic and therapeutic items differentiated between all three training levels, while investigational items could not readily distinguish between MS4s and EM residents.

Conclusions: The results of this research contest the assertion that SCTs measure a single common construct. These findings raise questions about the latent constructs measured by SCTs and challenge the overall utility of SCT scores. The outcomes of the concurrent validity study provide evidence that multiple scoring methods reasonably differentiate between medical training levels. Concurrent validity was also observed when considering item difficulty and item type.

Gary R. Pike, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Chapter One: Introduction

SCTs	Script concordance tests

Chapter Two: Literature Review

SCTs	Script concordance tests
OSCE	Objective structured clinical examination

Chapter Three: Experimental Design and Methods

SCT-PS	Problem solving script concordance test
SCT-EM	Emergency medicine script concordance test
MS2	Year two undergraduate medical student
MS4	Year four undergraduate medical student
CFA	Confirmatory factor analysis
RMSEA	Root mean square error of approximation
CFI	Comparative fit index
GFI	Goodness-of-fit index
EFA	Exploratory factor analysis
IUSM	Indiana University School of Medicine
ANOVA	Analysis of variance
MANOVA	Multivariate analysis of variance
KMO	Kaiser-Meyer-Olkin

Chapter Four: Findings and Discussion

SCT-PS	Problem solving script concordance test
SCT-EM	Emergency medicine script concordance test
MS2	Year two undergraduate medical student
MS4	Year four undergraduate medical student
CFA	Confirmatory factor analysis
EFA	Exploratory factor analysis
IUSM	Indiana University School of Medicine
ANOVA	Analysis of variance
MANOVA	Multivariate analysis of variance
KMO	Kaiser-Meyer-Olkin
SLS	Schmid-Leiman solution

Chapter Five: Conclusions

SCT-PS	Problem solving script concordance test
SCT-EM	Emergency medicine script concordance test
MS2	Year two undergraduate medical student
MS4	Year four undergraduate medical student
CFA	Confirmatory factor analysis
EFA	Exploratory factor analysis
SLS	Schmid-Leiman solution

Appendices: No abbreviations listed.

CHAPTER ONE

Introduction

Introduction

Overview of Research Questions, Hypotheses, and Specific Aims

Rationale and Significance

INTRODUCTION

Educators and investigators who are passionate about understanding the cognitive and developmental processes that mature throughout medical student and residency training have been somewhat perplexed by the complexity and elusivity of clinical reasoning. In particular, devising ways to objectively measure clinical reasoning has proven challenging. One instrument that has shown promise in assessing a component of clinical reasoning is the script concordance test (SCT). This instrument purportedly measures data interpretation abilities of examinees and gives insight into the organization of examinees' illness scripts (i.e., interconnected schema of illnesses, clinical features, memories, etc.). The following chapters and sections of this dissertation will further explore the psychometric properties and nuances of SCTs to advance the depth and breadth of SCT research. Specifically, this work (1) evaluated basic psychometric properties of SCT items, (2) directly assessed the construct validity of SCTs, and (3) explored the concurrent validity of six SCT scoring methods while also considering concurrent validity at the item difficulty and item type levels.

The item analysis adds insight into the reliability and fairness of SCTs. The construct validation study is the first to be reported in the SCT literature; it uses structural equation modeling techniques to empirically identify latent SCT constructs. Lastly, the need to explore the properties of various SCT scoring methods and to test concurrent validity from the perspective of item difficulty and item type emerged out of a secondary research question that was introduced as a consequence of the construct validation study. In its own right, the concurrent validity study is unique in that it compares three previously tested scoring methods to three methods that are absent or observed less frequently in the literature, and delves deeper to understand the discriminatory nature of SCTs at levels other than the test (i.e., composite score) level.

Overall, this dissertation is composed of five chapters and four appendices. Chapter one presents an overview of the three research questions and highlights their significance in medical education research. Chapter two provides a review of the

literature pertinent to the topics of script concordance tests and clinical reasoning. The third chapter describes the study design and methods used to accomplish this work. The contents of chapter four focus on the presentation, interpretation, and discussion of the research findings. Lastly, chapter five explores the impact of this research, offers recommendations for future studies, and provides a synopsis of this work in its entirety.

OVERVIEW OF RESEARCH QUESTIONS, HYPOTHESES, AND SPECIFIC AIMS

Research Question 1	What are the psychometric properties of script concordance test items?
Hypothesis 1	Script concordance test items will have moderate to high discrimination indices, discernible ranges of difficulty, and will demonstrate consistency across traits (e.g., examinee gender).
Specific Aim 1	This project aims to test and understand the psychometric properties of script concordance test items.
Research Question 2	To what extent does the factor structure of script concordance tests conform to the theory of the measure?
Hypothesis 2	A confirmatory factor analysis will reveal that script concordance tests measure a single clinical reasoning construct, data interpretation.
Specific Aim 2A	This project aims to determine how well a single construct model represents script concordance test scores.
Specific Aim 2B	This project aims to evaluate the relationships between the identified construct(s) and pre-existing empirical and theorized clinical reasoning models.
Research Question 3	How well do non-traditional SCT scoring methods, compared to 5-point aggregate scoring, differentiate between stages of medical training development, and to what extent are

	discriminatory differences heighted or lessened by considering item difficulty and item type?
Hypothesis 3	Non-traditional SCT scoring methods will closely reflect the properties of conventional methods, and at the level of item difficulty and item type, SCTs will retain their ability to differentiate between training levels.
Specific Aim 3	This project aims to compare non-traditional to conventional SCT scoring methods and to evaluate the ability of SCTs to retain their discriminatory power at the item difficulty and item type levels.

RATIONALE AND SIGNIFICANCE

The script concordance test (SCT), originally called the Diagnosis Script Questionnaire, was first developed in 1998 by the French-Canadian scholar Bernard Charlin.[1] Since its inception, a variety of healthcare fields and medical disciplines have shown interest in the SCT because of its properties to reliably assess clinical reasoning.[2-13] It is problem-solving centric and thought to more accurately reflect professional reality.[14] Despite the wealth of educational and psychometric research conducted on this instrument, pockets of underexplored SCT domains still remain, particularly in understanding how and why this instrument works. The overarching goal of this research was to report substantive evidence for or against the theorized construct validity of SCTs.

Research Question 1

What are the psychometric properties of script concordance test items?

Rationale. Fournier et al.[15] contend that SCTs can be optimized at the question and case level by item analysis. In good practice, item analyses are regularly performed to generate shorter more reliable instruments and to guide test development and maintenance. While such outcomes are advantageous and worth pursuing, in this

research an item analysis was utilized to optimize the available SCT datasets and to test construct biases.

Research Question 2

To what extent does the factor structure of script concordance tests conform to the theory of the measure?

Rationale. In a concurrent validity study, Seibert et al. inferred that the SCT "measures a dimension for which, as one should expect, experienced clinicians get better scores than less experienced subjects."[14] Others have similarly reported SCTs are intended to assess one dimension of clinical competence –data interpretation– in the context of clinical uncertainty,[1,2] though little evidence beyond anecdotal claims confirms this assertion. Content, concurrent, and convergent validity research does, to some degree, provide supplemental, yet indirect, evidence that SCTs measure a single construct[2,3,7-9,16] different from those constructs measured by knowledge-laden multiple choice exams. According to Lubarsky et al., "the SCT's claim to probe clinical data interpretation as an isolated construct requires more empirical substantiation."[2] The lack of sufficient evidence has prompted authors of recent literature reviews to propose a strategic SCT research agenda. At the forefront of this agenda is understanding whether examinees' thought and response processes align with intended constructs, as presently this ideology is based largely on theory.[2] A 2012 article by Dory et al. further affirms the need for additional SCT research in areas related to educational impact and inter-assessment correlations.[17] Using a multi-trait multi-method matrix to compare dimensions between like instruments would be one approach to answer Dory's call; though, it must first be made evident which dimensions (i.e., traits or constructs) SCTs measure.

Inferences concerning the unidimentionality of SCTs have been periodically reported and used to implicitly authenticate the instrument's construct validity. However, to the author's knowledge, no direct dimensional analysis has been conducted. This study will, therefore, perform a confirmatory factor analysis on SCT data to directly

explore whether the instruments' structure conforms to theorized assertions. Utilizing independent datasets permits instant replication of this work to verify study findings. This project will lay the necessary groundwork for future studies aimed at assessing dimensional correlations between SCTs and other established instruments and may reveal fundamental information pertinent to best scoring practices.

Research Question 3

How well do non-traditional SCT scoring methods, compared to 5-point aggregate scoring, differentiate between stages of medical training development, and to what extent are discriminatory differences heighted or lessened by considering item difficulty and item type?

Rationale. Concurrent validity of SCTs has been demonstrated on several accounts in a variety of settings.[3,7-9,16] In a thorough SCT review, Lubarsky et al.[2] wrote, "Script concordance hinges on an inference that examinees with more evolved illness scripts will interpret data and make decisions that increasingly concord with those of experts given the same clinical scenarios." It was therefore reasonable to postulate that non-traditional SCT scoring methods would closely reflect the discriminant properties exhibited by conventional methods and that SCTs, at more refined levels, would show similar discriminant characteristics.

While one benefit of SCTs is their ability to reliably distinguish between medical training levels, this trait is thought to be largely a consequence of the aggregate scoring approach.[18,19] Some even contest the aggregate scoring method altogether and advocate for single-best-answer scoring.[20] In an attempt to settle the controversy of aggregate versus consensus scoring, Bland et al. ran statistical analyses on five different scoring keys for marking SCTs. It was found that 5-point and 3-point aggregate scoring keys were similar, as reliability values were nearly identical and correlations of scores against levels of training were statistically significant and moderate in magnitude.[20] Results suggested that 5-point scaling systems added very little information and 3-point scales were sufficient. Three-point scoring methods that accounted for differences in distance

from either the mean or modal response were reasonably reliable and effective at distinguishing between levels of experience.[20] However, the study by Bland et al. did not explore all scoring method possibilities and their work was limited by moderate to small sample sizes. Theoretically, the best SCT scoring method(s) will exhibit multiple characteristics of sound validity, of which construct and concurrent validity are included. As readers will soon learn, one outcome of the present research was that no single scoring method demonstrated superior qualities of construct validity. Taking this information into account and considering that Bland's work did not test all plausible scoring solutions, it was of interest to assess the concurrent validity properties of the six scoring methods used in this research. Investigating the concurrent validity of these methods may help to discern if administrators' preferences for or against certain scoring solutions are more or less arbitrary and may shed light on whether it is valuable to have an SCT scoring approach that measures examinees' responses in terms of both direction and degree of impact.

Sensitive SCT instruments with sound psychometric properties, ostensibly, should be able to detect an increase in data interpretation abilities as experience is gained. However, the literature concerning data interpretation is cloudy as it has been alleged, "the ability to integrate and interpret data is independent of experience."[21] One analysis designed to evaluate the contributions of data interpretation errors to misdiagnosis using 'clinical reasoning problems' was inconclusive.[21] The outcomes reported that data interpretation errors increased with experience and, on difficult problems, general practitioners made more data interpretation mistakes than hypothesis generation mistakes. In part, the inability of 'clinical reasoning problems' to fully capture the essence and complexities of expert reasoning was to blame for inconclusive findings.[21] In another related study, Chimowitz et al.,[22] who investigated confirmed diagnostic errors in neurologists and neurology residents, reported no significant differences among junior residents, senior residents, and staff neurologists. Upon diagnostic error analysis, reasoning errors (of which data interpretation is an

underlying component) were thought to be a contributing factor to diagnostic mistakes, yet error rates were similar across training levels.[22]

If data interpretation does not improve with experience then how is it that SCT scores sensibly discriminate between training levels as demonstrated by previous concurrent validity findings? At face value, previous research findings and concurrent validity evidence appear to contradict one another. While it is not the intent of this work to deliberately debate whether data interpretation skills are a function of experience, this manuscript does take the position that some instruments or methodologies are not designed to measure data interpretation gains or disparities as well as others; hence, the presence of conflicting, seemingly illogical perspectives in the literature.

The outcomes of specific aim 3 may be partly dependent upon the construct findings from specific aim 2. It is plausible that SCTs measure more than one dimension of clinical reasoning. If this is true, training level disparities may not fully be explained by global differences in the ability to interpret data alone. Instead, the discriminatory power of SCTs may be better explained by differences in other, yet to be determined, measurable constructs or by pronounced differences in identified sub-constructs.

In review, the purpose of specific aim 3 was twofold. First, the objective was to test the concurrent validity properties of six SCT scoring conditions, three of which have been given little consideration to date. To understand the factors (or constructs) that drive the discriminatory power of SCTs, this study also assessed concurrent validity at the level of item difficulty and item type. In contrast to previously published works, data from two SCTs, that comprised a broader range and number of participants, were utilized.

Significance. Research on SCTs is far from complete. The call for additional studies in the areas of educational impact, inter-assessment correlations, and construct validity has been recently voiced.[2,17] The outcomes of this research contribute to the medical education literature on multiple levels. Broadly, this project serves as a stepping stone to accomplish more rigorous SCT and diagnostic reasoning research. Specifically,

this work (1) discusses the outcomes of an SCT construct validation study, (2) solidifies our understanding of various SCT scoring methods, (3) enhances the breadth of the SCT concurrent validity literature, and (4) offers practical implications as well as specific recommendations for future psychometric research.

CHAPTER TWO

Literature Review

Introduction

Overview of Clinical Reasoning Assessments

Script Theory

Script Concordance Test Structure

Script Concordance Test Scoring

Reliability and Validity of Script Concordance Tests

Advantages and Disadvantages of Script Concordance Tests

A Review of Clinical Reasoning Research

Summary

INTRODUCTION

The purpose of this research was threefold. First this project explored the psychometric properties of script concordance tests (SCTs) and investigated the presence of certain construct biases. Secondly, this research sought to understand the extent to which the factor structure of SCTs conformed to the theory of the measure. Thirdly, this work compared non-traditional scoring methods to conventional methods and appraised the discriminatory power of SCT according to item difficulty and type.

This review will highlight how the relationships between theory and empirical evidence have helped to shape the development of SCTs and current state of SCT research. The limitations and barriers of past diagnostic reasoning instruments will be explored to give context to the SCT and to provide rationale for its widespread popularity as a measure of clinical diagnostic reasoning. The concept of script theory will be discussed to lay the foundation for the theoretical model that will be referenced in this research. The exam structure and traditional scoring method of SCTs will be described in detail as aspects of this unique instrument warrant specialized modifications to statistical procedures. Moreover, traditional aggregate scoring will serve as the baseline by which to compare competing scoring methods. At the heart of this review, reliability and validity evidence will be analyzed to demonstrate how this project directly fills current research gaps. Finally, a broad overview of pertinent clinical reasoning research will be presented to provide additional context and perspective to the significance of this work.

OVERVIEW OF CLINICAL REASONING ASSESSMENTS

Some of the first attempts to assess clinical reasoning were carried out in the 1960s with, what are generically termed, “written simulation problems”. One early example was the ‘patient management problem’ in which patient issues were presented and examinees were tasked with selecting the appropriate history, physical exam, and/or investigation items from a list.[23] Due to the lack of reliability and validity on

multiple accounts, this instrument was soon abandoned.[23] Not long after the disappearance of patient management problems, the 'key feature' approach made its début. Key feature assessments demonstrated improved reliability, yet validity parameters remained an area of concern, as the intermediate effect (i.e., no observable difference in performance scores between developmental levels) persisted.[23,24] In the early 1980s, Feltovich and Barrows were among the first to pair the concept of mental "scripts" with clinical reasoning.[25] As the concept of script theory matured, Bernard Charlin developed the first rendering of the SCT in 1998 called the Diagnostic Script Questionnaire.[1] To the liking of many, SCT scores simultaneously measured one's level of reasoning competence and clinical experience.[1,4,26] SCTs continue to gain recognition due to their ability to reliably assess clinical reasoning and, to date, the majority of data gathered from this measure supports the ideology of script theory.

SCRIPT THEORY

Script theory stems from the roots of cognitive psychology. Psychologists define scripts as "schema-like representations that provide mental frameworks for proceduralized knowledge";[27] that is, practice knowledge that can be readily transferred and used in a variety of situations and settings. Script theory can be positioned within what cognitive theorists call a connectionist model of memory (also called a parallel distributed processing (PDP) model). This model promotes the ideology that cognitive tasks require parallel, as opposed to serial, processing because human cognitive systems operate in the face of numerous stressors and constraints. The concept of parallel processing implies that information is managed along multiple dimensions simultaneously, despite the presence of impediments like ambiguity in clinical reasoning.[27] Competing memory models view information and knowledge as being statically stored in the form of patterns. In the connectionist model, however, units of information and knowledge are not stored, but rather the strengths of the connections between units are stored.[28] Connection weights within or between networks become the

key feature that allows for information input to be linked with information output.[27] It is these linkages that help to transform clinical findings into coherent schemes that inadvertently trigger diagnostic notions.[29] Associations between clinical features, illnesses, and therapeutic options are the crossbars in an ornate scaffolding scheme that assembles together to form coherent structures called illness scripts.

In the context of clinical medicine, it has been postulated that experienced clinicians have more elaborate networks than novices which align with tasks they regularly perform.[23] Expert scripts are optimally organized so as to perform complex reasoning tasks (e.g., hypothesis generation, diagnosis, data interpretation, and treatment planning) with very little effort.[14] Scripts, for example, may be compartmentalized as groups of diseases, structured around conceptual models, and stored as representational memories of observed or experienced conditions.[30] Experts are known to amass memories of patient encounters, especially those that are rare or unforgettable. New patients that resemble prior cases can activate linkages between memories and elicit the recall of pertinent knowledge.[31]

Clinical scripts are said to develop early in medical training when students are first exposed to clinical encounters.[32] As experience is gained, scripts mature and are refined and the cognitive process grows more efficient. For example, during a think aloud protocol interns who are concentrating on a patient with knee pain were more apt to list possible causes of pain whereas experts fleshed out the similarities and differences between relevant hypotheses. According to Bowen, the comparisons that experts consider as they rule in or out hypotheses occurs during the data-acquisition phase of diagnostic reasoning and forms the basis for focused patient questioning, performing physical exams, and ordering diagnostics.[30] When tasked to reason diagnostically, connections between scripts are activated and diagnostic patterns are re-created and brought into working memory where problem solving and clinical reasoning transpire.

SCTs are theorized to measure the richness of knowledge networks called illness scripts and the connections between them. As previously described, illness scripts can be organized as disease states, syndromes, or disorders stored by clinicians and are connected through problem representations.[33-35] When problem representations are activated, they trigger clinical memories and patterns thereby allowing pertinent knowledge to become accessible for clinical reasoning.[30] It is commonly accepted that routine judgments made during the clinical reasoning process can be probed and subsequently measured.[36] Examinees are compelled to interpret data in order to make sound clinical decisions.[23] Therefore, an SCT "probes the organization of clinical knowledge"[23] and teases out the extent to which examinees have an elaborated versus minute and dispersed organization of knowledge.[37] Put another way, SCTs are presumed to mine and mobilize illness scripts pertinent to the problem at hand.[2]

Script theory operates on the principle that experienced clinicians, as opposed to novices or intermediates, have more elaborate illness scripts that align with tasks they regularly perform within their domain.[23] As such, the capacity of an expert to clinically reason is far greater and more refined than non-experts and often transpires more intuitively than analytically.[29] One's ability to interpret clinical data, an operationalization of illness scripts, is considered a function of basic science and clinical knowledge.[38] Because experts are skilled at integrating and encapsulating biomedical with clinical knowledge,[39] for script theory to hold, it would be expected that experts would outperform novices and intermediates as a whole and at each level of item difficulty. Despite this logic, some research findings presented momentarily contest the ideology of script theory.

SCRIPT CONCORDANCE TEST STRUCTURE

The script concordance test is an assessment for measuring clinical reasoning rather than factual clinical knowledge.[14] The SCT measures how the reasoning practices of examinees compare to a panel of experienced physicians in the field under

examination. In an attempt to better measure skills of clinical competence, SCTs were developed with the intention of assessing one's capacity for data interpretation. This skill alone requires the conjoined use of an adequate fund of knowledge and keen cognitive reasoning to make informed clinical decisions.[1]

A typical SCT will contain 60 to 90 questions nested within 20 to 25 cases.[2] A generalizability D-study, for minimizing reliability error, ascertained that reliability is enhanced when fewer cases, with a mean of three questions per case, are presented. Tests comprising 15-20 cases with 2-5 nested questions each is the best combination for obtaining sufficiently high reliability estimates.[40] Development of SCT questions follows a 'key features' approach in that they are reflective of those features that physicians find most pertinent for solving commonly encountered clinical scenarios.[23] Sets of test items are preceded by brief clinical vignettes that typically provide examinees with patient histories and pertinent information. On-line SCTs permit the use of enriched images, video clips, or audio bites in place of written vignettes thereby enhancing the authenticity of the experience.[9,23] In some SCT formats, an item table divided into three columns appears beneath clinical vignettes (Figure 2.1). Within the columns, examinees are told what to consider, are given new information that must be weighed against information pre-existing, and are required to indicate the effect the new information has had on the initial item under question. For example, in the first column the phrase, "If you were thinking of…." is followed by pre-generated hypotheses, therapeutic alternatives, investigative options, or ethical considerations. The second column begins, "…and then you find…" and is followed by new information in the form of test or lab results, pre-existing conditions, additional signs or symptoms, etc. The third column beginning with a phrase such as "…the hypothesis becomes:" contains a Likert-type response scale for examinees to indicate how new information is likely to affect listed hypotheses.[2] In a single response, direction and intensity of the effect are captured.[15] Each item associated with a vignette is independent of the other items.[15] SCTs are commonly structured to measure how well examinees make diagnostic, investigative,

and therapeutic decisions,[14] though SCTs are not confined to these three categories. Below are sample questions that precede an SCT administered in the emergency medicine clerkship at Indiana University School of Medicine.

Figure 2.1: Sample SCT case and items.

Case: A 60 year-old female presents to the Emergency Department with a chief complaint of dyspnea for the last two days. She reports having dyspnea on exertion. She denies chest pain. The patient reports a past medical history of Hypertension but is non-compliant with her medication.

Exhibit A: Item assessing 'diagnostic knowledge':

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes:
1	Congestive Heart Failure	Bilateral rales on lung exam	-2 -1 0 +1 +2
2	COPD Exacerbation	No history of smoking	-2 -1 0 +1 +2

-2-Highly Unlikely; -1-Less likely than before; 0-Neither more nor less likely; +1-More likely than before; +2-Very Likely

Exhibit B: Item assessing 'investigational knowledge':

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes:
3	Chest CT PE Protocol	Normal d-dimer	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally; -1-Not useful; possibly detrimental; 0-No less or more useful; +1-Useful; +2-Absolutely Necessary

Exhibit C: Item assessing 'therapeutic knowledge':

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes:
4	SL Nitroglycerine	BP 180/100	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally; -1-Not useful or possibly detrimental; 0-Neither less nor more useful; +1-Useful; +2-Necessary or absolutely necessary

The goal of SCTs is for examinees to choose an acceptable response from among multiple suitable answers.[15] This differs from a traditional multiple choice exam where examinees select a single best answer from among factually incorrect distracters.[2] Because SCT items provide hypotheses and new clinical information, the cognitive task of generating hypotheses and seeking/gathering data are omitted. "What remains, ostensibly, is the [hypothetico-deductive] stage of data interpretation, in which the

examinee is presumed to make a decision regarding the fit of the new data with the given hypothesis".[2]

SCRIPT CONCORDANCE TEST SCORING

The SCTs that were utilized in this project employed an aggregate scoring method in which points for a given response were assigned as a function of the proportion of experts who responded in the same manner. Scores therefore reflect the degree of similarity, or concordance, between examinees and experts.[20] The amount of agreement is hypothesized to be a measure of script development in the examinee.[41] Aggregate scoring works on the assumption that experienced clinicians are experts of their domain and therefore their opinions have validity as model responses.[14] In the face of uncertainty and ambiguity, experts may not agree in their responses. Items with such divergence are said to mirror clinical reality and should not be discarded as they likely strengthen the discriminate quality of the instrument.[1] Differences in interpretation are considered clinically valuable and as such merit proportional credit.[2] However, to some, this scoring method insinuates that 'experts never err',[2] though quality control measures have been established in light of such critiques.[42] Because there is no requirement to reach a one best answer consensus, SCT items can be viewed as having multiple correct answers.[18]

Exam answer keys are constructed based on the individual responses of experts (Table 2.1).[20] A score of 1.00 for a single test item signifies that an examinee's answer aligns perfectly with that of the modal expert response. Envision that out of 30 experts, 21 marked '-1' as an answer and the remaining 9 marked '0' on a 5-point Likert-type scale of responses. The modal score in this example is calculated by taking the quotient of the number of experts who answered '-1' (e.g., n=21) and the number of experts who answered the mode response (in this case '-1'; n=21), to arrive at a maximum item score of 1.00 (i.e., $\frac{21}{21}$). Therefore, the response of an examinee who answers '-1' aligns 100%

with the most frequent expert response. Partial credit is awarded when an examinee's response aligns with experts who gave an answer different from the majority. Examinees who answered '0' on the scale of responses would receive a score of 0.43 (i.e., $\frac{9}{21}$), because the response aligns only 43% of the time with the most frequent expert response. Partial credit is awarded within reason. To prevent incompetent answers from appearing on the answer rubric and to limit a superfluity of partiality, the responses of experts who score below two standard deviations from the pooled expert panel are considered outliers and are removed.[42] If an examinee's response does not align with that of any expert's, the examinee receives a score of 0.00 for that item.[14] The raw composite SCT score is the sum of the individual item scores.

Table 2.1: Sample SCT scoring matrix.

	Answer Key Derivation					
	Question 1		Question 2		Question 3	
Answer	No. of Expert Responses	Credit Awarded	No. of Expert Responses	Credit Awarded	No. of Expert Responses	Credit Awarded
-2	0	0.00	12	0.92	0	0.00
-1	21	1.00	13	1.00	0	0.00
0	9	0.43	0	0.00	0	0.00
+1	0	0.00	5	0.38	4	0.15
+2	0	0.00	0	0.00	26	1.00

	Score Assignment						
	Question 1		Question 2		Question 3		
Student	Student Response	Points Earned (p)	Student Response	Points Earned (p)	Student Response	Points Earned (p)	Score = Σp/3
1	-1	1.00	0	0.00	+2	1.00	67%
2	+1	0.00	+2	0.00	+2	1.00	33%
3	0	0.43	+1	0.38	+1	0.15	32%
4	0	0.43	-1	1.00	-1	0.00	48%
5	-1	1.00	-1	1.00	0	0.00	67%

The general concepts of classic test theory are just as true for SCTs as they are for other objective assessments. Items of moderate difficulty are preferred to optimize discrimination between examinees and maximize score variance.[15] Measurement error (or noise) is frequently observed when high variability exists among panel experts.[15]

Multiple SCT scoring procedures have been empirically tested in an attempt to settle a lively scoring debate. Much discussion has centered on whether SCTs should

utilize an aggregate or a single-best-answer scoring approach in which consensus among experts is attained. It has been well documented that SCTs exhibit reliability and tenets of construct validity (e.g., correlations between training level and SCT scores) without intermediate effects.[1,4,6,14,23] In a study comparing consensus and aggregate SCT scoring approaches, 59% of experts did not agree with the consensus response decided by a convened expert panel.[18] As a result of this study, Charlin et al.[18] posit that a single-best-answer approach should not be used in SCT testing. Upon closer investigation of the methods employed, Bland[20] asserts that Charlin's findings may have been an artifact of the 7-point response scale used. He also disputes the interpretation of Charlin's results because of ignored statistical violations and Bland arrives at an alternate conclusion that aggregate scoring contains more random error than a single-best-answer approach.[20] Valuable questions concerning scale range are raised from this conjuncture. For example, should a 3-, 5-, or 7-point scale be used, and do 1 point shifts (e.g., +2 vs. +3) on a 7- or 5-point scale represent meaningful differences? Studies aimed at giving meaning to composite SCT scores are perhaps of greatest necessity. Because this project explores the factor structure of SCTs, it may provide some insight into better scoring practices, such as replacing a single composite score with scores for each construct SCTs are found to measure.

In an attempt to settle the controversy of aggregate versus consensus scoring, Bland et al. ran statistical analyses on five different scoring keys for marking SCTs. It was found that 5-point and 3-point aggregate scoring keys were very similar, as reliability values were nearly identical and correlations against levels of training were significant and moderate in magnitude.[20] Results suggested that 5-point scaling systems add very little discriminative information and 3-point scales are sufficient. Single-best-answer scoring, with a 3-point scale, was demonstrated to be less reliable, but held similar validity coefficients. Scoring methods that accounted for differences in distance from either the mean or modal response were reasonably reliable and effective at distinguishing between levels of experience.[20]

A major disadvantage of using either 5- or 7-point Likert-scales in traditional aggregate scoring is that test administrators cannot readily distinguish those responses that were near the modal response from those that were distant from it.[20] For instance, if the mode response of the reference panel was '+2', examinees who answer '-1' receive the same score of 0.00 as those who answered '+1' (presuming no expert answered '-1' or '+1'). It is therefore possible for examinees who agree with experts on the response direction but not the intensity (or impact) to receive the same score of 0.00 as someone who fails to identify both the direction and the impact.[20] Employing a 3-point scaling system would all together eliminate 'degree of correctness' concerns. Qualitative data of student perceptions also implies that 5- or 7-point scaling systems should be avoided, as students reported at times arbitrarily choosing between '+1' and '+2' and '-1' and '-2'.[20] In addition to concerns regarding 'degree of correctness', Bland contends that "if a single best answer to an SCT does not exist, the SCT will be of limited use for in-course assessment".[20] The rationale is that novices are expected to perform like experts, to attain the best possible score. Customarily, course assessment instruments are designed to assess specific course objectives or behaviors. Without a single best answer, it becomes difficult to define attainable objectives. The complexities and intricacies of aggregate scoring are enough for some practitioners to forgo the use of aggregate scoring entirely.

With aggregate scoring come questions of panel reliability, panel size, and panel composition. Because answer rubrics are constructed from the responses of experts, it is pertinent to question the internal consistency of the panel itself. To attain a Cronbach's alpha coefficient greater than or equal to 0.70, a panel size ranging from 10-15 members is necessary.[43] In one study, it was determined that non-teaching and teaching physicians could sit on the expert reference panel without cause for concern.[44] Prior to assembling a panel of 15-20 experts, one must first define expertise; a task that is logical but complex in practice. Fournier[15] advocates for well-rounded physicians who have had ample clinical experience in their respective fields and who, in some way, have

been able to demonstrate sound clinical judgment. It is also recommended that panel members' expertise align with the intended assessment objectives.[15] The nature of SCTs requires little effort on the part of reference panel members because panel participants are asked to think through scenarios in a format they regularly use in practice. No preparation or review of content is needed for panel members to complete the test.[15]

With debates momentarily put aside, SCT scores must first demonstrate adequate reliability and construct validity to function as decision making devices and to hold meaning. Despite numerous SCT studies, disparities in the construct validity of SCTs continue to linger.

RELIABILITY AND VALIDITY OF SCRIPT CONCORDANCE TESTS

Reliability is an indicator of the consistency and reproducibility of a measurement procedure and is a necessary precondition for validity.[45] Indices of validity indicate the extent to which test scores are representative of the inferences made from those scores.[45] More simply, validity reveals whether purportedly measured constructs are in practice measured. The term construct "refers to something that is not observable but is literally *constructed* by the investigator to summarize or account for the regularities or relationships in observed behavior".[45]

Currently, SCT constructs have not been empirically identified using factor analytic techniques. Rather, theory and indirect measures have guided the assertion that SCT scores reflect "the ability to weigh clinical information in light of entertained hypotheses".[46] Demonstrating competence in the interpretation of clinical data in scenarios of uncertainty and ambiguity is considered highly valuable in clinical reasoning practice[47] and is central to sound clinical judgment.[48] However, it is not well understood if this is what SCTs actually measure.

Five sources of validity evidence, as recommended by the *Standards for Educational and Psychological Testing*,[49] will be used as the framework to discuss past research intended to authenticate SCT validity. The five sources of construct validity are:

content, response process, internal structure, relationships to other variables, and consequences.

Content. Content validity is "the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose".[50] To measure content validity, a panel of judges rate items in categories of relevance, representativeness, specificity, clarity, and overall technical quality.[50] Items with high inter-rater reliabilities and large content validity indices are retained and utilized in the assessment instrument. Because the targeted construct of the SCT is presumed to be clinical data interpretation, the content of SCTs must be relevant to and representative of the processes that evoke data interpretation. As such, SCT items are ill-defined and genuine with inconsistencies, uncertainties, and imperfections much like clinical practice.[15] Moreover, the non-essential features of a case should be unfamiliar to learners so as not to elicit hindsight (retrospective) bias.[51] Additional guidelines for SCT construction advise that, while examinees should rely upon factual knowledge, recall of factual information alone should not be sufficient to answer SCT questions.[15] To validate SCT items, the distribution of expert responses is closely examined. SCT items that induce identical responses from all experts may be excluded as they do not satisfy the concept of being ill-defined and are likely to exhibit lower discrimination indices.[15] Similarly, items with a broad distribution of expert responses are closely reviewed to discern if they are confusing or too vague and may be subsequently removed from the test if these undesirable features are found.[2]

Response Process. Response process gathers evidence to explain the relationships between an examinee's thought processes or behaviors and an instrument's intended constructs.[2] This source of validity also examines how responses are influenced by contextual nuances. Data in support of response process validity is presently lacking, in part, because SCT constructs have not been directly explored or confirmed through statistical modeling. It has been reported, however, that the

relationships between new clinical information and activated scripts affect information processing times and the accuracy of clinical judgments.[52]

Internal Structure. Internal structure provides evidence of instrument reliability. That is, how consistently constructs are measured within a single instrument, between instruments, or over time. Many who have studied the internal structure of SCTs have reported moderate to high alpha coefficients ranging from 0.60-0.90, indicating dependable reliability.[1,3,8,9,16,53] The evidence that SCTs demonstrates robust reliability across multiple medical disciplines further supports the argument that SCTs probe a single common construct.[2]

Relationships to Other Variables. To gauge the extent to which an instrument measures targeted constructs, it is either compared to instruments that measure similar or identical constructs or is compared to unrelated instruments that measure considerably different constructs. The relationship between instruments will ideally converge or diverge. Comparing SCTs with multiple choice exams has resulted in only minor success at establishing convergent validity. A study by Collard et al.[54] reported significant positive correlations between an adapted SCT and true/false test scores for less experienced students, but not for advanced trainees. From these results the authors speculate that factual knowledge and clinical reasoning grow independent of one another as experience is gained.[54] In a 2011 study using data from emergency medicine residents, SCT scores were compared to an in-training exam, and scores from year-4 medical students were compared to the United States Medical Licensing Exam-Step 2, Clinical Knowledge (USMLE-Step 2, CK). In both instances significant positive correlations were identified with correlations among residents being higher than those among students.[16] The above studies provide limited evidence for convergent validity. More rigorous analyses (e.g., multi-trait multi-method matrix[45]) are needed to clarify the meaning of these findings.

Concurrent and predictive validity also fall within the scope of 'relationships to other variables'. A study by Seibert et al.,[14] whose focus was in urology, reported the SCT

satisfied parameters of concurrent validity. Novices, residents, and experts demonstrated significantly different levels of reasoning as theoretically expected, and SCT scores were positively correlated to training level; though the study lacked sufficient statistical power. Numerous other studies testing concurrent validity at the composite score level have found similar outcomes.[3,7-9,16]

In addition to assessing concurrent validity, the predictive power of SCTs has also been explored, but to a lesser extent. By comparing clerkship SCT scores to SCT and Objective Structured Clinical Exam (OSCE) scores at the end of residency, Brailovsky et al.[23] demonstrated that SCTs can moderately predict future SCT scores, but fail to adequately predict OSCE scores. Another work explored whether SCT performance predicted scores on a reliable and valid three part Canadian licensing examination.[46] The board exam was comprised of short-answer management problems (i.e., a measure of problem solving and clinical decision making capacity), an OSCE (for assessing clinical skills), and simulated office orals[55] (a form of standardized patient simulations for assessing social competence, problem identification, problem management, and interview processes and organization).[23] In this study, the SCT was found to be successful at predicting short-answer management problems and simulated office orals, but unsuccessful at predicting OSCE scores.[23] It was concluded that SCTs can effectively predict performance on other high stakes exams considered to measure clinical reasoning.[23] This provides additional evidence in favor of construct validity, as has been supported by other studies.[1,4,26] It also suggests that a common construct exists between the SCT and the Canadian licensing exam, excluding the OSCE. If SCTs are capable of measuring, in general, the organization of clinical knowledge, it may have additional practical applications such as identifying early reasoning deficits or as a benchmark for acceptance into residency programs.[23]

Consequences. The role of the final validity source, 'consequences,' is to gather information on the consequences of the assessment method, whether positive or negative. The impact an assessment procedure has on society, learning and teaching, the

effectiveness of the procedure's scoring format, and establishing scoring thresholds are all elements enfolded into this source of construct validity. In general, significant evidentiary gaps remain in the domain of SCT scoring. For example, little has been written on the effectiveness or meaning of SCT scoring procedures and no consensus has been reached on the pass/fail thresholds of SCTs.[2] The use of standardized scores has been proposed,[42] but this recent recommendation has been minimally adopted.[16] One benefit of SCTs is their ability to reliably distinguish between medical training levels, thought to be largely a consequence of the aggregate scoring approach.[18,19] There are some that contest the aggregate scoring method altogether and advocate for single-best-answer scoring.[20] Overall, it is clear that additional research is needed before SCTs can be employed as high stakes examinations. Of primary concern are establishing discipline specific and cross-discipline standards for optimal SCT scoring and clearly communicating what each standard represents. By directly investigating SCT constructs, this study will provide information necessary for interpreting SCT scores and may prove to be useful in establishing more appropriate scoring strategies.

ADVANTAGES AND DISADVANTAGES OF SCRIPT CONCORDANCE TESTS

Writing an SCT is thought to be straight forward and often intuitive, requiring little to no training and few resources.[14] The test can be administered in paper-based format, or easily transcribed for electronic distribution with basic exam writing software. One perk of an aggregate scoring system is that test items can be used even when consensus among experts is not reached.[14] An advantage of SCTs over other clinical reasoning instruments is the large number of items used to assess examinees. It takes roughly an hour for examinees to answer 60-90 questions and the issue of case-specificity is practically obsolete.[2]

Because SCTs focus on assessing reasoning capacity, the test is devoid of skill oriented items and items intended to measure other equally important aspects of clinical reasoning. SCTs, as a whole, are not well suited for assessing collaborative

reasoning, other aspects of hypothetico-deductive reasoning, procedural skills, cognitive errors, interviewing skills, reasoning efficiency, or physical exams. In general, current clinical reasoning assessments have several shortcomings. Traditional multiple choice exams and SCTs are often limited in that they oversimplify cases, they do not allow examinees to direct a clinical encounter or generate hypotheses,[56] and some items inadvertently cue students toward the correct response.[57] As Charlin et al.[1] state, "Clinical competence is a multidimensional entity. No single assessment method can adequately measure it." In addition to using SCTs to understand students' reasoning capacity, educators also rely on observations, document analyses, as well as patients and other healthcare providers to draw conclusions about the reasoning aptitude and general clinical performance of students.[30]

A REVIEW OF CLINICAL REASONING RESEARCH

Although a wealth of clinical reasoning information has been collected and analyzed for more than 30 years, gray areas and unsettled theoretical debates on how best to teach and assess this abstract, yet instrumental, domain still linger. This seemingly elusive phenomenon has captured the interest of a consortium of medical educators worldwide who have collectively contributed to our present understanding of this complex, multifaceted subject. Clinical reasoning is a specialized form of problem solving, sometime used interchangeably with terms like decision making or judgment.[34] The literature on clinical reasoning offers accounts and descriptions of multiple theories, models, and approaches thought to be integral components of this complex process. Examples include dual-process theory, hypothetico-deductive theory, probabilistic reasoning, pattern recognition, heuristics, expert development theory, and script theory.[34,58-60] In-depth investigations of the aforementioned theories have provided useful insight and a necessary foundational platform for further research. It is from this rich historical groundwork that medical educators can expand upon and improve widespread, yet unrefined, clinical reasoning theories, as well as explore supplementary

perspectives and assessments yet to be considered. In the context of SCTs, there is likely some degree to which the above theories comingle, but the interdependency of such theories is not fully understood. The key to unlocking the connections between examinees' thought processes (or behaviors) and how they respond to SCT items will likely be reliant on understanding the interrelated dynamics of multiple theories and concepts. As such, this section provides a brief introduction to prior clinical reasoning research and select clinical reasoning theories.

According to an early cross-sectional study, from the time medical students enter undergraduate medical programs until they are practicing on their own as trained physicians, the majority of their clinical reasoning thought processes remain relatively constant.[61] It should be noted, however, that the retrospective method used to extract the thought processes of participants in this particular study has its limitations. A different study released in 2011 that sampled students at every level within five different medical schools across the U.S. reported substantial gains in clinical reasoning performance from one academic/developmental level to another.[62] These more recent findings make it more challenging to accept that clinical reasoning thought processes do not evolve. However, clinical reasoning gains observed in the third-year were not as extensive as anticipated.[62] This particular finding, echoed by another study,[47] may suggest that students reach a reasoning plateau during their clinically intensive years. In spite of this, script concordance test data has frequently demonstrated linear correlations between reasoning performance and level of training[20] and suggests that a third-year plateau is gradually overcome through real-world clinical experience as novices continue to develop and hone their clinical expertise.

Discrepancies between the above studies are likely the product of not having well developed standardized instruments that assess multiple components of clinical reasoning. Perhaps data interpretation skills develop at a different rate and more profoundly than hypothesis generation skills. Yet without adequate measures to detect these subtle differences, such idiosyncrasies in clinical reasoning development will go

unnoticed. As an interesting yet related aside, programs that stress and apply horizontal and vertical integration of the basic and clinical sciences early within a curriculum do not yield students with superior gains in clinical reasoning, as measured by currently available instruments.[62] This finding is somewhat counterintuitive and challenges educators to explore more deeply what it means to teach clinical reasoning and promote clinical reasoning growth.

It is generally accepted that expert physicians reason through a forward reaching approach, from data to diagnosis (i.e., data to final solution), using a network of causal rules.[63] Conversely, beginners commonly reason in the direction from hypothesis(es) (i.e., tentative solutions) to data, through a backward reaching process.[63,64] Forward and backward reasoning involves both inductive and deductive reasoning strategies. Interestingly, the use of forward and backward reasoning approaches was found to be related to the amount of information available and the degree of confidence respondents had as a result of previous feedback.[63] These early observations foreshadowed two notions: (1) "…the context within which a problem is being addressed has a major impact on the accuracy of the decisions reached and the optimal balance between potential reasoning strategies"[59] and (2) psycho-affective factors (e.g., confidence, complacency, disillusionment, etc.) that influence intuition are not only present during clinical problem solving, but are vital to keep in check so as not to override the equally important fact orientated, procedurally driven analytical system, as described in the dual-process theory.[60]

It is also accepted that medical students use biomedical knowledge more so than experts to causally reason through clinical problems.[34,58] However, in rare or complex cases, experts may resort to using basic science knowledge to explain an unfamiliar phenomenon and may also arrive at their answers through a backward reaching process that compares tentative solutions to available data.[34] The findings of these studies also suggest that an expert's ability to succinctly and meaningfully organize knowledge is instrumental in developing clinical expertise, perhaps more so than one's knowledge of

basic science. To teach or practice diagnostic reasoning processes, some educators have students complete problem representations (or problem statements) that succinctly convey a patient's chief complaint and key presenting features. Inherent within problem statements are semantic qualifiers that are used as descriptors to compare and contrast diagnostic considerations. When evaluating novice versus expert problem statements, the use of descriptive semantic qualifiers was strongly associated with advanced clinical reasoning skills.[30,65]

Expert development theory has explored a plethora of factors and unique skills that experts possess over their less experienced protégées. Probing the ability of experts to minimize cognitive errors is another area that holds promise for understanding how best to address reasoning shortcomings and teach toward expertise.

Cognitive Errors. Some hold the perspective that clinical reasoning errors transpire as a result of one's inability to adequately collect appropriate, meaningful information.[66,67] Others place blame on incompetence, inadequacies in knowledge, or incorrect integration and interpretation of data.[21,68,69] However, Scott argues it is neither of these factors. Instead, he posits that human thinking is simply more fragile under conditions of uncertainty, complexity, and the demanding pressures of time.[29] "All decision making is vulnerable to different forms of cognitive and affective (emotional) bias or error."[29]

The reality and prevalence of cognitive errors is noteworthy and should not be taken lightly. In acute care settings, diagnoses can be missed or delayed 5-14% of the time.[29] At autopsy, 25% of undiagnosed cases are diagnosed, reaffirming the prevalence of diagnostic errors.[29,70] It is also estimated that roughly 45% of patients who are correctly diagnosed do not received evidence based care and as many as 30% of the tests ordered and drugs administered are unnecessary.[29,71] Furthermore, for those cases about which clinicians were certain they took the appropriate actions, 40% of fatalities were shown at autopsy to have resulted from incorrect diagnosis.[72] This is not to say that reasoning errors directly cause or impact these statistics, as some errors may be beyond

the control of the clinician. However, according to an Australian study that investigated the causes of adverse events, nearly half of recorded health care errors involved poor clinical reasoning or decision making.[73] This particular study researched human errors made by healthcare providers and was not restricted to mistakes made only by clinicians.

An initial step toward correcting cognitive and affective errors is understanding their source and how they occur. When practitioners reflect on errors they have knowingly made, they will often offer explanations in an attempt to rationalize wrong decisions. It is from these so called 'excuses' that we gain a glimpse of the many sources of cognitive errors. In some instances mental heuristics are to blame.[29] That is, common short cuts or rules of thumb do not play out as they were intended to. Internal and external biases can also affect the likelihood of committing a cognitive or emotional error. Examples of internal biases include value bias (i.e., being partial towards or against internalized values and beliefs), agency bias (i.e., placing one's own interests ahead of patients' interests), expectation bias (i.e., having a distorted perspective of the patient-doctor relationship and the expectations that accompany that relationship), and affective bias (influenced by emotion and personality).[29] Beyond intrinsic factors, external variables also shape the context in which cognitive errors are made. Social bias is one example in which the opinions of others and socialization may sway one's judgment. When time, resources, or limited skill sets affect reasoning, external bias is accused.[29] Ones mental and physical health, as well as distractions and interruptions, should not be overlooked as they can become cognitive stressors with the potential to impinge on one's cognitive integrity.[29]

Although the kinds of factors that contribute to cognitive errors are documented, many educators do not thoroughly assess/track cognitive mistakes or use such information as a teaching tool to reduce the recurrence of errors. Whether errors are caused as a result of faulty integration and data interpretation or due to external pressures, strategies and activities to correct such errors are worth exploring. The third

aim of this research may bring some reassurance that data interpretation, as measured by SCTs, is a function of experience. If this is the case, then it is logical to infer that some developmental process is taking place that moves a learner from a state of high cognitive error rates to a more experienced state of lower error rates. Therefore, this research may help to determine whether the potential for teaching clinical reasoning skills exists, and more specifically whether data misinterpretation can be overcome through experience or perhaps purposeful teaching.

SUMMARY

The works of Bernard Charlin and other instrumental educational researchers have laid the necessary groundwork on which to build more in-depth SCT and clinical reasoning studies. Despite fourteen years of rich educational investigations, sizable knowledge gaps and best-practice controversies still remain. For example, the number and nature of constructs measured by SCTs is currently based on speculation. Therefore, the principal aim of this work was to empirically test the factor structure of SCTs. In preparation to discuss the findings of this research, the bulk of this chapter reviewed SCT exam format and structure, scoring methods, and validity evidence. All in all, this research adds to the depth of medical education literature by contributing additional SCT validity evidence and adds to the breadth of literature by setting the stage for other quantitative analyses and follow-up investigations.

CHAPTER THREE

Research Design and Methods

Introduction

Data Collection and Instrumentation

Item Analysis

Construct Validation Study

Concurrent Validity Study

Summary

INTRODUCTION

This study was a large-scale retrospective data analysis that made use of SCT scores from undergraduate medical students, residents, and experienced physicians collected at Indiana University School of Medicine (IUSM). A variety of statistical procedures were performed to answer the research questions central to this project. In total, three questions and corresponding hypotheses were posed:

Research Question 1. What are the psychometric properties of SCT items?

Hypothesis 1. Script concordance test items will have moderate to high discrimination indices, discernible ranges of difficulty, and will demonstrate consistency across traits (e.g., examinee gender).

Research Question 2. To what extent does the factor structure of the script concordance test conform to the theory of the measure?

Hypothesis 2. A confirmatory factor analysis will reveal that script concordance tests measure a single clinical reasoning construct, data interpretation.

Research Question 3. How well do non-traditional SCT scoring methods, compared to 5-point aggregate scoring, differentiate between stages of medical training development, and to what extent are discriminatory differences heighted or lessened by considering item difficulty and item type?

Hypothesis 3. Non-traditional SCT scoring methods will closely reflect the properties of conventional methods, and at the level of item difficulty and item type, SCTs will retain their ability to differentiate between training levels.

This chapter presents a description of data collection methods and an account of the SCT instruments and datasets used. Subsequent sections, organized in research question order, provide an explanation of data analyses and statistical procedures employed, as well as a description of the assumptions that underlie each statistical

approach. The institutional review board of Indiana University - Purdue University Indianapolis granted this study approval under exempt status.

DATA COLLECTION AND INSTRUMENTATION

The three datasets used in this research consisted of test scores from undergraduate medical students, emergency medicine (EM) residents, and practicing board certified EM physicians who completed either a problem solving script concordance test (SCT-PS; Appendix A) and/or an emergency medicine script concordance test (SCT-EM; Appendix B) in its entirety.

The SCT-PS was administered to undergraduate medical students (n=522) at IUSM twice during their enrollment; once during year-two (as MS2s) while students were dispersed at one of nine IUSM centers and once during year-four (as MS4s) at which time all medical students studied at the main centrally located campus.

In their fourth year, students were required, per clerkship mandates, to complete an SCT-EM in emergency medicine. Scores of undergraduate medical students (n=988) comprised the majority of SCT-EM data. SCT-EM scores from EM residents (n=40) and experienced EM physicians (n=12) were also included in the analysis. EM residents (postgraduate year 1-3) participated on a voluntary basis and were not incentivized for their time. Local EM physicians comprised the reference panel and also participated voluntarily.

Instrument Blueprint. The SCT-PS, taken by MS2s and MS4s since 2008, had a total of 75 diagnostic oriented questions nested within 16 cases. From February 2008 to May 2011, the department of emergency medicine administered the SCT-EM to students on the EM clerkship. The SCT-EM was composed of 59 items nested within 12 cases. Twenty-three items were of diagnostic orientation to assess the appropriateness of predetermined hypotheses, 16 items were ‘investigational’ to assess the suitability of diagnostic tests, and 20 items evaluated the aptness of therapeutic interventions.

The SCT-PS was created by faculty of IUSM and the State University of New York-Stonybrook as an assessment of problem solving competence.[6] SCT-PS reference panel participants (n_{panel}=13) were experienced physicians from family medicine and general internal medicine. The SCT-EM was created by IUSM faculty who participated voluntarily. For the purpose of SCT-EM answer key creation, a panel of board certified EM physicians (n_{panel}=12), who had at minimum 5 years clinical experience, were recruited and utilized.[5]

Both instruments followed a traditional SCT format in which examinees responded to items using a five-point Likert-type scale ranging from '-2' to '+2'. Responses indicated the influence of new information on a given diagnosis, test, or treatment, as outlined above. Negative answer choices were associated, for example, with hypothesis elimination or test or treatment contraindication. A selection of '0' indicated neutrality. Useful and absolutely necessary information was designated '+1' and '+2', respectively. Examinees were allotted 2 hours to complete the SCT-PS and 90 minutes to complete the SCT-EM. SCT scores were initially computed using an aggregate scoring method.[15] Incomplete SCTs in which examinees failed to respond to one or more items were excluded from the study.

ITEM ANALYSIS

The first aim of this study focused on assessing the psychometric properties of SCT items. Item analyses are commonly performed to enhance instrument reliability and to reduce the number of items required to measure targeted constructs. Item-total correlations were computed to evaluate the extent to which examinees' responses to individual items were representative of, or consistent with, differences in their total test scores.[74] Item discrimination indices were calculated to isolate and subsequently discard items that demonstrated poor discriminatory power between high and low scoring examinees. Pearson's product-moment coefficient was used to evaluate the existence of gender construct bias.[74] Scores of male examinees were correlated against scores of female examinees to assess whether males and females scored equivalently on each

exam. Pearson's correlation coefficients greater than 0.85 were considered to represent a fair exam.

CONSTRUCT VALIDATION STUDY

The second aim of this research was to empirically validate the construct validity of SCTs. A confirmatory factor analysis (CFA) was performed to evaluate how well SCT scores conformed to the theory that SCTs are unidimensional (i.e., assess a single dimension of clinical reasoning competence). In general, factor analysis is a data reduction technique that can also be used to identify unobservable (latent) factors. In psychometrics, factor analyses are commonly used to provide evidence for or against construct validity. According to Cronbach and Meehl, construct validity is "a measure of some attribute or quality which is not 'operationally defined'" (i.e., measured directly).[75]

CFA is a type of structural equation model built on the foundations of theory and prior research. This work utilized CFA to test a one-factor solution as suggested by theory and inference. CFA assesses whether the parameters of a measurement model align with pre-specified or pre-existing data, in the form of a 'sample' variance-covariance matrix.[76,77] The parameters of a model generally include factor loadings, factor variance/covariance, and error/uniquenesses. CFA computes an estimate for each parameter, and uses those estimates to generate an 'implied' variance-covariance matrix. This second 'implied' matrix is an approximation of the first 'sample' matrix contrived from prior analyses or estimated via software programs. Therefore, CFA tests how well an estimated model represents sample data. In this instance, the researcher was interested in testing how well a unidimensional model represented SCT data. The process of fitting an 'implied' model to 'sample' data is iterative in nature and requires an initial fit. "LISREL uses the instrumental variables and two-stage least squares methods to compute starting values,"[76] and thereafter factor loadings were freely estimated.

Several methods exist to estimate CFA models, the most common of which is maximum likelihood. This approach works by evaluating the likelihood of each parameter (e.g., factor loadings, factor variance/covariance, and error), per a given dataset, and sets parameter values at their maximum likelihood.[76] Two main benefits to using maximum likelihood are (1) standard errors and levels of significance can be calculated and (2) model goodness-of-fit can be computed. However, before a maximum likelihood estimation can be used, three assumptions must be satisfied. First, maximum likelihood estimations require large sample sizes, often greater than 200. The SCT datasets that were analyzed in this project had sample sizes of over 500 examinees. Secondly, the maximum likelihood procedure requires that data conform to a multivariate normal distribution. To detect multivariate normality, univariate normality and the absence of outliers were verified. Histograms, normal probability plots, skewness, and kurtosis measures were also observed to confirm that SCT composite scores followed a univariate normal distribution. An evaluation of Z-scores (i.e., standardized scores) was conducted to identify the presence of extreme outliers. Examinees with Z-scores greater than 4.0 would have been treated as outliers had outliers been identified. Finally, maximum likelihood assumes that data have continuous levels of measurement. Data that is categorical, dichotomous, or ordinal cannot be used in maximum likelihood estimation. The traditional 5-point aggregate method for scoring SCTs generates scores that are "pseudo-ordinal." A true ordinal variable is one that has a definitive rank or order, yet the distance between ranks is immeasurable because "ordinal variables do not have origins or units of measurement."[78] With the traditional 5-point aggregate approach, SCT scores are pseudo-ordinal because it is permissible to have a modal response and two secondary responses with the same partial score. However, in many instances a modal response, secondary response, and tertiary response are assigned in a rank order as a function of the number of experts who responded to that item. It is imperative that ordinal variables in structural equation models be metrically transformed into variables of a continuous type. An approach

described by Jöreskog[78] was used to transform SCT item scores into interval data. *PRELIS 2* calculated variances, covariances, and correlations, as wells as estimated necessary thresholds for converting ordinal to interval measures. *LISREL 8.8* computed all confirmatory factor models.

Model goodness-of-fit was assessed on four fronts. Absolute goodness-of-fit indices were calculated with a model chi-square (χ^2).[76,77] However, because model chi-square is less dependable with large sample sizes,[77] the root mean square error of approximation (RMSEA), a parsimony correlation index, was also computed. RMSEA is less sensitive to large samples and was used to assess reasonable model fit within the population. Thirdly, a comparative fit index (CFI) was used to evaluate the fit of a conservative baseline (or 'null') model, that assumed no relationships among variables, to the solution specified by the investigator.[76] RMSEA and CFI are both insensitive to sample size effects and are robust against departures from multivariate normality.[79] A RMSEA model fit of 0.06 or less and CFI coefficients of 0.95 or greater were considered indicators of acceptable models.[76] Lastly, *LISREL's* goodness-of-fit index (GFI) for measuring the amount of mutual variance and covariance within a model was computed and compared to the conventional threshold of adequate fit (i.e., GFI greater than 0.90).[80]

Exploratory Factor Analysis. This study also utilized an exploratory factor analysis (EFA) as an alternate procedure for identifying SCT constructs. Within *SPSS* (version 20) a principle components solution with orthogonal (Varimax) rotation was used to attain simple structure for uncorrelated factors. A classical EFA that utilized tetrachoric correlations was performed in *LISREL 8.8* to explore the number of factors extracted under conditions of dichotomous (right/wrong) scoring solutions. Within *LISREL*, factors were extracted using principle components analysis and simple structure was attained via Varimax rotation. To evaluate sampling adequacy and whether sample correlation matrices were appropriate for factor analytic methods, a Kaiser-Meyer-Olkin (KMO) coefficient was computed and values greater than or equal to 0.60 were considered sufficient for conducting an EFA.[81]

Items with factor loadings less than 0.4 are considered non-significant[77,82] and as such were ignored to enhance factor interpretation. Factorially ambiguous items that exhibited high loadings on two or more factors or factorially complex items that crossed factors following construct replication were managed as needed.

The unanticipated emergent nature of this study led to the computation of a second-order factor analysis estimated via a principle components solution. Second-order factors present in the data were identified using Kaiser's criterion (i.e., Eigenvalues ≥ 1.00), and Promax rotation was used to interpret factor structure. The analysis was conducted in *SPSS* (version 20) as described by Thompson.[83] To further facilitate factor interpretation, Schmid-Leiman solutions were performed to assess the direct relationships between SCT items and higher-order factors.[84,85] This solution was used to understand how well individual SCT items measured first- versus second-order factors.

CONCURRECT VALIDITY STUDY

The third aim of this research tested the concurrent validity of various SCT scoring methods and conducted an item difficulty and item type analysis to understand the internal nature of SCTs and to explore additional uses of the instrument. A repeated measure analysis of variance (ANOVA), performed on the SCT-PS dataset, assessed whether scoring methods could longitudinally discriminate between training levels. A multivariate analysis of variance (MANOVA) tested whether all scoring methods could differentiate between training levels on the SCT-EM dataset. A repeated measures ANOVA and one-way ANOVA compared within and between training level effects on item scores, grouped by level of item difficulty. In addition, SCT-EM items were categorized into three types including diagnostic, investigational, and therapeutic items. Repeated measures ANOVA assessed differences in item types within and between training levels.

Repeated Measures. Assumptions of the repeated measures model, including homogeneity of variance and sphericity, were tested to confirm the appropriateness of

the statistical approach. In a repeated measures design, observations (e.g., scores grouped by level of difficulty) are naturally dependent within samples and independent across samples. Levene's test was computed to assess homogeneity of variance (i.e., whether the variance of populations were equal). A p-value greater than 0.05 was used to signify the assumption had been satisfied. Histograms and normal probability plots were used to detect non-normality. An additional assumption of repeated measures is sphericity, assessed via Mauchly's test. Sphericity assumes the differences of paired scores have the same variance.[86] By definition, study designs with variables containing only two levels of a repeated measure always satisfy the sphericity assumption. Because SCT scores were grouped by difficulty (i.e., easy, moderate, difficult) and item type (i.e., diagnostic, investigational, therapeutic) with three levels each, sphericity was tested.

The level of significance (alpha) was set at 0.05 and eta squared (η^2) was utilized to measure the magnitude of the reported effects. Eta squared values of 0.01, 0.06, and 0.14 were used to discern small, medium, and large effects, respectively.[86]

MANOVA. In a similar fashion, assumptions of multivariate analysis of variance (MANOVA) were tested to ensure the appropriateness of the statistical approach. The main assumptions of a one-way MANOVA are as follows: (1) Two or more dependent variables must consist of interval or ratio data. All six scoring methods, composed of percentage scores (ratio data), were included as dependent variables thereby satisfying this assumption. (2) The independent variable should consist of two or more categorical groups. In this study the independent variable was training level which included MS4s, EM residents, and experienced EM physicians. (3) The independence of observations assumption specifies that members of a group cannot hold simultaneous or dual membership in another group. (4) The sample size should be adequate in that the number of cases per group ought to exceed the number of dependent variables analyzed. Assumptions three and four were satisfied. (5) An absence of univariate or multivariate outliers is recommended. No outliers were identified. (6) Data should adhere to multivariate normal distributions. This assumption was tested and found to be satisfied

by verifying univariate normality. (7) Moderate correlations between each pair of dependent variables are compulsory. This assumption was satisfied as correlations among scoring methods were greater than or equal to 0.548. (8) The homogeneity of variance-covariance matrices was assessed using Box's M test of equality of covariance. Values greater than 0.001 indicated the assumption had been satisfied for the unbalanced research design.[87]

Item Difficulty Derivation. Item difficulty was established according to natural breaks revealed via histogram analysis of student SCT scores. For the SCT-PS, items in which partial or full credit was given to 85.00% or more of examinees were classified as easy items (n_{easy}=35). Moderate items ($n_{moderate}$=15) were answered correctly, in part or in full, by 60.00% to 84.99% of examinees. Items answered correctly, in part or in full, by 59.99% of examinees or less were labeled difficult ($n_{difficult}$=8). Histogram analysis of the SCT-EM indicated 31 easy items, 16 moderate items, and 2 difficult items. Items answered correctly, in part or in full, by 80.00% or more of student examinees were easy, between 40.00% and 79.99% were moderate, and 39.99% or less were difficult.

SCT-EM Survey. In an attempt to further explain and understand the quantitative findings of the concurrent validity study, an SCT-EM item survey (Appendix C) was distributed to emergency medicine faculty. The survey probed faculty perceptions on a random sampling of SCT-EM items. Some participants (i.e., those with an odd birth month) were asked to respond to the level of difficulty they felt items exhibited (Appendix C, Part A). Others (i.e., those with an even birth month) were asked to respond to the ambiguity level of items (Appendix C, Part B). Items evaluated by both groups of faculty were identical. Distribution of the survey and data collection was achieved via the RedCap survey system.

SUMMARY

A combination of univariate and multivariate techniques including, but not limited to, confirmatory and exploratory factor analyses and repeated measures ANOVA

were employed to answer the research questions posed. The assumptions of all statistics were carefully tested and considered to ensure accuracy and to minimize confounding effects.

CHAPTER FOUR

Findings and Discussion

Introduction

Item Analysis

Discussion: Item Analysis

Construct Validation Study

Discussion: Construct Validation Study

Concurrent Validity Study

Discussion: Concurrent Validity Study

Summary

INTRODUCTION

The culmination of previous research has failed to assess SCTs at a more refined level and has neglected to directly explore the latent dimensions of SCTs. Filling these gaps will provide additional insight into the nature of SCTs and will lay the groundwork for additional more advanced investigations. The research questions and hypotheses addressed in this section include:

Research Question 1. What are the psychometric properties of SCT items?

Hypothesis 1. Script concordance test items will have moderate to high discrimination indices, discernible ranges of difficulty, and will demonstrate consistency across traits (e.g., examinee gender).

Research Question 2. To what extent does the factor structure of the script concordance test conform to the theory of the measure?

Hypothesis 2. A confirmatory factor analysis will reveal that script concordance tests measure a single clinical reasoning construct, data interpretation.

Research Question 3. How well do non-traditional SCT scoring methods, compared to 5-point aggregate scoring, differentiate between stages of medical training development, and to what extent are discriminatory differences heighted or lessened by considering item difficulty and item type?

Hypothesis 3. Non-traditional SCT scoring methods will closely reflect the properties of conventional methods, and at the level of item difficulty and item type, SCTs will retain their ability to differentiate between training levels.

This chapter is divided into three main sections: item analysis (research question 1), construct validation study (research question 2), and concurrent validity study (research question 3). Each section is further subdivided into a results section that will report the findings of this research and a discussion section that will cover the interpretation and meaning of the outcomes.

ITEM ANALYSIS

Item-total correlations in conjunction with item discrimination indices were used to optimize the internal consistency of the SCT instruments prior to testing study hypotheses. Optimization was performed to minimize measurement error and statistical inflation induced by the instruments themselves.

The SCT-PS was optimized by discarding 17 items identified as having negative or modest (i.e., <0.100) item-total correlations and/or negative discrimination indices on both (SCT-PS-MS2 and SCT-PS-MS4) administrations of the exam (Table 4.1). Using Cronbach's alpha, the optimized 58-item SCT-PS was calculated to have a reliability of 0.745 and 0.802 for the first and second administration of the SCT-PS, respectively. SCT-EM optimization was attained by removing 10 items that demonstrated low/negative item-total correlations and/or low/negative item discrimination indices (Table 4.1). Using scores from undergraduate medical students only, the optimized 49-item SCT-EM was calculated to have a reliability of 0.556. A Pearson product-moment correlation revealed no construct biases concerning gender at any level of item difficulty for either the SCT-PS-MS2 ($r \geq 0.896$; $r^2 \geq 0.803$; $p \leq 0.003$; Figure 4.1), SCT-PS-MS4 ($r \geq 0.952$; $r^2 \geq 0.906$; $p < 0.001$; Figure 4.2) or SCT-EM exam ($r \geq 0.941$; $r^2 \geq 0.886$; $p < 0.001$; Figure 4.3).

Table 4.1: Item properties of un-optimized instruments

	SCT-PS-MS2		**SCT-PS-MS4**			**SCT-EM**	
Item	Item-total correlation coeff.	Discrim. Index	Item-total correlation coeff.	Discrim. Index	Item	Item-total correlation coeff.	Discrim. Index
Q01C01	.067	0.141	.241	0.275	~~Q31C01~~	-.006	0.081
Q02C01	.254	0.355	.295	0.314	Q32C01	.161	0.148
~~Q03C01~~	-.070	-0.012	-.040	-0.016	Q33C01	.110	0.179
Q04C01	.097	0.168	.128	0.077	Q34C01	.134	0.159
~~Q05C01~~	-.002	0.090	.075	0.077	Q35C02	.190	0.258
Q06C01	.160	0.167	.267	0.204	Q36C02	.108	0.240

~~Q07C02~~	-.051	-0.012	-.002	-0.004	Q37C02	.046	0.151
~~Q08C02~~	.016	0.037	-.032	0.051	Q38C02	.059	0.068
~~Q09C02~~	.066	0.076	.081	0.069	Q39C02	.072	0.174
Q10C02	.180	0.139	.221	0.089	Q40C03	.063	0.157
~~Q11C03~~	.060	0.079	.005	-0.013	Q41C03	.254	0.304
~~Q12C03~~	-.148	-0.096	-.113	-0.070	Q42C03	.098	0.166
Q13C03	.083	0.216	.324	0.249	Q43C03	.163	0.205
Q14C03	-.101	-0.027	.115	0.241	~~Q44C03~~	-.033	0.021
Q15C03	.189	0.224	.178	0.178	Q45C04	.024	0.133
Q16C04	.309	0.388	.336	0.343	Q46C04	.134	0.297
Q17C04	.131	0.243	.252	0.326	Q47C04	.234	0.449
~~Q18C04~~	.125	0.061	.068	0.000	~~Q48C04~~	-.040	0.056
Q19C04	.055	0.085	.114	0.042	Q49C04	.112	0.198
~~Q20C04~~	-.154	-0.090	-.240	-0.186	~~Q50C04~~	-.079	0.003
Q21C05	.216	0.189	.290	0.261	Q51C05	.109	0.211
Q22C05	.071	0.168	.214	0.347	~~Q52C05~~	.014	0.106
Q23C05	.250	0.176	.249	0.148	Q53C06	.234	0.404
Q24C05	.121	0.057	.091	0.078	Q54C06	.118	0.177
Q25C05	.260	0.338	.293	0.432	Q55C06	.046	0.145
Q26C06	.147	0.174	.086	0.058	Q56C06	.113	0.220
Q27C06	.121	0.213	.382	0.319	~~Q57C06~~	-.147	-0.052
Q28C06	.218	0.321	.170	0.272	Q58C07	.103	0.077
Q29C06	.280	0.352	.195	0.250	Q59C07	.050	0.184
Q30C06	.011	0.062	.319	0.177	Q60C07	.083	0.157
Q31C07	.346	0.407	.362	0.385	Q61C07	.122	0.251
Q32C07	.125	0.127	.187	0.159	Q62C07	.092	0.145
Q33C07	.245	0.363	.142	0.140	Q63C08	.208	0.255
Q34C07	.243	0.415	.224	0.310	Q64C08	.018	0.145
Q35C08	-.003	0.021	.140	0.165	Q65C08	.136	0.209

Q36C08	.155	0.213	.308	0.242	Q66C08	.032	0.066
Q37C08	.184	0.216	.072	0.105	Q67C08	.055	0.128
Q38C08	.215	0.310	.231	0.330	~~Q68C09~~	-.028	0.069
Q39C09	.254	0.202	.298	0.205	Q69C09	.036	0.096
Q40C09	.253	0.336	.326	0.400	Q70C09	.086	0.078
Q41C09	.209	0.218	.285	0.174	Q71C09	.098	0.166
Q42C09	-.024	0.037	.127	0.203	Q72C10	.116	0.249
Q43C10	.098	0.139	.083	0.128	Q73C10	.129	0.109
Q44C10	.193	0.237	.265	0.307	Q74C10	.048	0.149
Q45C10	.238	0.322	.158	0.163	Q75C10	.072	0.128
Q46C10	.193	0.249	.197	0.223	~~Q76C10~~	-.026	0.080
Q47C10	.308	0.365	.351	0.399	Q77C10	.069	0.150
~~Q48C11~~	-.094	-0.041	-.132	-0.069	~~Q78C11~~	.014	0.140
~~Q49C11~~	-.126	-0.070	-.109	-0.064	Q79C11	.138	0.267
Q50C11	.145	0.093	.144	0.045	Q80C11	.065	0.061
~~Q51C11~~	-.120	-0.070	-.087	-0.051	Q81C11	.047	0.051
Q52C11	.066	0.093	.069	0.105	Q82C11	.026	0.152
Q53C12	.082	0.030	.149	0.128	Q83C12	.134	0.073
Q54C12	.209	0.258	.159	0.072	Q84C12	.159	0.151
Q55C12	.263	0.401	.265	0.366	Q85C12	.152	0.243
Q56C12	.089	0.159	.238	0.241	Q86C12	.035	0.103
Q57C12	.266	0.234	.299	0.240	~~Q87C12~~	-.045	0.046
Q58C13	.224	0.373	.174	0.287	Q88C12	.104	0.089
Q59C13	.121	0.209	.345	0.274	Q89C12	.118	0.183
~~Q60C13~~	.056	0.070	.048	0.060			
Q61C13	.217	0.207	.243	0.154			
Q62C13	.285	0.262	.364	0.309			
~~Q63C14~~	.071	0.087	.070	0.069			
~~Q64C14~~	-.015	0.010	-.132	-0.078			

Q65C14	.195	0.358	.234	0.393	
Q66C14	.146	0.199	.120	0.132	
Q67C14	.227	0.408	.251	0.318	
Q68C15	.211	0.161	.047	0.064	
~~Q69C15~~	-.068	-0.043	-.095	-0.059	
Q70C15	.188	0.296	.124	0.210	
Q71C15	.194	0.356	.258	0.360	
~~Q72C16~~	-.020	-0.004	.000	-0.002	
Q73C16	.217	0.138	.182	0.135	
Q74C16	.258	0.385	.279	0.351	
Q75C16	.213	0.313	.266	0.278	

Q=question number
C=case number
~~Q##C##~~=discarded item

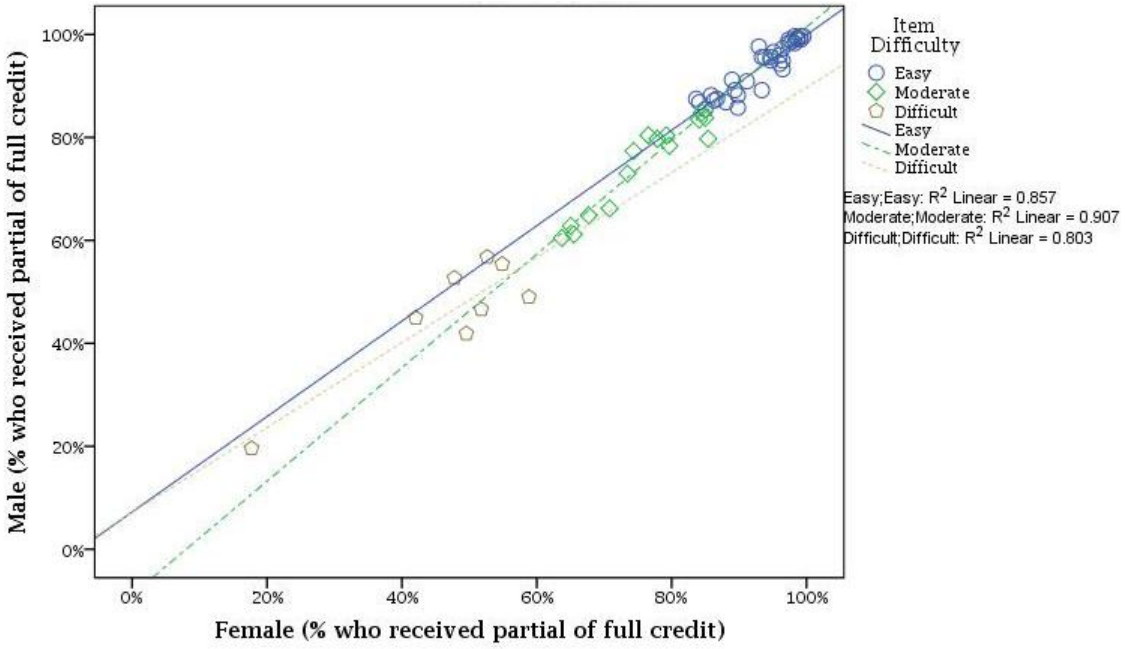


Figure 4.1: SCT-PS-MS2 Item Performance: Male vs. Female - Scatter plot of SCT-PS-MS2 items correlating the percentage of males who received partial or full credit on an item to the percentage of females who received partial or full credit on the same item. At each level of difficulty, Pearson's correlation coefficient (r) was greater than or equal to 0.896 and the coefficient of determination (r^2) explained 80.3% of the variability or greater.

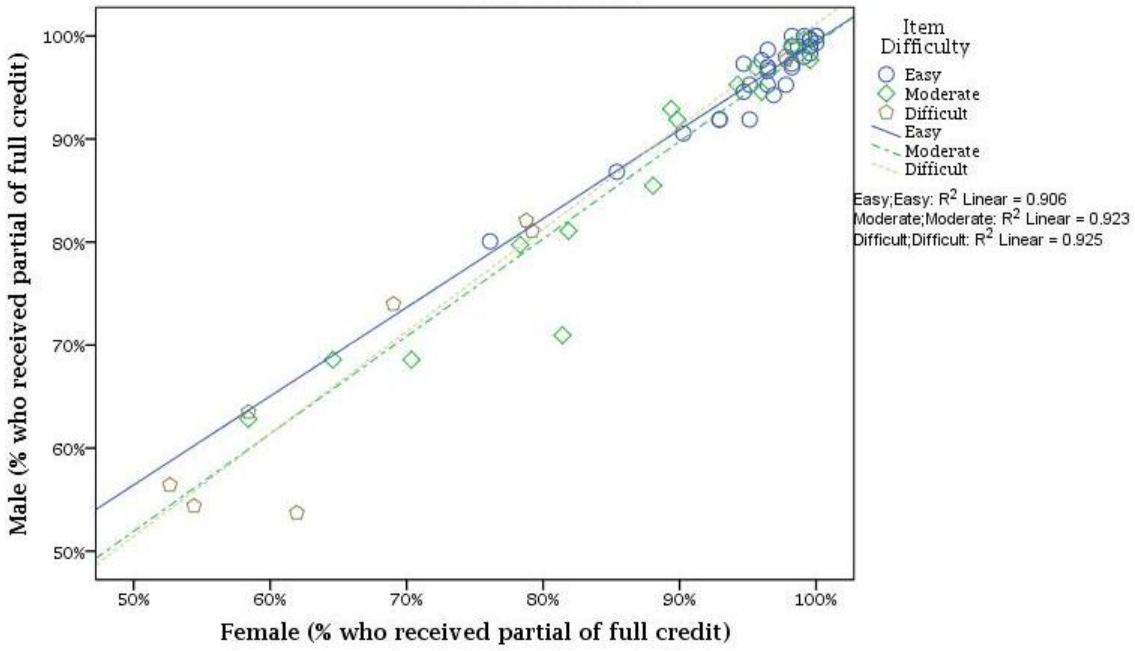


Figure 4.2: SCT-PS-MS4 Item Performance: Male vs. Female - Scatter plot of SCT-PS-MS4 items correlating the percentage of males who received partial or full credit on an item to the percentage of females who received partial or full credit on the same item. At each level of difficulty, Pearson's correlation coefficient (r) was greater than or equal to 0.952 and the coefficient of determination (r^2) explained 90.6% of the variability or greater.

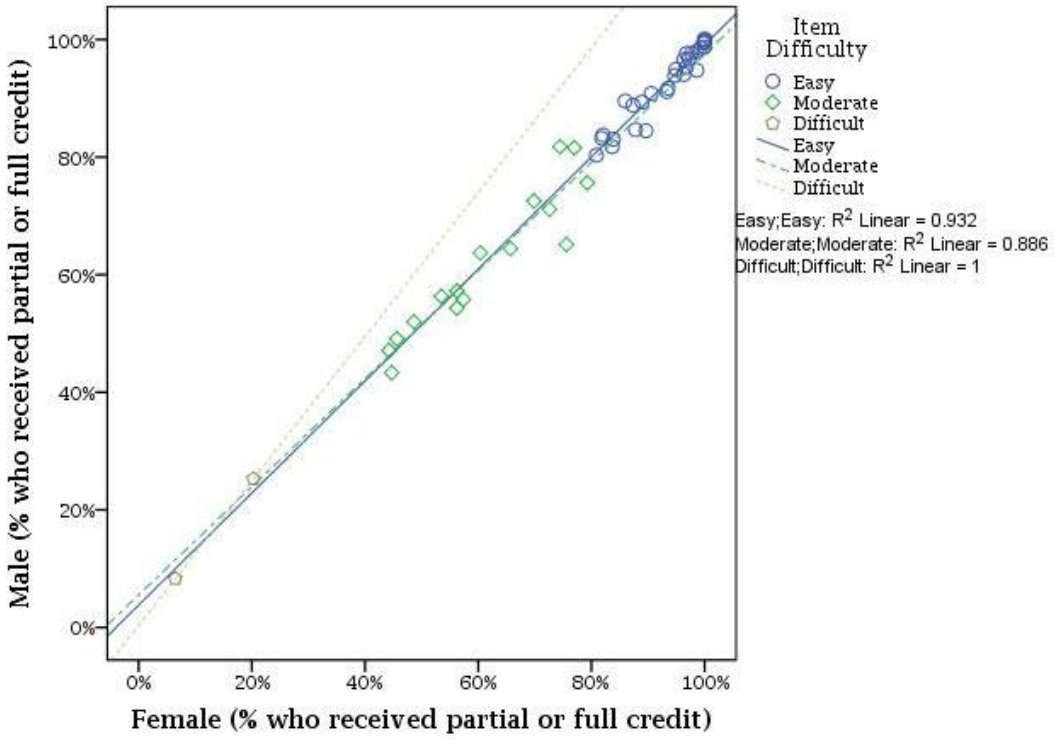


Figure 4.3: SCT-EM Item Performance: Male vs. Female - Scatter plot of SCT-EM items correlating the percentage of males who received partial or full credit on an item to the percentage of females who received partial or full credit on the same item. At each level of difficulty, Pearson's correlation coefficient (r) was greater than or equal to 0.941 and the coefficient of determination (r^2) explained 88.6% of the variability or greater.

DISCUSSION: ITEM ANALYSIS

The main purpose of testing the item properties of each instrument/dataset was to identify and discard items with poor psychometrics so that error, induced by the instruments themselves, and ultimately statistical inflation would be minimized. Not only did these findings contribute to instrument/dataset optimization, but provided evidence that the investigated SCTs had no gender biases; a sought after attribute of any test. Unfortunately, no previous studies have reported on the presence or absence of gender biases on SCTs. As such, no comparisons to other findings could be made. While potential race or ethnicity biases were not tested, I would hypothesize that other construct biases are unlikely. However, additional research is needed to confirm this.

CONSTRUCT VALIDATION STUDY

A confirmatory factor analysis using a maximum likelihood estimation was performed to test the hypothesis that SCT scores conform to a unidimensional model. All factors were freely estimated and polychoric correlations were used. Maximum likelihood estimation of optimal factor loadings revealed that SCT data did not conform to the hypothesized unidimensional model. While model-fit indices were adequate, or near to adequate, unidimensional factor loadings were weak, compromising construct interpretation (Table 4.2).

Table 4.2: Results of CFA testing unidimensionality of SCTs

Dataset	No. of items with factor loadings ≥0.4 (%)	χ^2	RMSEA	CFI	GFI
SCT-PS-MS2 (n=522)	9 of 58 (16)	5508.017 (p=0.0)	†0.0190	‡0.968	0.764
SCT-PS-MS4 (n=522)	20 of 58 (34)	5558.335 (p=0.0)	†0.0116	‡0.992	0.782
SCT-PS-EM (n=1040)	3 of 49 (6)	5663.496 (p=0.0)	†0.0318	0.851	0.818

RMSEA: Root Mean Square Error of Approximation; †indicates acceptable model at ≤ 0.06
CFI: Comparative Fit Index; ‡indicates acceptable model at ≥ 0.95
GFI: Goodness-of-Fit Index; *indicates acceptable model at > 0.90

Factor Loadings. The standard maximum likelihood solution reported mostly positive factor loadings that ranged from 0.051 to 0.550, 0.039 to 0.610, and 0.014 to 0.597 for the SCT-PS-MS2, SCT-PS-MS4, and SCT-EM, respectively. In the SCT-PS-MS2 dataset, only 9 of 58 test items (16%) loaded strongly on a single factor. The number of factor loadings increased to 20 of 58 (34%) on the SCT-PS-MS4 dataset. Only 3 of 49 items (6%) were found to have sizable loadings on the SCT-EM.

Model Fit Evaluation. The collective interpretation of model fit indices provides weak evidence to support the hypothesized 1-factor model. For all datasets, the absolute fit index (χ^2) was significant ($p<0.01$), an expected finding because chi-squared indices are less dependable with large sample sizes.[80,88] The root mean square error of approximation (RMSEA) reported that the investigator specified model represented a reasonable approximation of the data (Table 4.2). A comparative fit index (CFI) was also calculated to evaluate the fit of the conservative baseline model to the investigator specified solution. Comparative fit indices were acceptable with the SCT-EM dataset being the exception (Table 4.2). *LISREL's* goodness-of-fit index (GFI) was unsatisfactory as no dataset reported GFI values greater than 0.90.

CFA Summary and Rationale for EFA. CFAs conducted on three SCT datasets revealed that SCT items, scored via traditional aggregate scoring techniques, poorly loaded on a single factor. While a majority of model fit measures were indicative of a good fitting model, the number of items with substantive loadings on a single factor was minimal. Collectively, the above CFA findings suggest that SCTs do not measure a single first-order construct (thought to be data interpretation). At this juncture, two questions were raised: 1) Are SCTs multidimensional and, if so, what constructs do they measure? and 2) To what degree are discrete constructs a product of different SCT scoring methods? To answer these questions, the study was continued by performing exploratory factor analyses (EFAs) on each dataset under six different scoring conditions for a total of 18 analyses. Table 4.3 provides a description of the six scoring methods

employed (labeled A-F). Table 4.4 showcases a sample of recoded scores, and Table 4.5 presents a summary of EFA findings.

Table 4.3: Summary and explanation of scoring methods used for EFA

Scoring Method	Description
A. 5-point aggregate	This is the traditional aggregate scoring method that awards full credit to examinees who select the modal response from a 5-point response scale. Proportional credit is awarded when examinees' responses align with reference panel members who gave an alternate (non-modal) answer.
B. 5-point single answer	The only response for which examinees receive full credit is the modal response. No partial credit is awarded.
C. 5-point distance from mode	This scoring method renders a weighted penalty to examinees who do not give a modal response. Penalty points are a function of the number of steps examinees are away from the modal response. For example, examinees who answer -1 or +1 are 1 step from the modal response (0) and thus receive a 1 point penalty (e.g., Table 4.4, Item 3, C). Examinees who answer -2 receive a 4 point penalty when the modal response is +2 (e.g., Table 4.4, Item 1, C) Penalty points are translated into a score in which credit is awarded for being closer to the modal response. Equation: $C= 1-(\delta/\Delta)$ Where C=scoring method C; δ=distance penalty point; Δ=maximum distance from mode (e.g., 2, 3, or 4)
D. 5-point aggregate with distance penalty	This scoring method blends methods A and C. In addition, to receiving full and partial credit from traditional aggregate scoring, penalties were also calculated to account for the distance from the modal response. Instating a penalty prevents examinees who were near to the modal response from receiving the same score of 0.00 as an examinee who was distant from the modal response. For example, with traditional aggregate scoring, if the mode response of the reference panel was '+2', examinees who answer '-1' receive the same score of 0.00 as those who answered '+1' (presuming no expert answered '-1' or '+1'). With 5-point aggregate penalty scoring, if the panel mode response was '+2', examinees who answer '-1' receive a score of 0.13 and those who answered '+1' are awarded a score of 0.38 (Table 4.4, Item 1, D). Equation: D=(A+C)/2 Where A, C, and D=designated scoring methods
E. 3-point aggregate	Responses on a 5-point response scale were recoded to generate a 3-point aggregate score. Responses of +1 and +2 were condensed into a single positive score. Likewise, -1 and -2 were combined to represent a single negative score. Partial credit remained feasible to attain.
F. 3-point single answer	Scoring method E without partial credit.

Table 4.4: Scoring samples from 3 of 49 SCT-EM items

	Response Options					Student 1	
Item 1	-2	-1	0	+1	+2	Response	Score
A. 5-point aggregate	0.00	0.00	0.71	0.00	1.00	0	0.71
B. 5-point single answer	0.00	0.00	0.00	0.00	1.00	0	0.00
C. 5-point distance from mode	0.00	0.25	0.50	0.75	1.00	0	0.50
D. 5-point aggregate with distance penalty	0.00	0.13	0.61	0.38	1.00	0	0.61
E. 3-point aggregate	0.00		0.71		1.00	0	0.71
F. 3-point single answer	0.00		0.00		1.00	0	0.00

	Response Options					Student 1	
Item 2	-2	-1	0	+1	+2	Response	Score
A. 5-point aggregate	0.10	1.00	0.10	0.00	0.00	-2	0.10
B. 5-point single answer	0.00	1.00	0.00	0.00	0.00	-2	0.00
C. 5-point distance from mode	0.67	1.00	0.67	0.33	0.00	-2	0.67
D. 5-point aggregate with distance penalty	0.38	1.00	0.38	0.17	0.00	-2	0.38
E. 3-point aggregate	1.00		0.09		0.00	-2	1.00
F. 3-point single answer	1.00		0.00		0.00	-2	1.00

	Response Options					Student 1	
Item 3	-2	-1	0	+1	+2	Response	Score
A. 5-point aggregate	0.00	0.00	1.00	0.33	0.00	+2	0.00
B. 5-point single answer	0.00	0.00	1.00	0.00	0.00	+2	0.00
C. 5-point distance from mode	0.00	0.50	1.00	0.50	0.00	+2	0.00
D. 5-point aggregate with distance penalty	0.00	0.25	1.00	0.42	0.00	+2	0.00
E. 3-point aggregate	0.00		1.00		0.33	+2	0.33
F. 3-point single answer	0.00		1.00		0.00	+2	0.00

	Item Number						Student 1	
Composite Score	1		2		3		Σ Points (Σp)	Score = (Σp/3)(100)
A. 5-point aggregate	0.71	+	0.10	+	0.00	=	0.81	27.00%
B. 5-point single answer	0.00	+	0.00	+	0.00	=	0.00	0.00%
C. 5-point distance from mode	0.50	+	0.67	+	0.00	=	1.17	39.00%
D. 5-point aggregate with distance penalty	0.61	+	0.38	+	0.00	=	0.99	33.00%
E. 3-point aggregate	0.71	+	1.00	+	0.33	=	2.04	68.00%
F. 3-point single answer	0.00	+	1.00	+	0.00	=	1.00	33.33%

EFA on Scoring Methods. Kaiser-Meyer-Olkin (KMO) coefficients for evaluating sampling adequacy were found to be adequate at ≥ 0.60. Inter-factor correlations were small to non-existent (except where noted in Table 4.5). For each of the three datasets (SCT-PS-MS2, SCT-PS-MS4, and SCT-EM) Kaiser's criterion (i.e., factors with Eigenvalues ≥ 1.00) suggested extracting 19 or more factors. However, Catell's Scree test and investigation of non-redundant residuals provided additional information that helped to reduce extracted factors to a more parsimonious number. Overall a range of 3-5 factors were extracted for the SCTs under investigation. The amount of total variance explained was less than 50.00% and rarely did the majority of items load on the extracted factors.

Examination of factor loading patterns across SCTs and all scoring methods revealed that items nested within cases were frequently distributed across various first-order factors. That is, all items within a single case rarely loaded on the same factor (Table 4.6). Because "differences in item response level might spuriously produce difficulty factors,"[89] an analysis of the percentage of easy, moderate, and difficult items per factor was conducted. Outcomes suggested that items did not load according to difficulty (Table 4.6). On the SCT-EM, items were categorized as diagnostic, investigational, or therapeutic questions. Therefore, the proportion of diagnostic versus investigational versus therapeutic questions per factor was also explored. Under the condition of scoring method D (5-point aggregate with distance penalty) and assuming a 3 factor model, 100% (4 of 4) of SCT-EM items that loaded significantly on the first factor were classified, a priori, as diagnostic questions. Similarly, 66.67% (2/3) of items were investigational on factor two, and 100% (3 of 3) of items on factor three represented therapeutic questions. This factor structure is logical and was consistent across other scoring methods. However, because the number of salient loadings per factor was few, further analyses on equivalently structured SCTs, that yield more items with significant loadings per factor, are needed to confirm this result.

Table 4.5: Summary of EFA results

Scoring Method A: 5-point aggregate

	SCT-PS-MS2	SCT-PS-MS4	SCT-EM
Number of extracted factors	5	3	4
Total variance explained by extracted factors	19.28%	15.55%	15.54%
No. of items with factor loadings ≥ 0.4 (%)	16 of 58 (28)	17 of 58 (29)	11 of 49 (22)
Cronbach's alpha	0.745	0.802	0.556

Scoring Method B: 5-point single answer

	SCT-PS-MS2	SCT-PS-MS4	SCT-EM
Number of extracted factors	3	3	3
Total variance explained by extracted factors	19.82%	23.81%	18.77%
No. of items with factor loadings ≥ 0.4 (%)	17 of 58 (29)	24 of 58 (41)	13 of 49 (27)
Kuder-Richardson 20 Coefficient (KR-20)	0.809	0.778	0.464

Scoring Method C: 5-point distance from mode

	SCT-PS-MS2	SCT-PS-MS4	‡SCT-EM
Number of extracted factors	3	3	4
Total variance explained by extracted factors	26.59%	15.87%	16.19%
No. of items with factor loadings ≥ 0.4 (%)	23 of 58 (40)	14 of 58 (24)	14 of 49 (29)
Cronbach's alpha	0.876	0.745	0.478

Scoring Method D: 5-point aggregate with distance penalty

	SCT-PS-MS2	SCT-PS-MS4	SCT-EM
Number of extracted factors	4	4	4
Total variance explained by extracted factors	26.76%	18.37%	15.87%
No. of items with factor loadings ≥ 0.4 (%)	29 of 58 (50)	17 of 58 (29)	15 of 49 (31)
Cronbach's alpha	0.859	0.798	0.561

Scoring Method E: 3-point aggregate

	SCT-PS-MS2	SCT-PS-MS4	SCT-EM
Number of extracted factors	4	4	3
Total variance explained by extracted factors	16.35%	24.00%	11.91%
No. of items with factor loadings ≥ 0.4 (%)	13 of 58 (22)	21 of 58 (36)	6 of 48* (13)
Cronbach's alpha	0.590	0.667	0.322

Scoring Method F: 3-point single answer

	‡SCT-PS-MS2	SCT-PS-MS4	SCT-EM
Number of extracted factors	3	3	3
Total variance explained by extracted factors	33.80%	49.64%	25.66%
No. of items with factor loadings ≥ 0.4 (%)	19 of 58 (33)	33 of 58 (57)	5 of 49 (10)
Kuder-Richardson 20 Coefficient (KR-20)	0.518	0.549	0.278

*Item removed due to lack of variance
‡Promax rotation; factor correlation > 0.50.

Table 4.6: Sample of a representative component matrix: Rotated component matrix for optimized SCT-PS-MS4 (Scoring method A) and loading percentages by item difficulty.

	Component				Component				Component			
	I	II	III		I	II	III			I	II	III
Q01C01E	.274			Q40C09E	.316	.276		$n_{\geq 0.4}$=12	E	25%	58.3%	16.7%
Q02C01E	.321			Q41C09E		.441		$n_{\geq 0.4}$=4	M	50%	25%	25%
Q04C01M		.292	-.241	Q42C09M			.259	$n_{\geq 0.4}$=1	D	0%	100%	0%
Q06C01E		.316		Q43C10E								
Q10C02E		.401		Q44C10E	.535							
Q13C03D		.452		Q45C10D	.239							
Q14C03D				Q46C10E			.337					
Q15C03M	.255			Q47C10E	.240	.464						
Q16C04E	.254	.342		Q50C11E		.289						
Q17C04M	.247		.337	Q52C11E								
Q19C04E				Q53C12E								
Q21C05E	.233	.342		Q54C12E	-.210	.441						
Q22C05M		.242		Q55C12M	.334	.249						
Q23C05E	.201			Q56C12E	.342							
Q24C05E				Q57C12E	.352	.288						
Q25C05E	.418			Q58C13D	.273							
Q26C06E			.343	Q59C13M	.444	.245						
Q27C06M		.526		Q61C13E	.392							
Q28C06E			.203	Q62C13E	.323	.315						
Q29C06E	.320		.226	Q65C14D	.253	.231						
Q30C06E		.519		Q66C14E			.412					
Q31C07E	.248	.401		Q67C14D	.270		.271					
Q32C07E		.254	-.201	Q68C15E	.288		-.336					
Q33C07M	.233			Q70C15M	.296							
Q34C07D	.368			Q71C15D	.262	.277						
Q35C08E			.415	Q73C16M		.272						
Q36C08E	.430	.226		Q74C16M	.457							
Q37C08M			.422	Q75C16M		.278	.220					
Q38C08M	.368											
Q39C09E		.478										

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Loadings < 0.200 were suppressed. Loadings ≥ 0.4 are underlined.
Q##=question number; C##=case number; E=easy; M=moderate; D=difficult

Higher-order Factor Analysis. A second-order factor analysis, using a principle components solution and Promax rotation, was computed on the identified primary (first-order) factors extracted from the SCT-PS and SCT-EM. All second-order procedures factor analyzed three first-order factors on each instrument (e.g., Figure 4.4), under each scoring condition. Kaiser's criterion discarded second-order factors with Eigenvalues less than 1.00.

On the SCT-PS, one second-order factor was specified for the majority of scoring methods. Two second-order factors were identified for the SCT-PS under scoring method F. The second-order, one factor, solutions on the SCT-PS explained 49.39% or less of the total variance between first-order factors (Table 4.7). Two second-order factors were regularly extracted on the SCT-EM. Two second-order factors explained between 69.18% and 78.11% of the variance between first-order factors.

A Schmid-Leiman solution (SLS) was also computed to understand more fully the relationships between the exam items and the higher-order factors (e.g., Figure 4.5). SLS probes how well observed variables measure second-order factors.[83,85] Specifically, the variance explained between each item and each factor, regardless of factor level, was explored. In interpreting a SLS, items that have a greater second-order loading than first-order loadings are considered to be better measures of second-order factors.[85] Conversely, SLS can delineate which items are purer measures of first-order factors. Appendix D.1-D.6 presents SLS outputs for the SCT-PS. In general, the Schmid-Leiman solutions reported that SCT-PS items were often better measures of the second-order factor than of the first-order factors. For example, a SLS conducted under conditions of traditional 5-point aggregate scoring (Appendix D.1) found seven SCT-PS items to measure the second-order factor, while no items were found to measure the first-order factors (Figure 4.5). For scoring method E (3-point aggregate; Appendix D.5), 14 items were found to be better measures of the second-order factor, whereas one item was reported to be a more robust measure of factor one. After having extracted one second-order factor for the SCT-PS and SCT-EM and removing the unique variance of the

second-order factor from the first-order factors, little detectable variance remained in the first-order factors. These findings are largely inconclusive because numerous items failed to load on the second- or first-order factors. Under ideal SLS circumstances, all exam items would load on the first-order factors, second-order factor, or both. These results did not conform to a stereotypical higher-order model.

Table 4.7: Second-order factor analysis results

Scoring Method A: 5-point aggregate

	SCT-PS	SCT-EM	
	Second-order Component	Second-order Component	
	1	1	2
First-order factor-I	**0.780**	**0.794**	0.295
First-order factor-II	**0.741**	**0.814**	-0.207
First-order factor-III	**0.569**		**0.957**
Total variance explained by extracted components	49.39%	78.11%	

Loadings <0.2 are suppressed. Loadings ≥ 0.4 are in bold.

Scoring Method B: 5-point single answer

	SCT-PS	SCT-EM	
	Second-order factor	Second-order factors	
	1	1	2
First-order factor-I	**0.661**	**0.736**	
First-order factor-II	**0.754**		**0.974**
First-order factor-III	**0.655**	-0.727	
Total variance explained by extracted components	47.82%	69.18%	

Loadings <0.2 are suppressed. Loadings ≥ 0.4 are in bold.

Scoring Method C: 5-point distance from mode

	SCT-PS	SCT-EM	
	Second-order factor	Second-order factors	
	1	1	2
First-order factor-I	**0.528**	**0.823**	
First-order factor-II	**0.761**	**0.529**	-0.669
First-order factor-III	**0.709**	**0.456**	**0.759**
Total variance explained by extracted components	45.39%	74.36%	

Loadings <0.2 are suppressed. Loadings ≥ 0.4 are in bold.

Scoring Method D: 5-point aggregate with distance penalty

	SCT-PS	SCT-EM	
	Second-order factor	Second-order factors	
	1	1	2
First-order factor-I	**0.561**	**0.665**	0.396
First-order factor-II	**0.777**	**0.692**	0.229
First-order factor-III	**0.719**	-0.487	**0.867**
Total variance explained by extracted components	47.87%	70.63%	

Loadings <0.2 are suppressed. Loadings ≥ 0.4 are in bold.

Scoring Method E: 3-point aggregate

	SCT-PS	SCT-EM†	
	Second-order factor	Second-order factors	
	1	1	2
First-order factor-I	**0.632**	**0.754**	
First-order factor-II	**0.509**	**0.731**	-.260
First-order factor-III	**0.766**		**0.964**
Total variance explained by extracted components	41.51%	70.76%	

Loadings <0.2 are suppressed. Loadings ≥ 0.4 are in bold.
†1 item was removed due to lack of variance (i.e., all examinees received full credit on the item). A total of 48 items were analyzed in the first-order factor analysis.

Scoring Method F: 3-point single answer

	SCT-PS		SCT-EM†	
	Second-order factors		Second-order factors	
	1	2	1	2
First-order factor-I	**0.828**		**0.723**	**0.417**
First-order factor-II		**0.934**	**0.772**	-0.310
First-order factor-III	**0.709**	**0.407**		**0.890**
Total variance explained by extracted components	74.47%		72.67%	

Loadings <0.2 are suppressed. Loadings ≥ 0.4 are in bold.
†1 item was removed due to lack of variance (i.e., all examinees received full credit on the item). A total of 48 items were analyzed in the first-order factor analysis.

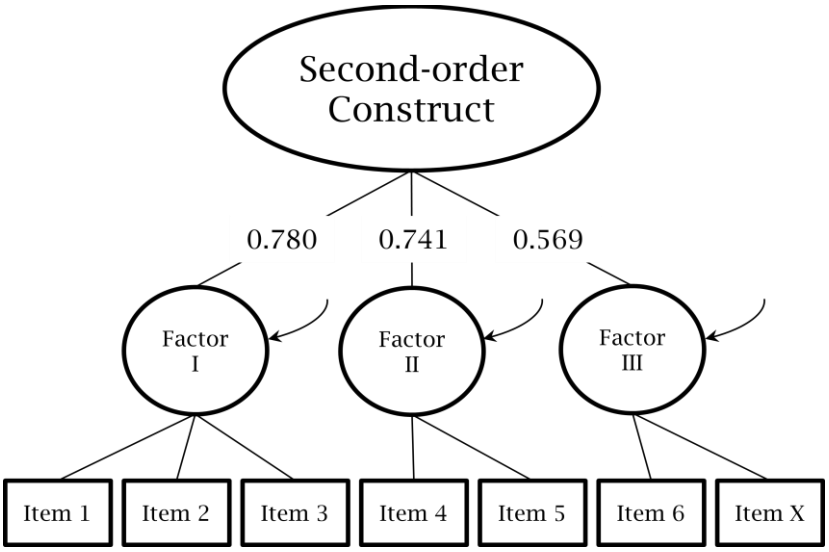


Figure 4.4: Sample second-order factor model - SCT-PS, scoring method 'A'. Floating arrows represent unique extraneous contributions to each factor. First-order factor loadings are not displayed.

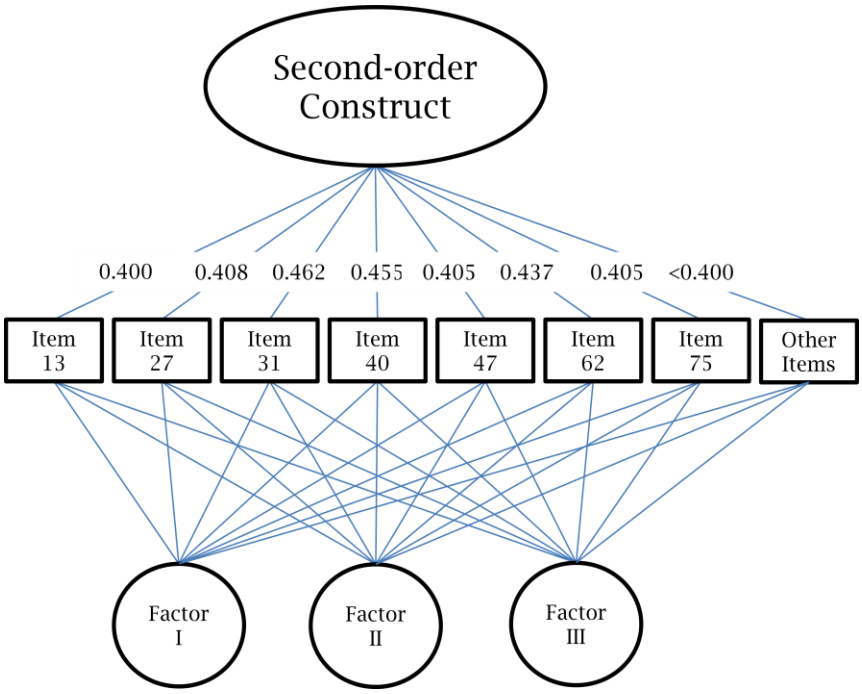


Figure 4.5: Schmid-Leiman solution for SCT-PS - scoring method 'A'. Contributions of seven items to the second-order factor were ≥0.400. Contributions of all items to the first-order factors were non-significant (≤0.381).

DISCUSSION: CONSTRUCT VALIDATION STUDY

This research investigated whether SCTs measure a single construct, an assumption driven by theory and indirect validity evidence. The absence of literature on SCT factor structure and the prospective use of SCTs as high-stakes assessments were the impetuses for conducting this study. In review, unfavorable confirmatory factor analysis findings prompted additional investigation via exploratory factor analytic techniques to understand the number and nature of SCT constructs and their influence on various scoring practices.

CFA Outcomes. Many who have studied the internal structure of the SCT have reported moderate to high Cronbach's alpha coefficients ranging from 0.60-0.90, indicating dependable reliability.[1,3,8,9,16,53] While evidence that the SCT demonstrates robust reliability across multiple medical disciplines supports the argument that SCTs probe a single common construct,[2] the CFA findings contested this claim. We found model-fit indices to marginally represent a unidimensional structure. In addition, examination of factor loadings revealed impracticalities as a limited number of items (34% or less) significantly loaded on one factor. Uniform discrepancies between model-fit indices and factor loadings among all three datasets suggest that SCTs are not unidimensional.

EFA Outcomes. In a study comparing consensus and aggregate SCT scoring approaches, 59% of experts did not agree with the consensus response decided by a convened expert panel [18]. As a result, Charlin et al. [18] posited that a single-best-answer approach should not be used in SCT testing. Bland,[20] however, asserts that Charlin's findings may have been an artifact of the 7-point response scale used. Bland also disputed the interpretation of Charlin's results because of ignored statistical violations. Bland subscribes to an alternate conclusion that aggregate scoring contains more random error than a single-best-answer approach.[20] Our analyses of the effects of scoring methods on factor structure indicate that no scoring method is superior to another, in terms of construct validity. Despite testing a variety of scoring methods and

having ample power, this study was unable to explicitly identify constructs measured by SCTs. While a range of 3-5 factors were routinely extracted, the total variance explained by extracted factors was less than desirable (i.e.,<0.60), and oftentimes less than half of an instrument's items loaded on the extracted factors. In the case of 3-point single answer scoring (scoring method F), the total variance explained and the number of significant factor loadings unpredictably increased at the cost of reliability. Also, our findings (Table 4.5, scoring method E) contradict Bland et al. who previously reported equivalent reliability coefficients between traditional 5-point aggregate scoring and 3-point aggregate scoring approaches.[20] We attribute this discrepancy to inadequate sampling, as the data utilized by Bland et al. contained only 85 examinees.

These findings represent the first attempt to directly investigate the factor structure of SCTs. From our evidence we propose two interpretations: Either 1) SCTs do not consistently measure the same latent constructs or 2) the factor structure of SCTs is more complex or multifarious than could be recognized by the analyses conducted.

Because extracted factors reported insignificant factor loadings for the majority of items and explained a nominal amount of the total variance, it is inconsequential to give meaning or labels to the presumed latent constructs. Given that the findings are largely inconclusive, no speculations about the nature of SCT constructs were made. Undeniably, exploration of SCT constructs deserves more rigorous study as extant literature on this topic is rare.

As Cronbach and Meehl[75] contend, research evidence that undermines an instrument's validity can be interpreted in three ways: (1) "The test does not measure the construct variable," (2) "The theoretical network which generated the hypothesis is incorrect," or (3) "The experimental design failed to test the hypothesis properly." In knowing that SCTs exhibit moderate to high reliability across multiple disciplines,[1,3,8,16,53] that SCT scores converge in moderation with local[54] and nationally standardized instruments,[16] that SCTs are resistant to intermediate effects,[1,4,6,14,23] and that SCTs have modest predictive power,[23,46] I argue that the theoretical networks of SCTs are rich. The

effects of sampling error in this study were reduced by utilizing datasets with large sample sizes. Also, a small ratio of factors to variables was extracted, and results were cross-validated with different SCT instruments and with multiple scoring methods. The remaining explanation is that the performed analyses were unable to detect distinct and stable SCT constructs; either because they are non-existent or they follow a more complex factor structure.

Assuming that constructs follow a more complex structure, to a small degree it is probable that items could load according to the response processes (e.g., pattern recognition, episodic memory, Bayesian reasoning, etc.) utilized by examinees. Alternatively, perhaps the way in which items load could be explained by Brunswik's lens model. In this model, a relationship exists between the weight (i.e., relative importance) of information and the final judgment made.[90] Therefore the interpretation of an outcome (i.e., an examinees response) is related to the weight that each examinee assigns to the cues that are present in the new information and the clinical vignette. Items, therefore, may load according the weights that examinees assign to cues within an item. Those items that examinees perceive as having strong meaningful cues may hang together; while those items with weak non-consequential cues may be more likely to load on the same factor. Both of these posited theories will require more substantial investigation.

Higher order factor analysis. The inconclusiveness of this research warranted the execution of additional higher-order factor analyses. If a higher-order factor structure had been identified than the goal of specific aim 2B (i.e., to evaluate the relationships between the identified construct(s) and pre-existing empirical and theorized clinical reasoning models) could have been addressed. However, in the absence of a convincing factor structure, adequate theoretical comparisons and practical connections could not be made.

The higher-order factor analysis suggested one second-order construct for the SCT-PS and two second-order constructs for the SCT-EM. With the higher-order factor(s)

contributing to more than 40% of the total (extracted) variance, their impact was notable.[91] However, interpretation of the Schmid-Leiman solution computed for the SCT-PS implied that very few exam items made a strong contribution to the second- or first-order factor(s). This research demonstrated that SCT items with significant loadings (though few) have a clear hierarchical structure, yet the underlying commonality that adjoins these items is not evident. Qualitative investigation of exam items failed to uncover overlapping qualities between well performing items. When comparing items with significant loadings against items with non-significant loadings, no explicit differences in item characteristics or quality were identified. A combination of weak Schmid-Leiman solution outcomes, few salient primary loadings, and a scarcity of shared item characteristics renders the higher-order factor analysis findings inconclusive.

A more thorough investigation of reliability was revealing and, in part, helped to explain the outcomes of this research. Reliability, as calculated using Cronbach's coefficient alpha, is a function of the number of items on an instrument and the degree of covariance among items. Instruments with items that have small inter-item correlations can demonstrate reasonable reliability in the presence of large numbers of items. For instance, the SCT-PS instrument had small inter-item correlations (mean inter-item correlation=0.067, min=-0.113, max=0.283), yet showed reasonably high reliability (0.802; scoring method A) due to the presence of 58 items. In this instance, decreasing the number of exam items from 58 to 10 would drastically reduced the reliability of the instrument, because of the lack of strong inter-item correlations. In contrast, instruments with items that exhibit consistently strong inter-item correlations would require fewer items to attain high reliability. To complicate matters, items with weak inter-item correlations tend not to factor analyze well (i.e., they rarely yield a clear consequential factor structure). The culmination of this information suggests that large numbers of SCT items are required to produce high reliability coefficients and also explains why a clear factor structure was not observed. The natural question to ask is,

"Why do SCT items have small inter-item correlations?" Could something inherent in the format or exam structure of SCTs explain this finding?

A commentary by Clarence Kreiter and the results of this research have caused me to critically rethink the exam structure of SCTs. In the commentary, Kreiter[92] argues that examinees must first assess the probability (**P1**) that the hypothesis (or investigative action, or therapeutic action, etc.) is reasonable in the context of the problem, and then they must calculate the likelihood and usefulness (**P2**) of the hypothesis given both the scenario and the new information. When examinees respond to an item they are therefore subjectively rating the magnitude of the difference (**P2-P1**) on a 5-point Likert scale. The interplay, or lack thereof, between P1 and P2 can cause response confusion. If a diagnostic test (**P1**) is undeniably useful based solely on the information in the case scenario, and the new information adds little to no insight (**P2**), an examinee is left to determine whether the final response reflects the usefulness of **P1** or the difference between **P2** and **P1**; thereby creating response confusion. Here is an example from the SCT-EM:

Case: A 35-year-old female patient present to the emergency department with the chief complaint of chest pain and shortness of breath for the last 2 days. The symptoms began suddenly and the pain is worse with deep breathing and located on the left side.

If you were considering asking for a…	*and you find the patient has a…*	*…this investigation becomes…*
chest X-ray	productive cough	-2 -1 0 +1 +2

-2: Not useful at all; -1: Less useful; 0: Neither more nor less useful; +1: Useful; +2: Absolutely necessary

In this example, a chest X-ray may appear to be a useful diagnostic test. However, the new information concerning a productive cough may not have any impact or may have only little impact on one's decision to order a chest X-ray. An examinee's response may therefore reflect the usefulness of the chest X-ray in the context of the scenario instead of reflecting the effect that the new information had on ordering an already useful chest X-ray; in which case it is no longer clear what is being measured. Is the investigative action (i.e., ordering a chest X-ray) in the context of the scenario alone being measured, or is the investigative action in the context of both the scenario and the

new information being measured? This type of discrepancy may begin to explain why inconsistencies and irregularities in SCT factor structure were observed in this research.

Restructuring the format of SCT items may prove useful in minimizing or eliminating response confusion. For example, requiring an extra yet separate response for the hypothesis (investigative action, therapeutic action, etc.) may bring clarity to the response process. Using the previous case and item, I propose a new SCT format as follows:

Case: A 35-year-old female patient present to the emergency department with the chief complaint of chest pain and shortness of breath for the last 2 days. The symptoms began suddenly and the pain is worse with deep breathing and located on the left side.

Asking for a…	*… will…*
chest X-ray	a b c

a: Not provide useful information; b: provide some useful information; c: provide very useful or necessary information

If you then found the patient has…	*…ordering a chest X-ray becomes…*
a productive cough	-1 0 +1

-1: Less useful than it already was; 0: Neither more nor less useful than it already was; +1: More useful than it already was

A SCT formatted in this way would require that credit be awarded based on paired scoring. For example, the modal response may be recorded as 'c,0'. All examinees who marked both 'c' and '0' would receive full credit. Partial credit would be awarded according to the non-modal paired responses of experts. Response choices were reduced to a 3-point scale because the total number of response combinations increased to nine. This revised formatting structure may be valuable to consider in future iterations of SCT research. Minimizing response confusion by restructuring how examinees respond to items may strengthen inter-item correlations and enhance the overall factor structure of SCTs. For the aforementioned item format to be effective will likely require that SCT items be administered electronically. Doing so will prevent the second part of an item from being viewed prior to answering the first part of the item.

The above example is one proposed recommendation. It may also be of worth to explore the use of Guttman scales and Thurstone scales to measure clinical reasoning,

as opposed to the use of Likert scales. Alternate scaling systems, such as the Guttman or Thurstone scales, would likely require the use of item response theory and more specifically a procedure called Mokken scale analysis. It is my hope that future investigations will, at minimum, consider such possibilities.

CONCURRENT VALIDITY STUDY

Scoring Method Analysis. Correlations between scores and training level for the purposes of concurrent validity were not conducted on the SCT-PS dataset that compared students as MS2s to students as MS4s because observations across samples were not independent. However, the SCT-PS demonstrated moderate predictive validity as correlations between MS2 and MS4 scores were significant ($p<0.001$) but modest ($r=0.381$). A repeated measures analysis also reported that all scoring methods discriminated between training levels. MS4s consistently scored higher than they did as MS2s ($p\leq0.001$, $\eta^2\geq0.093$), despite the scoring method employed. Table 4.8 summarizes these findings and presents reliability coefficients for each scoring method.

Table 4.8: SCT-PS descriptive statistics of all scoring methods

	Training Level (n=522)	Reliability	Mean Percentage Score (SD)	p-value (MS2s vs. MS4s)	Partial Eta Squared (η^2)
Scoring Method A (5-point aggregate)	MS2s MS4s	0.745 0.802	60.2 (10.0) 68.8 (10.5)	<0.001	0.361
Scoring Method B (5-point single answer)	MS2s MS4s	0.809 0.778	51.1 (14.3) 56.4 (12.3)	<0.001	0.102
Scoring Method C (5-point distance from mode)	MS2s MS4s	0.876 0.745	78.3 (9.8) 81.5 (5.2)	<0.001	0.093
Scoring Method D (5-point aggregate with distance penalty)	MS2s MS4s	0.859 0.798	70.4 (11.3) 75.9 (7.7)	<0.001	0.173
Scoring Method E (3-point aggregate)	MS2s MS4s	0.590 0.667	82.9 (5.6) 88.7 (4.8)	<0.001	0.408
Scoring Method F (3-point single answer)	MS2s MS4s	0.518 0.549	73.8 (6.4) 79.7 (5.8)	<0.001	0.371

Table 4.9: SCT-EM summary of descriptive statistics and correlation coefficients

SCT-EM					
	Reliability	Correlation with training level	Training Level (n)	Mean Percentage Score (SD)	Range (as percentage scores)
Scoring Method A (5-point aggregate)	0.556	0.784	EM Physician (12)	82.8 (3.1)	77.7-87.4
			PGY-3 (14)	72.0 (4.2)	66.3-82.5
			PGY-2 (15)	68.8 (5.5)	58.2-74.7
			PGY-1 (11)	63.1 (6.5)	53.5-75.1
			MS4 (988)	60.4 (8.0)	36.2-82.6
Scoring Method B (5-point single answer)	0.464	0.720	EM Physician (12)	68.5 (4.9)	61.2-75.5
			PGY-3 (14)	58.3 (5.6)	49.0-71.4
			PGY-2 (15)	54.8 (6.6)	40.8-65.3
			PGY-1 (11)	48.6 (7.6)	38.8-61.2
			MS4 (988)	47.3 (8.8)	20.4-73.5
Scoring Method C (5-point distance from mode)	0.478	0.721	EM Physician (12)	86.3 (2.7)	81.6-91.2
			PGY-3 (14)	80.6 (3.8)	72.8-88.4
			PGY-2 (15)	80.0 (3.5)	72.8-84.4
			PGY-1 (11)	74.0 (4.7)	69.4-83.7
			MS4 (988)	73.9 (5.1)	57.1-88.4
Scoring Method D (5-point aggregate with distance penalty)	0.561	0.765	EM Physician (12)	84.6 (2.7)	80.3-88.7
			PGY-3 (14)	76.3 (3.8)	70.0-85.0
			PGY-2 (15)	74.3 (4.4)	65.4-79.4
			PGY-1 (11)	68.5 (5.7)	60.5-79.5
			MS4 (988)	67.3 (6.4)	45.7-84.8
Scoring Method E (3-point aggregate)	0.332	0.678	EM Physician (12)	88.7 (4.1)	82.3-96.5
			PGY-3 (14)	84.5 (4.6)	76.8-93.0
			PGY-2 (15)	81.8 (5.1)	72.9-91.2
			PGY-1 (11)	77.7 (3.5)	72.7-84.4
			MS4 (988)	77.0 (5.3)	58.1-90.8
Scoring Method F (3-point single answer)	0.278	0.556	EM Physician (12)	78.6 (6.3)	67.4-89.8
			PGY-3 (14)	75.8 (6.9)	65.3-89.8
			PGY-2 (15)	73.3 (6.6)	61.2-85.7
			PGY-1 (11)	67.5 (4.7)	61.2-75.5
			MS4 (988)	67.8 (6.5)	46.9-85.7

Table 4.9 presents reliability coefficients, correlation coefficients, mean percentage scores, and the range of scores for all scoring methods computed on the SCT-EM dataset. To elicit a balanced design, composite scores derived for each scoring method were first weighted so that each of the five training levels equally represented 20% of the population. Composite scores were then correlated with training level to test the strength of their associations. All scoring methods demonstrated a significant, positive correlation with level of training ($r=0.556$-0.784, $p<0.001$; Table 4.9). Correlations among the six scoring methods were moderate to high ($r=0.675$-0.990, $p<0.001$). A one-way MANOVA that included all six scoring methods as dependent variables and training level (MS4s, EM residents, EM physicians) as the independent

variable reported significant differences between training levels ($p<0.001$, Wilk's $\lambda=0.864$ $\eta^2=0.071$). A power analysis using G*power indicated a 73.9% chance of detecting a medium effect size (as defined by Lomax[86]) at the 0.05 level. A follow-up post hoc test revealed significant pair-wise differences between training levels (i.e., MS4s vs. EM Residents, EM Residents vs. EM Physicians, and MS4s vs. EM Physicians) for each scoring method employed ($p\leq0.016$). A Box's M test, at $\alpha=0.001$ per the unbalanced design[87], was non-significant ($p=0.004$) indicating that homogeneity of variance-covariance matrices was satisfied.

Item Difficulty and Item Type Analysis. Univariate analyses were conducted on the SCT-PS and SCT-EM datasets to study the effects of item difficulty and medical training level on clinical data interpretation and to explore training level differences by item type. Only data generated with the traditional 5-point aggregate scoring method was used to conduct the item difficulty and item type analyses.

SCT-PS (MS2s vs. MS4s). Scores arranged by level of difficulty were normally distributed with the exception of scores on easy items captured from the second administration of the SCT-PS. Sphericity, tested because items were grouped into three difficulty levels, was violated warranting the use of the Greenhouse-Geisser correction. The repeated measures analysis reported a significant effect for time, with MS4s outperforming their scores as MS2s, net the effects of item difficulty. Controlling for time, differences in performance between easy, moderate, and difficult items were also found to be statistically significant ($p<0.001$, $\eta^2=0.509$, Table 4.10). A Scheffé procedure revealed statistically significant differences ($p<0.001$) for each pair-wise comparison (i.e., easy vs. moderate, moderate vs. difficult, and easy vs. difficult) on both administrations of the SCT-PS. Scores on easy items were significantly greater than those on moderate items which were significantly greater than those on difficult items (Table 4.11). A significant two-way interaction between time and difficulty was also observed ($p<0.001$, $\eta^2=0.159$, Table 4.10). That is, the change in mean performance scores (Δ mean), from

MS2s to MS4s, grew in magnitude as item difficulty increased. A post hoc power analysis [with model parameters of η^2=0.159, n=522, α err probability=0.05, 2 groups (i.e., MS2s and MS4s), and 3 measures (i.e., easy, moderate, difficult)] revealed a statistical power of 0.911. Figure 4.6 summarizes the above findings.

Table 4.10: SCT-PS repeated measures ANOVA summary table (with Greenhouse-Geisser correction)

	Type III Sum of Squares	**Degrees of Freedom**	**Mean Squares**	**F Ratio**	**p-value (sig)**	**Partial Eta Squared** (η^2)
Time (MS2 vs. MS4)	9.710	1.000	9.710	363.850	<0.001	0.411
(error)	13.903	521.000	0.027			
Item Difficulty	15.596	1.653	9.432	539.353	<0.001	0.509
(error)	15.065	861.420	0.017			
Time × Item Difficulty	2.360	1.669	1.413	98.673	<0.001	0.159
(error)	12.459	869.761	0.014			

Table 4.11: SCT-PS percentage scores by training level and item difficulty

	SCT-PS				
Items	1st Administration (as MS2s)		2nd Administration (as MS4s)		
	Mean	Std Dev.	Mean	Std Dev.	Δ Mean
Easy (n=35)	65.26	±10.07	71.64	±10.24	6.38
Moderate (n=15)	58.51	±13.83	66.41	±14.28	7.90
Difficult (n=8)	41.97	±20.04	60.80	±18.72	18.83

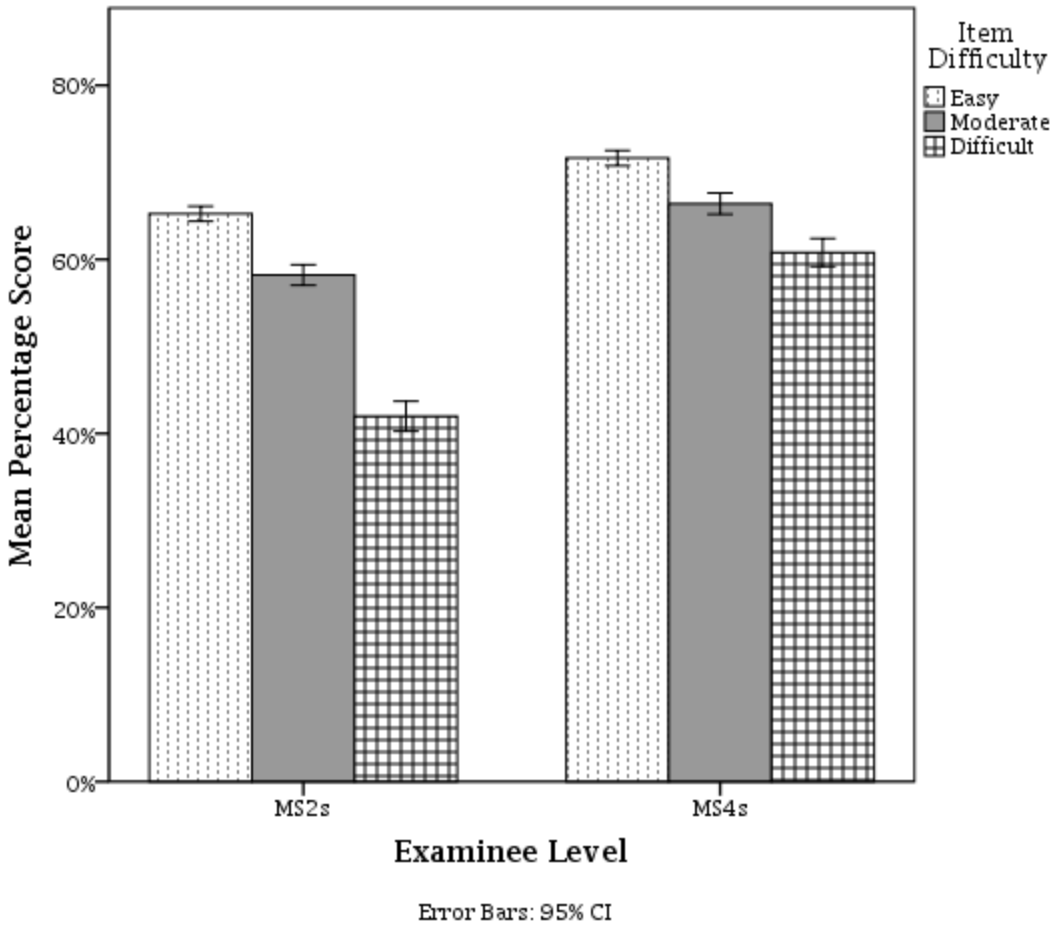


Figure 4.6: SCT-PS Performance by Item Difficulty - Bar graph comparing MS2 and MS4 mean percentage scores on the SCT-PS. Overall and within each difficulty category, MS2s performed significantly lower than MS4s. For MS2s and MS4s, scores on easy items were significantly higher than scores on moderate items which were significantly higher than scores on difficult items.

SCT-EM (MS4s vs. Residents vs. Experienced Physicians). Repeated measures between subjects analysis and an LSD multiple comparisons procedure[86] reported a statistically significant difference ($p<0.001$, $\eta^2=0.213$, Table 4.12) in overall SCT-EM scores between each training level. EM experts significantly outperformed EM residents who significantly outperformed MS4s, net the effects of item difficulty (Table 4.13). LSD is a commonly performed post hoc procedure for assessing pair-wise contrasts between three groups.[86]

Normality and sphericity assumptions were also violated on the SCT-EM dataset. As such a Greenhouse-Geisser correction was used to assess differences in item difficulty scores and the interaction between item difficulty and training level. Overall, scores on easy, moderate, and difficult items differed significantly ($p<0.001$, $\eta^2=0.064$, Table 4.12), irrespective of training level. A Scheffé multiple comparisons procedure[86] revealed that each pair-wise comparison of item difficulty scores was significant ($p<0.001$), net the effects of training level. Bonferroni (Dunn) procedures were also performed independently for MS4s, residents, and experienced physicians. MS4s performed significantly higher ($p<0.001$) on easy items compared to moderate items and significantly higher on moderate items compared to difficult items. Residents performed in a comparable manner ($p\leq0.036$). In the case of EM experts, no differences in performance between easy, moderate, or difficult items were identified ($p=0.801$; Tables 4.13).

A one-way ANOVA and an LSD post-hoc test were conducted to assess differences between training levels for each difficulty category (Figure 4.7). While homogeneity of variance was violated for easy ($p=0.007$) and difficult items ($p<0.001$), under large sample conditions ANOVA is robust with respect to departures. On easy items, experienced EM physicians generated significantly higher SCT-EM scores than EM residents ($p=0.001$) who in turn yielded significantly higher scores than MS4s ($p=0.043$).

Significant differences ($p<0.001$) between each medical training level were also reported for moderate and difficult items.

A significant interaction was observed between item difficulty and training level ($p<0.001$, $\eta^2=0.070$). That is, the combination of the main effects resulted in experienced physicians scoring higher than residents and MS4s at any level of difficulty (Table 4.13). The magnitude of the difference in mean performance scores increased as the gap between training level and item difficulty increased (Table 4.14).

Table 4.12: SCT-EM ANOVA summary table

	Type III Sum of Squares	**Degrees of Freedom**	**Mean Squares**	**F Ratio**	**p-value (sig)**	**Partial Eta Squared** (η^2)
Training Level	8.681	2	4.340	140.581	<0.001	0.213
(error)	32.017	1037	0.031			
Item Difficulty	3.342	1.393	2.399	71.116	<0.001	0.064
(error)	47.738	1444.560	0.034			
Level × Item Difficulty	3.689	2.786	1.324	39.245	<0.001	0.070

Table 4.13: SCT-EM percentage scores by training level and item difficulty

	SCT-EM					
	MS4s (n=988)		EM Residents (n=40)		EM Experts (n=12)	
	Mean	Std Dev.	Mean	Std Dev.	Mean	Std Dev.
Easy Items (n=31)	70.43	±8.40	73.15	±6.77	81.99	±4.64
Moderate Items (n=16)	46.77	±12.98	61.78	±12.20	84.61	±6.65
Difficult Items (n=2)	12.90	±22.64	46.16	±36.33	81.79	±22.71

Table 4.14: SCT-EM change in percentage scores (between training levels organized by item difficulty)

	SCT-EM		
	MS4s vs. EM Residents	EM Residents vs. EM Experts	MS4s vs. EM Experts
	|Δ mean|	|Δ mean|	|Δ mean|
Easy Items (n=31)	2.72	8.84	11.56
Moderate Items (n=16)	15.01	22.83	37.84
Difficult Items (n=2)	33.26	35.63	68.89

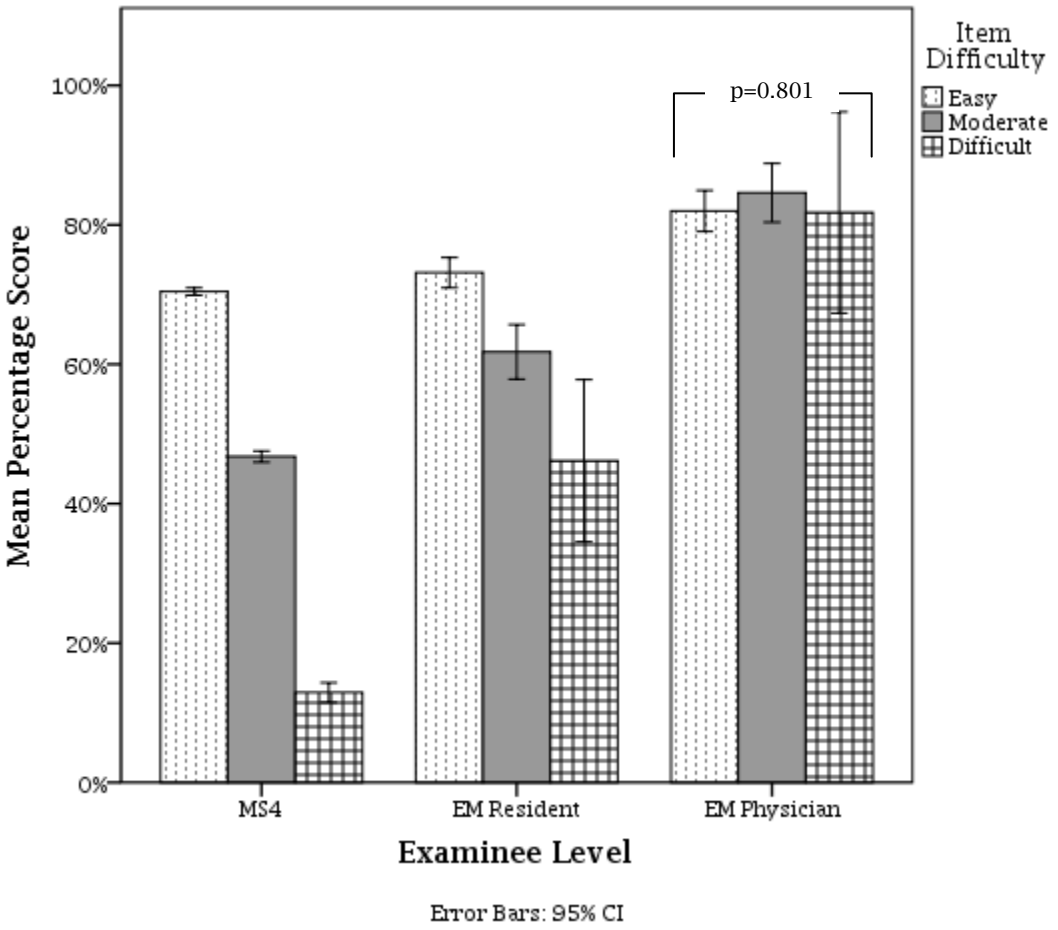


Figure 4.7: SCT-EM Performance by Item Difficulty - Bar graph comparing MS4, EM resident, and EM physician mean percentage scores on the SCT-EM. Overall and within each difficulty category, experienced EM physicians scored significantly higher than residents who scored significantly higher than MS4s. Among MS4s and EM residents scores on easy items were significantly higher than scores on moderate items which were significantly higher than scores on difficult items. No differences in scores categorized by difficulty were observed for experienced EM physicians.

Item Type Analysis. A repeated measures analysis on SCT-EM items categorized by type (i.e., diagnostic (n=21), investigational (n=11), or therapeutic (n=17)) reported significant differences in scores between item types within each training level ($p<0.001$, $\eta^2=0.014$) and between training levels ($p<0.001$, $\eta^2=0.111$). No interaction effect was observed ($p=0.066$). Irrespective of training level, performance on diagnostically oriented items was significantly higher ($p\leq0.002$) than investigational or therapeutic items. Overall, no performance differences ($p=0.094$) were observed between investigational and therapeutic items. A Scheffé multiple comparisons procedure reported that diagnostically oriented items discriminate between EM physicians, EM residents, and MS4s ($p\leq0.003$). On investigational items, MS4s scored as well as EM residents ($p=0.090$), whereas EM physicians scored higher ($p<0.001$) than MS4s and EM residents. Therapeutic items also exhibited discriminant properties as EM physicians scored significantly higher than EM residents who performed better than MS4s ($p<0.001$). A Cronbach's alpha calculation revealed that items categorized as diagnostic items ($\alpha=0.521$) had a higher reliability than those categorized as investigational ($\alpha=0.168$) or therapeutic ($\alpha=0.212$) items. For a visual summary of the item type analysis refer to Figure 4.8.

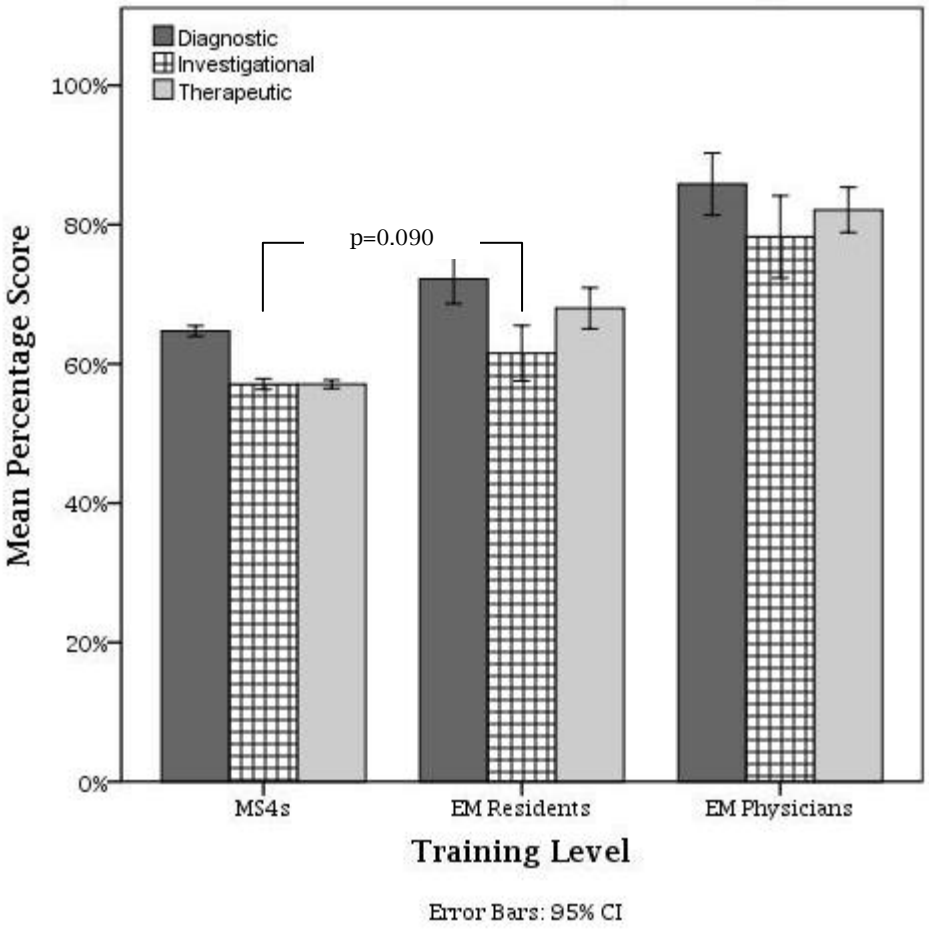


Figure 4.8: SCT-EM Performance by Item Type - Examinee scores on the SCT-EM grouped by item type. Diagnostic and therapeutic items were successful at discriminating between training levels ($p \leq 0.003$), whereas intermediate effects were observed on investigational items between MS4s and residents ($p=0.090$).

SCT-EM Item Survey. A survey was distributed to EM faculty to acquire their perceptions on the difficulty and ambiguity of randomly selected SCT-EM items. The survey response rate was 35.8% (29 of about 81). The item that was labeled difficult by the item analysis was perceived by 7.69% of faculty (1 of 13) to be 'very ambiguous or very ill-defined'. Some (38.46%, 5 of 13) faculty perceived the difficult item to be 'somewhat ambiguous or somewhat ill-defined', while most (53.85%, 7 of 13) perceived it to be 'straight forward or non-ambiguous' (Table 4.15A). Easy items were routinely classified by most faculty (53.85% or greater, 7 of 13) as 'straight forward or non-ambiguous'. Only one moderate item was classified by most faculty (53.85%, 7 of 13) as being 'somewhat ambiguous or somewhat ill-defined'. The remainder of moderate items were classified by most faculty (53.85% or greater, 7 of 13) as 'straight forward or non-ambiguous' (Table 4.15A).

A second group of EM faculty was responsible for labeling the same SCT-EM items as 'easy', 'moderate', or 'difficult'. Items that were found to be easy according to histogram analysis were labeled by most EM faculty (56.25%, 9 of 16) as 'easy', with the exception of one item that was labeled moderate by faculty (75.00%, 12 of 16). Moderate items were labeled 'easy' by faculty (50.00% or greater, 8 of 16) with the exception of one labeled 'moderate'.

When participants were forced to identify items within a case that were 'most ambiguous or ill-defined', more challenging items (as defined by item analysis) were not always perceived as being more ambiguous or ill-defined (Table 4.15B). For example question 83 on the SCT-EM was categorized as 'easy' according to the item analysis. However, more EM faculty perceived this item to be the 'most ambiguous or ill-defined' item for case 12 over an item that was categorized as 'moderate' by the item analysis.

Table 4.15A: Results of SCT-EM item survey

Item	Difficulty according to item analysis	EM Faculty Perceptions n=16			EM Faculty Perceptions n=13		
		Easy	Moderate	Difficult	straight forward or non-ambiguous	somewhat ambiguous or somewhat ill-defined	very ambiguous or very ill-defined
Q80C11	Difficult	37.50%	50.00%	12.50%	53.85%	38.46%	7.69%
Q81C11	Easy	56.25%	37.50%	6.25%	84.62%	15.38%	0.00%
Q82C11	Moderate	56.25%	37.50%	6.25%	92.31%	7.69%	0.00%
Q83C12	Easy	25.00%	75.00%	0.00%	53.85%	38.46%	7.69%
Q85C12	Easy	56.25%	37.50%	6.25%	69.23%	30.77%	0.00%
Q89C12	Moderate	43.75%	50.00%	6.25%	53.85%	46.15%	0.00%
Q53C6	Moderate	62.50%	37.50%	0.00%	92.31%	7.69%	0.00%
Q54C6	Easy	68.75%	31.25%	0.00%	69.23%	30.77%	0.00%
Q56C6	Moderate	56.25%	31.25%	12.50%	30.77%	53.85%	15.38%
Q40C3	Moderate	50.00%	43.75%	6.25%	61.54%	30.77%	7.69%
Q41C3	Easy	81.25%	18.75%	0.00%	69.23%	30.77%	0.00%
Q42C3	Easy	56.25%	43.75%	0.00%	61.54%	30.77%	7.69%

Q##=question number, C##=case number

Table 4.15B: Results of SCT-EM item survey continued

Item	Difficulty according to item analysis	EM Faculty Perceptions n=13
		Most ambiguous or ill-defined per case (check all that apply)
Q80C11	Difficult	37.50%
Q81C11	Easy	25.00%
Q82C11	Moderate	0.00%
None are ambiguous or ill-defined		18.75%
Q83C12	Easy	37.50%
Q85C12	Easy	18.75%
Q89C12	Moderate	31.25%
None are ambiguous or ill-defined		18.75%
Q53C6	Moderate	6.25%
Q54C6	Easy	31.25%
Q56C6	Moderate	56.25%
None are ambiguous or ill-defined		12.50%
Q40C3	Moderate	25.00%
Q41C3	Easy	6.25%
Q42C3	Easy	25.00%
None are ambiguous or ill-defined		18.75%

Q##=question number, C##=case number

DISCUSSION: CONCURRENT VALIDITY STUDY

This project utilized scores from independent SCT instruments to cross-validate and more intimately explore the concurrent validity of SCTs under six scoring conditions. Furthermore, this study also assessed concurrent validity at the level of item difficulty and item type in an attempt to understand the factors (or constructs) that drive the discriminatory power of SCTs. The results of this study supported hypothesis three that stated, "Non-traditional SCT scoring methods will closely reflect the properties of conventional methods, and at the level of item difficulty and item type, SCTs will retain their ability to differentiate between training levels." SCT scores discriminated between levels of experience under each of the six scoring conditions. In addition, disparities between developmental stages were observed at the item difficulty and item type levels.

Scoring Method Analysis. A study by Seibert et al.,[14] whose focus was in urology, reported the SCT satisfied parameters of concurrent validity. Novices, residents, and experts demonstrated significantly different levels of reasoning, and SCT scores were positively correlated to training level. Because numerous studies[3,7-9,16] testing concurrent validity at the composite score level have reported similar findings, it was not surprising to observe that all six scoring methods differentiated between levels and that scores strongly correlated with level of development.

Unexpectedly, the reliability of the 3-point scoring methods was consistently lower than that of all 5-point scoring methods. This finding contradicted reports by Bland et al., whose study contained a comparatively smaller sample of 85 examinees. Of the more reliable 5-point scoring methods, methods A and D regularly reported moderate to large measures of effect size ($\eta^2 \geq 0.104$) and demonstrated the highest correlation coefficients. This suggested that the efficacy of methods A and D to discriminate between training levels was marginally superior to other methods.

One disadvantage of using either 5- or 7-point Likert-scales in traditional aggregate scoring is that test administrators cannot readily distinguish examinee

responses that were near the modal response versus those that were distant from it.[20] For instance, if the mode response of the reference panel was '+2', examinees who answer '-1' receive the same score of 0.00 as those who answered '+1' (presuming no other panel members answered '-1' or '+1'). It is therefore possible for examinees who agree with panel members on the response direction but not the impact to receive the same score of 0.00 as someone who fails to identify both the direction and the impact.[20] This contingency was the impetus for testing the efficacy of scoring method D (5-point aggregate with distance penalty). The properties of scoring method D were similar to traditional aggregate scoring (method A) with the benefit of simultaneously measuring both response direction and impact.

Employing a 3-point scaling system would all together eliminate 'degree of correctness' concerns. However, our findings demonstrated that 3-point scoring methods were less reliable, and with 3-point scoring procedures the value of differing expert opinions is minimized. Qualitative data of student perceptions implies that 5- or 7-point scaling systems should be avoided, as students reported at times arbitrarily choosing between '+1' and '+2' and '-1' and '-2'.[20] In addition to concerns regarding 'degree of correctness', Bland contends that "if a single best answer to an SCT does not exist, the SCT will be of limited use for in-course assessment".[20] The rationale is that novices are expected to perform like experts, to attain the best possible score. Customarily, course assessment instruments are designed to assess specific course objectives or behaviors. Without a single best answer, it becomes difficult to define attainable objectives. The complexities of aggregate scoring are enough for some practitioners to refrain from the use of this method entirely. While I do acknowledge the above legitimate concerns, based on the findings of our research, I recommend using either a 5-point aggregate (method A) or 5-point aggregate with distance penalty (method D) approach when scoring SCTs because they exhibited stronger internal consistency and validity coefficients than the other tested methods; keeping in mind that scoring method D accounts for 'degree of correctness' unlike scoring method A.

Item Difficulty and Item Type Analysis. Our findings demonstrated that MS4s, who have greater clinical knowledge and exposure to patients through clerkship experiences, performed significantly higher at all difficulty levels than they did as MS2s. Likewise, on the SCT-EM, experienced physicians outperformed residents who outperformed MS4s in all difficulty categories. Because residents and practicing physicians have increased exposure to rare and atypical presentations, they are theoretically able to build, refine, and link illness scripts in a more organized, purposeful manner than undergraduate medical students.[58]

The present retrospective study was performed at a large multicenter institution. Each of the nine IUSM centers autonomously delivers instruction to medical students during years one and two of undergraduate training. As such, I believe that aspects of this study are representative of a large-scale multi-institutional study. Our results do echo a cross-sectional multi-institutional study that investigated differences in clinical reasoning skills of undergraduate medical students. In the aforementioned study, Williams et al.[62] reported clinical data interpretation gains at each level of undergraduate medical training, though gains in the third year were not as substantial. It was also reported that medical school elements (e.g., curriculum, instructional delivery systems, faculty, etc.) account for only a small percentage of variation in data interpretation scores. The study by Williams et al., however, did not explore the nuances of item difficulty or item type on clinical reasoning performance.

Items categorized by type were also found to distinguish between training levels, with the exception of investigational items not being able to differentiate between MS4s and EM residents. The outcomes of the item type analysis suggested that: (1) residents within the EM program at IUSM do not perform as well on investigational items as might be expected, (2) MS4s are not as well trained on investigational and therapeutic items as they are on diagnostic items, (3) categorizing items into sub-constructs may prove useful for evaluating specific cognitive skill sets and holds promise as an additional marker for program evaluation (provided that the reliabilities of items categorized by

type can be enhanced), and (4) clustering items into meaningful types modestly substantiates the construct validity and three dimensional factor structure of the SCT-EM. If a well constructed tripartite SCT instrument conformed to a more consequential factor structure than was observed in this research, it would likely be pertinent to reframe the customary reporting of composite scores into three subscores that represent the corresponding item types. While the concurrent validity properties and partial factor loading evidence point to three distinct constructs on the SCT-EM, additional work in this area is necessary to confirm these findings and determine whether such outcomes can be extrapolated to other SCTs.

SCT-EM Item Survey. The SCT-EM item survey was conducted to understand whether difficult items were also perceived as ambiguous and to cross check that the statistically derived levels of difficulty were reflective of faculty perceptions. On the survey, EM faculty mostly perceived SCT-EM items to be easier than was specified by the item analysis, which could not distinguish between difficulty levels (p=0.801) using EM faculty scores. This finding was not surprising considering that consensus on item difficulty is not easily reached, nor is consensus commonly expected when examining the content validity of items.[93] In addition, no items were considered to be 'very ambiguous or ill-defined' by most faculty. As in the instance of item difficulty, consensus concerning ambiguity is likely not easily reached. If ambiguity is clinician specific then how one clinician defines ambiguity may not align with how other clinicians define ambiguity. It is also possible cases were not ambiguous enough to elicit a response from participants.

SUMMARY

While research questions 2 and 3 resulted in two independent studies, their topics and content are intimately connected. It could be reasonably argued that because of the inconclusive nature of the construct validity study, the findings of research question 3 are arbitrary, relative, or invalid. If the constructs of SCTs are unknown, then

making assertive statements about the effects of item difficulty, item type, and experience on the assumed construct of data interpretation may resonate with some as being farfetched. To this end, readers are reminded that construct validity is not an all or nothing proposition, but is a matter of degree. While the construct validity findings of research question 2 challenge much of the SCT literature to date, there remains the possibility that SCT constructs exist in a more complex form than was originally anticipated. There is also a chance that the current SCT structure imposes a degree of response confusion that ultimately affects the construct validity of the instrument. Interestingly, sub-constructs of the SCT-EM (i.e., diagnostic items, investigational items, therapeutic items) were found to reasonably discriminate between developmental levels. This finding was in alignment with the construct validation study and suggests the SCT-EM likely has three first-order constructs with meaningful properties. Until additional evidence is gathered, the assertions and inferences concerning the ability of SCTs to measure clinical data interpretation will hold moderate value because of the abundance of prior SCT research that supports this argument.

In summation, this chapter reported and discussed the findings of this research. The outcomes of research question 1 provided evidence that SCTs are resistant to gender biases. The results of research question 2 that explored the construct validity of SCTs reported that SCTs do not adhere to a unidimensional factor structure as literature has previously suggested. Moreover, no single scoring method outshines another in terms of conveying clear identifiable constructs. Lastly, findings from research question 3 provided evidence that data interpretation ability and medical training level could be measured concurrently by all six scoring methods and by items categorized by difficulty and those clustered by type.

CHAPTER FIVE

Conclusions

Introduction

Significant Findings and Conclusions

Implications

Research Limitations and Strengths

Recommendations for Future Research

Research Synopsis and Closing

INTRODUCTION

Psychometric evaluations contribute abundantly to the improvement of medical education and quality of formative and summative assessments. The goals of this academic research project were to analyze the construct validity of SCTs and to explore the nuances of clinical data interpretation as it pertained to SCT item difficulty, item type, and non-traditional scoring methods. Basic psychometric properties of SCTs were also investigated to assess the presence of construct biases and to optimize the available datasets prior to analysis.

The remainder of this chapter is divided into five sections. The first section will review the significant findings of the three research questions and will summarize major research conclusions. The second section will address the implications of this work. Under the heading 'Research Limitations and Strengths' the shortcomings and virtues of each study will be discussed, and the fourth section will cover future plans for additional educational research. Finally, the last section of this chapter will comprehensively summarize the work of this composition in its entirety and will serve as a conclusion to bring the efforts of this dissertation to a close.

SIGNIFICANT FINDINGS AND CONCLUSIONS

Item Analysis. Common classical test theory procedures were used to examine the presence or absence of construct biases and to optimize the instruments under investigation. As a result of inadequate item-total correlations and poor item discrimination indices, 17 items were removed from the SCT-PS and 10 items discarded from the SCT-EM. Consequently, reliability of each instrument increased. Data from the optimized instruments were then used to conduct the main analyses of this study. A Pearson's product–moment correlation, to assess gender biases, found both SCT instruments to be fair. Males and females performed equally well at all levels of item difficulty. Collectively, optimization and absence of construct biases led the researcher

to infer that confounding effects were minimized negating concern for statistical inflation from instrument error.

Construct Validation Study. Previous psychometric research on SCTs has repeatedly demonstrated that SCTs have high reliability consistent across medical disciplines. This and other implicit evidence has lead researchers to infer that SCTs measure a common single construct, perceived to be data interpretation. A combination of confirmatory and exploratory factor analyses was used to investigate the veracity of the above assertions. A confirmatory factor analysis reported moderate model-fit indices and unconvincing factor loadings, as a majority of items presented with factor loadings below the significant salient threshold. The outcomes of this research, therefore, suggested that SCTs do not conform to a unidimensional factor structure. This opposes what has previously been theorized. Furthermore, examination of SCT scores via exploratory factor analysis provided evidence that latent constructs do not follow a simple first-order factor model, as no constructs were substantively extracted. Majority of items did not load on the extracted factors and the total variance explained by the extracted factors was less than 50%. Similar trends were observed across all six scoring methods investigated. Results of a higher-order factor analysis in conjunction with a Schmid-Leiman solution echoed the abovementioned findings in that no substantive factor structure with ample item loadings could be identified. This research underscores the need for more rigorous psychometric evaluation of SCTs and accentuates concerns relative to the meaning of SCT scores.

Concurrent Validity Study. A closer look at SCT scores was warranted to more fully understand the discriminant nature of this uniquely constructed instrument. Assessment of six mathematically contrived SCT scoring methods was informative in that all methods were found to discriminate between medical training levels, but methods on a 3-point scale were less reliable than 5-point scoring procedures. Additional outcomes disclosed that experienced clinicians outperformed residents who outperformed medical students on easy, moderate, and difficult clinical data

interpretation problems. Likewise MS4s outperformed their own MS2 scores on a problem solving SCT at every level of item difficulty. From this, it was concluded that data misinterpretation at any point along a continuum of difficulty decreases as a function of clinical experience and extended practice. Differentiation between training levels was also observed in items arranged by type. The above outcomes raise the question of whether or not data interpretation skills can be proactively taught and improved upon through intentional instruction. With optimism, I believe deliberate and strategic interventions can be implemented to promote and advance clinical reasoning in learners at multiple levels; though best practices for achieving such outcomes are still being explored.

IMPLICATIONS

Construct Validation Study. The outcomes of research question 2, that focused on the direct analysis of SCT constructs, offered valuable psychometric information that has a propensity to influence scoring practices and the meaning held by SCT instruments. Although results were deemed inconclusive in answering the questions posed, they were conclusive in other respects. The findings of this research bear worth because it is now known that SCTs do not conform to a unidimential model, nor do SCTs fully conform to a first-order or higher-order factor model. It was also concluded that different scoring procedures had no effect on construct validation. The implication is that medical schools and programs should consider this evidence and proceed with caution when attempting to draw meaning or make decisions from SCT scores. Until a sound empirically driven factor structure is identified, it is unclear what SCTs measure. The repercussions of an instrument having poor construct validity are problematic, because appreciating the makeup of an instrument's constructs tends to foreshadow and influence the meaning and subsequent structure of performance scores. While the writing and development of SCTs is thought to be a straight forward intuitive process,[14] future extrapolations of this research may challenge this dogma. Understanding more

fully the commonalities between items with good psychometric properties versus those with poor properties may lend itself to more steadfast rules that bolster the development of psychometrically sound SCTs.

Messick's multi-faceted theory of validity focuses on the importance of score meaning, the relevance and utility of test scores, and the consequences of test interpretation and use.[94] Without strong evidence of construct validity, the meaning of scores may become misconstrued, affecting the relevance and utility of scores, thereby distorting the interpretation and use of such data. Because "validity is not an all ('valid') or nothing ('invalid') proposition; rather it is a matter of degree," educators are charged with appraising the value and weight of all validity evidence in light of underlying theories. Until clear evidence of meaningful constructs is reported, it is advised that the interpretation and use of SCT scores be met with caution. Institutions or programs that are currently utilizing or contemplating the use of SCTs are encouraged to carefully and fully consider the presented evidence when interpreting the value of such an assessment.

Concurrent Validity Study. The findings of research question 3 consistently demonstrated gains in data interpretation performance from one training level to the next. It is therefore inferred that data interpretation skills are being learned whether direct efforts to promote the acquisition of such skills are employed or not. Regardless of how this phenomenon is occurring, these outcomes offer a glimpse of hope for educational practitioners that perhaps clinical environments and learning opportunities can be created to cultivate the development of data interpretation and clinical reasoning skills. Rather than leaving data interpretation abilities, and ultimately script development, to the random variability of clinical exposure, I advocate for the construction of, and student exposure to, authentic and standardized interventions aimed at promoting diagnostic reasoning growth. A study by Nabil et al.[95] has shown moderate success in improving reasoning skills after exposing learners to problem-solving schemes that mimic the cognitive tasks of physicians. However, best practices

for developing data interpretation and diagnostic reasoning skills, as well as remediating reasoning deficiencies in novice learners and residents, is a growing area of study that deserves more rigorous investigation.

Another implication of this research is that SCTs could prospectively be used as markers to measure the success of programs that are teaching skills related to or dependent upon data interpretation abilities. For example, a clerkship director overseeing a third or fourth-year clerkship may benefit from knowing that the skills of medical students on diagnostic items are far superior to their skills on investigational and therapeutic items. Access to this knowledge may result in a shift in emphasis toward investigational and therapeutic items, so as to enhance the well-roundedness of students. Because data interpretation is more or less a universal trait common across all clinical disciplines, it stands to reason that SCTs could be used to gauge the productivity and effectiveness of the teaching of diagnostic reasoning skills between medical specialties.

RESEARCH LIMITATIONS AND STRENGTHS

Item Analysis. Testing male against female SCT-EM performance by item difficulty was unrevealing for difficult items because only two difficult items were reported for the SCT-EM. Consequentially, Pearson's correlation coefficient was observed to be 1.00, because two data points naturally produce a perfectly correlated line. Repeating this test with a greater number of difficult items may yield different results.

Demographic information beyond gender was not tested. For example, construct biases towards a specific race or ethnic group were not evaluated. Also, this research was not able to control for the effects of past experience or variability in curriculum on data interpretation ability. As MS2s, students take the SCT-PS exam after having been exposed to different curricula and pedagogies, a consequence of the nine campus system unique to Indiana University. In addition, MS4s rotate through the EM clerkship

at different points along the continuum of their training. Therefore, confounding variables (e.g., the month in which the SCT-EM was taken) may have mildly influenced examinee performance and overall test outcomes.

Construct Validation Study. Perhaps the greatest limitation of the construct validation study was the modest reliability of the SCT-EM instrument. It has been documented that poor reliability can result in low communalities because variance from random error cannot be explained by common factors.[96] This principle elucidates why the total explained variance of the SCT-EM was consistently lower than that of the SCT-PS when the same number of factors was extracted. Furthermore, when individual items are the unit of analysis, principle component-based estimation can over extract factors as data reliability decreases and under conditions in which items are categorized (e.g., Likert-scale responses).[89] Also, this research was conducted at a single institution of medicine and may not be generalizable to other healthcare professions or locals. Because this study was bounded by the sample of examinees tested, other SCT instruments with different items may yield different results. The design of this study was strengthened by utilizing dataset with large sample sizes, by cross-validating results with independent SCT instruments, and by testing the effects of multiple scoring methods on construct validation.

Concurrent Validity Study. While this large-scale study included data from two SCTs for the majority of analyses, it was not without limitations. Correlation coefficients, providing evidence of concurrent validity, could not be cross-validated with the SCT-PS dataset because it consisted of only two training levels and observations were not independent across samples. The number of difficult items identified on each exam was restrictive. Also, performance of experienced physicians on the SCT-EM was not ideal as they performed equally well on easy, moderate, and difficult items. This may suggest that either greater disparity between difficulty categories could have been attained or a natural clinical reasoning plateau was reached by experienced physicians. This finding may have been a result of using only student SCT scores to identify natural breaks

between levels of item difficulty. Finally, the presented outcomes may not translate to all SCT instruments.

It is thought that this research as a whole was largely resistant to the effects of case specificity due to the presence of multiple cases (e.g., 16 cases for the SCT-PS and 12 cases for the SCT-EM). Case specificity occurs when problem solving ability is dependent on the attributes of a specific case.[97] According to Norman et al.,[98] the overall effects of case variance are smaller than the effects of item variance. The case-to-test and item-to-case ratios that showed the highest reliability estimates in Norman's study were 15-20:1 and 2-3:1, respectively.[98] The SCT-PS instrument aligned with these recommendations. The number of SCT-EM cases (i.e., 12) fell just short of the 15-20 case-to-test ratio. This may explain why the reliability of the SCT-EM instrument was lower than that of the SCT-PS.

RECOMMENDATIONS FOR FUTURE RESEARCH

Item Analysis. Further explorations of SCTs and future attempts to optimize this instrument should take into consideration additional construct biases and the potential effects of confounding variables. Secondly, because this work was viewed through the lens of classic test theory, additional analyses from the perspective of item response theory could also be considered.

Construct Validation Study. Despite the contributions of this research, knowledge gaps still remain in the domain of SCT scoring and construct validation. This research accentuates the need to explore SCT constructs in more depth and with deliberate intentions to enhance scoring procedures. Replication of this work is strongly encouraged. This work could also be supplemented by evaluating additional SCT scoring methods, such as those that utilize a 7-point scale or methods that rely on genuine consensus scoring rather than mathematically derived consensus scoring. It is also recommended that other higher-order factor models, such as a group-factor model and a bi-factor model that Rindskopf and Rose describe, be explored.[99] Last of all, it may

prove beneficial to investigate clusters of conceptually related SCT items referred to as 'testlets'; a concept predominantly used in computer adaptive testing.[100] Bundling items into content related areas may strengthen the instrument by localizing item ordering effects influenced by item difficulty and by minimizing context effects such as the balance and representativeness of content (thereby preventing repeated emphasis of a particular subject or theme). The use of testlets may prove more effective in identifying and defining latent constructs than we were able to accomplish via item-level analyses.

The aspects that contributed to an incomprehensible factor structure were unclear. It may be that response confusion was partly to blame for the inconclusive findings. As such, future iterations of SCT research may find value in reformatting the structure of SCTs. My recommendations for a revised SCT structure and scoring approach were recorded in the discussion section of the concurrent validity study in chapter four. Comparing the outcomes of this research to the future outcomes attained from the newly formatted SCT may prove revealing.

Concurrent Validity Study. To compliment the concurrent validity findings and to further enrich the body of validity evidence, future studies aimed at directly evaluating reasoning and cognitive processes used to respond to SCT items are needed. An adaptation of a protocol used by Boshuizen and Schmidt[39] that employs naturalistic observations, think-aloud methodologies, and focused probing to explore the order or simultaneous extraction of the cognitive layers involved in interpreting ambiguous and ill-defined data may prove useful in generating meaningful information to reach more substantial conclusions regarding the differences in approaches used to answer SCT items.

Learning, in part, from this research and the research of others that data interpretation skills mature with experience and clinical exposure brings optimism that perhaps the growth and learning of such skills can be fostered in more controlled environments. However, best practices and examples of such learning environments are few. Future research aimed at eliciting the direct development of data interpretation

skills and diagnostic reasoning would, therefore, be welcomed by members of the medical education community.

RESEARCH SYNOPSIS AND CLOSING

In spite of a history rich in educational measurement and psychometric research, the work of academic investigators on SCT related topics is far from complete. In review, this work explored central yet foundational questions related to the utility, meaning, and interpretation of script concordance test scores. The first of the three research questions was integrated into two major studies related to construct and concurrent validity.

Confirmatory and exploratory factor analyses were used in a SCT construct validity investigation to test the theorized assertion that SCTs are unidimensional. Outcomes of this research contradicted this assertion. The investigated SCTs did not conform to a unidimensional model for assessing clinical reasoning competence. Methodically performed exploratory analyses that considered the potential impact of various scoring methods were also unsuccessful at identifying the number and nature of latent constructs measured by SCTs.

The second study, that mostly explored categorization effects, was focused on testing general properties of the exams. Of interest was assessing whether the discriminatory power of SCTs would be heighted or lessened when considering items categorized by difficulty or type. As anticipated, SCT scores varied significantly between training levels irrespective of how items were categorized, with one exception. Analysis of non-traditional and conventional composite scoring procedures revealed similar outcomes in that no matter the scoring method employed, each of the six methods retained the ability to differentiate between medical training levels.

In its entirety, this dissertation has brought added clarity to the discriminatory nature of SCTs and has provided direct empirical evidence that contests the construct validity of SCT instruments. As such, this psychometric evaluation has made meaningful

contributions to the fields of clinical diagnostic reasoning and SCT research. It is my hope that medical education scholars will benefit from these findings and build upon this work to perpetuate the advancement of clinical reasoning and other related educational enterprises inherent to the practice of undergraduate and graduate medical education.

APPENDICES

APPENDIX A

Name: _____ Score: _____/ _____

MS2/MS4 Problem Solving Script Concordance Test

Cases were removed to protect the integrity of this exam, as this test is currently utilized by Indiana University School of Medicine.

Case #1

1. If you were thinking of the diagnosis of urinary tract infection and you find out that the patient has a history of dysuria then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

2. If you were thinking of the diagnosis of appendicitis and you find the following evidence, right lower quadrant pain then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

3. If you were thinking of the diagnosis of biliary colic and you find the following evidence, no gallstones on ultrasound scan then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

4. If you were thinking of the diagnosis of renal colic and you find the following evidence, hematuria then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

5. If you were thinking of the diagnosis of appendicitis and you find the following evidence, the patient is hungry then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

6. If you were thinking of the diagnosis of gastroenteritis and you find the following evidence, the patient has diarrhea then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

Case #2

7. If you were thinking of the diagnosis of strep throat and you find out the following evidence, negative bacterial swab for strep, then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

8. If you were thinking of the diagnosis of esophageal cancer and you find the following evidence, the patient has no dysphagia but did recently cough up blood, then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

9. If you were thinking of the diagnosis of lung cancer and you find the following evidence, the patient quit smoking five years ago, then this diagnosis becomes?
   A. -2=almost ruled out
   B. -1=somewhat less probable
   C. 0= neither less or more probable
   D. +1=somewhat more probable
   E. +2=almost certain

10. If you were thinking of the diagnosis of recurrent laryngeal nerve compression and you find the following evidence, patient has a family history of aortic aneurysm, then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #3

11. If you were thinking of the diagnosis of simple back strain (lumbago) and you find out that the patient has a decrease in sensation over the plantar and lateral aspect of the left foot and left calf, absent ankle reflexes then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

12. If you were thinking of the diagnosis of spinal cord compression and you find the following evidience, the patient has decreased rectal tone and urinary incontinence then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

13. If you were thinking of the diagnosis of vertebral osetomyelitis and you find the following evidence, a recent history of injecting drug use then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

14. If you were thinking of the diagnosis of nephrolithiasis and you find the following evidence, urinalysis positive for microscopic hematuria then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

15. If you were thinking of the diagnosis of muscle strain and you find the following evidence, localized tenderness over L2 spinous process then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #4

16. If you were thinking of the diagnosis of COPD and you find out that the patient has a 50 pack-year smoking history then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

17. If you were thinking of the diagnosis of congestive heart failure and you find the following evidience, the patient has clear lungs on physical exam then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

18. If you were thinking of the diagnosis of anemia and you find the following evidence, history of colon cancer in the patient's brother then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

19. If you were thinking of the diagnosis of acute coronary syndrome and you find the following evidence, ST depressions in anterior leads on EKG then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

20. If you were thinking of the diagnosis of congestive heart failure and you find the following evidence, bilateral pulmonary edema on chest x-ray then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #5

21. If you were thinking of the diagnosis of asthma and you find out that the history of multiple episodes of wheezing then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

22. If you were thinking of the diagnosis of bronchiolitis and you find the following evidience, the patient has siblings at home with similar symptoms then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

23. If you were thinking of the diagnosis of bacterial pneumonia and you find the following evidence, rales in right lower lobe then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

24. If you were thinking of the diagnosis of aspirated foreign body and you find the following evidence, normal chest x-ray then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

25. If you were thinking of the diagnosis of cystic fibrosis and you find the following evidence, the patient is <5th percentile for weight and history of chronic diarrhea then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #6

26. If you were thinking of the diagnosis of diverticulitis and you find the following evidence, white blood cell count of 13,000 cells/µL (normal reference of 4,500-10,000 cells/µL then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

27. If you were thinking of the diagnosis of acute pancreatitis and you find the following evidence, history of alcohol binge drinking then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

28. If you were thinking of the diagnosis of diabetic ketoacidosis and you find the following evidience, blood glucose level of 375 mg/dL (normal reference of 70-150 mg/dL) then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

29. If you were thinking of the diagnosis of acute appendicitis and you find the following evidence, WBC count of 7,000 cells/µL (normal reference range 4,500 - 10,000/µL) then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

30. If you were thinking of the diagnosis of bowel obstruction and you find out the following evidence, midline abdominal scar, then this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #7

31. If you were thinking of the following diagnosis cirrhosis and you find the following evidence he has a history of alcoholism this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

32. If you were thinking of the following diagnosis ascites and you find the following evidence serum protein of 64 g/L (normal reference of 62-76 g/L) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

33. If you were thinking of the following diagnosis hepatic carcinoma and you find the following evidence carcinoembryonic antigen (CEA) of 2ng/ml (normal reference of <2.5 ng/ml) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

34. If you were thinking of the following diagnosis hemochromatosis and you find the following evidence serum iron of 55 µmol/L (normal reverence of 12-30 µmol/L) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #8

35. If you were thinking of the following diagnosis anemia and you find the physical exam shows normal conjunctiva this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

36. If you were thinking of the following diagnosis iron deficiency and you find the following evidence mean corpuscular volume (MCV) of 70 femtoliters (normal reference range 80-90 femtoliters) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

37. If you were thinking of the following diagnosis peptic ulcer and you find the following evidence heartburn does not improve with oral antacids this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

38. If you were thinking of the following diagnosis colon cancer and you find the following evidence occult blood negative this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #9

39. If you were thinking of the following diagnosis hepatitis and you find scleral icterus this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

40. If you were thinking of the following diagnosis pneumonia and you find the following evidence decreased breath sounds, right lower lung this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

41. If you were thinking of the following diagnosis duodenal ulcer and you find the following evidence occult blood in stool this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

42. If you were thinking of the following diagnosis cholangitis and you find the following evidence hypotension this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #10

43. If you were thinking of the following diagnosis simple back strain (lumbago) and you find the following evidence absence of point tenderness over the left side spine on physical exam this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

44. If you were thinking of the following diagnosis herniated intervertebral disc and you find the following evidence inability to produce pain in either lower extremity during straight leg raising maneuver this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

45. If you were thinking of the following diagnosis a spinal/paraspinal malignancy and you find the following evidence normal X-rays of the lumbar spine and pelvis this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

46. If you were thinking of the following diagnosis a paraspinal abscess and you find the following evidence an erythrocyte sedimentation rate (ESR) exceeding 100 mm/hr (the upper limit of normal in the elderly is 35-40 mm/hr) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

47. If you were thinking of the following diagnosis a lumbar vertebral compression fracture and you find the following evidence the patient gives a prior history of osteoporosis for which he takes a bisphosphonate this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #11

48. If you were thinking of the following diagnosis bowel blockage and you find the following evidence plain film shows no dilated loops of small bowel or air-fluid levels this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

49. If you were thinking of the following diagnosis obstructed common bile duct and you find the following evidence liver function tests normal this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

50. If you were thinking of the following diagnosis pancreatitis and you find the following evidence tenderness in upper right and left quadrant this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

51. If you were thinking of the following diagnosis acute lead poisoning and you find the following evidence 24-hour delta-ALA of 14 mg (normal reverence of 1.5-7.5 mg/24 hours) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

52. If you were thinking of the following diagnosis duodenal ulcer and you find the following evidence no tarry stools this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #12

53. If you were thinking of the following diagnosis pleurisy and you find the following evidence chest wall tenderness this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

54. If you were thinking of the following diagnosis pulmonary embolism and you find the following evidence history of oral contraceptive use this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

55. If you were thinking of the following diagnosis congestive heart failure and you find the following evidence jugular venous distention (JVD) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

56. If you were thinking of the following diagnosis pulmonary embolism and you find the following evidence increased alveolar-arterial gradient this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

57. If you were thinking of the following diagnosis asthma and you find the following evidence bilateral wheezing on lung exam this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #13

58. If you were thinking of the following diagnosis community acquired pneumonia and you find the following evidence normal lung sounds on exam this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

59. If you were thinking of the following diagnosis pulmonary embolism and you find the following evidence swelling and pain in left leg for one week this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

60. If you were thinking of the following diagnosis acute bronchitis and you find the following evidence bilateral supraclavicular lymphadenopathy this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

61. If you were thinking of the following diagnosis costochondritis and you find the following evidence pain reproducible with palpation of chest wall this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

62. If you were thinking of the following diagnosis spontaneous pneumothorax and you find the following evidence sudden onset of pain this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #14

63. If you were thinking of the following diagnosis splenic sequestration and you find the following evidence hemoglobin (Hgb) of 13.5 g/dL (normal reference of 11-16 g/dL for children) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

64. If you were thinking of the following diagnosis new onset menses and you find the following evidence no history of vaginal bleeding this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

65. If you were thinking of the following diagnosis sickle cell crisis and you find the following evidence grater than 25% hemoglobin S (normal reference range is 0% Hgb S) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

66. If you were thinking of the following diagnosis sickle cell crisis and you find the following evidence patient and mother state that pain is unlike previous crisis this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

67. If you were thinking of the following diagnosis cholelithiasis and you find the following evidence gallbladder ultrasound shows “sludge” this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #15

68. If you were thinking of the following diagnosis food poisoning and you find the following evidence no one at home has similar symptoms this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

69. If you were thinking of the following diagnosis metabolic acidosis and you find the following evidence blood pH 7.2 (normal reference of pH 7.35-7.45) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

70. If you were thinking of the following diagnosis pregnancy and you find the following evidence patient has breast fed 3 month old infant this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

71. If you were thinking of the following diagnosis diabetic ketoacidosis and you find the following evidence urine ketones positive (normal negative) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

Case #16

72. If you were thinking of the following diagnosis hepatitis and you find the following evidence AST of 37 U/L (normal reference of 0-35 U/L) and ALT of 35 U/L (normal reference of 0-35 U/L) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

73. If you were thinking of the following diagnosis intermittent porphyria and you find the following evidence urine bilirubin positive (normal negative) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

74. If you were thinking of the following diagnosis acute cholecystitis and you find the following evidence Murphy's sign on physical exam this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

75. If you were thinking of the following diagnosis blocked bile duct and you find the following evidence icteric sclera (normal-no discoloration of the sclera) this diagnosis becomes?
    A. -2=almost ruled out
    B. -1=somewhat less probable
    C. 0= neither less or more probable
    D. +1=somewhat more probable
    E. +2=almost certain

APPENDIX B

Emergency Medicine Script Concordance Test

CASE 1: A 52 yo Hispanic female with a past medical history of hypercholesterolemia, COPD and hypertension presents to the Emergency Department with a chief complaint of chest pain for 3 hours. The pain is sharp, sub-sternal, radiates to both arms and is associated with diaphoresis and nausea.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
31	Acute Myocardial Infarction	Normal Stress Test 6 months ago	-2 -1 0 +1 +2
32	Aortic Dissection	Normal Mediastinum on Chest X-Ray	-2 -1 0 +1 +2

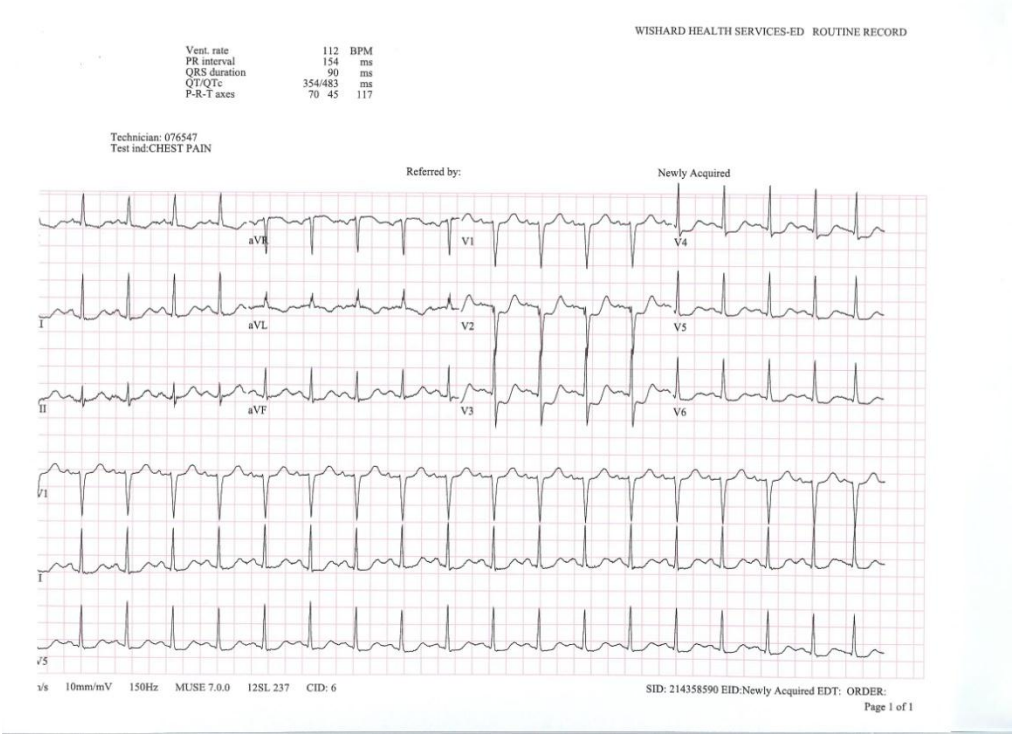
-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
33	IV Thrombolytics	The EKG seen below in Figure 1	-2 -1 0 +1 +2
34	Aspirin	History of GERD	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

Figure 1 EKG



CASE 2: A 35 yo female patient presents to the Emergency Department with the chief complaint of chest pain and shortness of breath for the last 2 days. The symptoms began suddenly and the pain is worse with deep breathing and located on the left side.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
35	Pulmonary Embolism	History of OCP use and smoking	-2 -1 0 +1 +2
36	Pneumonia	Normal lung sounds on exam	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
37	d-dimer (elisa)	Tachycardia and room air oxygen sat of 92%	-2 -1 0 +1 +2
38	Chest X-ray	Productive cough	-2 -1 0 +1 +2
39	Ultrasound of lower extremities for DVT	Normal PE-protocol chest CT	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

CASE 3: 53 yo male patient presents with one day of abdominal pain, nausea and abdominal bloating. He denies fever, vomiting and prior episodes of symptoms. He has had 3-4 episodes of non-bloody diarrhea over the last 12 hrs.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
40	Gastroenteritis	Bilateral lower quadrant tenderness and guarding on exam	-2 -1 0 +1 +2
41	Diverticulitis	History of diverticulosis on colonoscopy	-2 -1 0 +1 +2
42	Partial small bowel obstruction	No prior abdominal surgeries	-2 -1 0 +1 +2
43	Appendicitis	Normal appetite	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
44	CT Scan of the Abdomen/Pelvis	left lower quadrant tenderness with voluntary guarding	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

CASE 4: A 22 yo female presents to the Emergency Department complaining of lower abdominal pain for the last 12 hours. She describes the pain as sharp in the right lower quadrant. She has some nausea with one episode of vomiting. Her LMP was 6 weeks prior but she is irregular. She has only one sexual partner, who is male.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
45	Appendicitis	Normal WBC count	-2 -1 0 +1 +2
46	Tubo-Ovarian Abscess	Unilateral right adenexal tenderness with palpable mass on pelvic exam	-2 -1 0 +1 +2
47	Urinary Tract Infection	History of dysuria and frequency	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
48	Pelvic Ultrasound	Serum hCG = 650 mIU/ml	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering asking for…	…and you find the following evidence….	…this treatment becomes…
49	IV morphine	Positive urine pregnancy test	-2 -1 0 +1 +2
50	Ceftriaxone and azithromycin	Bilateral adnexal and cervical motion tenderness	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

CASE 5: 36 yo male patient presents to the Emergency Department with a chief complaint of headache for the last 2 days. He states he has never had a similar headache in the past. It is described as sharp and throbbing. The pain started suddenly and was maximal in intensity at the onset. He has no past medical history.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
51	Subarachnoid Hemorrhage	Normal non-contrast Head CT	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
52	Lumbar Puncture	Temperature of 102.4 orally	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

CASE 6: A 17 yo male patient was brought to the Emergency Department after being apprehended by the local police. He was driving a stolen car and wrecked the car into a telephone pole, exited the car and fled on foot. He was brought down by the police dog. The officer with him states he thinks he is "on something."

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
53	Cocaine Intoxication	Tachycardia, and dilated pupils bilaterally	-2 -1 0 +1 +2
54	Traumatic Intracranial Hemorrhage	Scalp contusion and GCS 12	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
55	Wound closure with sutures	5 cm gaping facial laceration from dog bite	-2 -1 0 +1 +2
56	Amoxicillin/clavulanate	superficial wounds from a dog bite	-2 -1 0 +1 +2
57	Benzodiazepines	Agitation, tachycardia, and diaphoresis	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

CASE 7: A 2 ½ month old female child is brought in to the Emergency Department with a fever of 102.6 F axillary at home for one day. The parents report no other symptoms. Patient was not taking formula well earlier but just took 4 oz in the ED. Temp on arrival was 102.8 F rectal. The child's immunizations are up to date.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
58	Coxsackie Virus Infection	Ulcerative lesions in the mouth and on the tongue	-2 -1 0 +1 +2
59	Urinary Tract Infection	Non toxic child with normal exam	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
60	Chest X-ray	Room air oxygen saturation of 98%	-2 -1 0 +1 +2
61	Lumbar puncture	Positive RSV nasal wash	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
62	IM Ceftriaxone and discharge home	WBC count 22,000	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

CASE 8: A 44 yo male patient presents to the Emergency Department with a chief complaint of "not acting right." The patient seems slightly confused when you talk to him. He has diffuse tremor. He reports chronic heavy alcohol use.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
63	Hypoglycemia	History of Diabetes on oral hypoglycemic agents	-2 -1 0 +1 +2
64	Ethanol Withdrawal Syndrome	Blood alcohol 96 mg/dl	-2 -1 0 +1 +2
65	Stroke	Asterixis	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
66	Non Contrast Head CT	Non-focal neurologic examination	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
67	IV magnesium	EKG with QT interval of 510 ms	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

CASE 9: A 60 yo male presents to the Emergency Department in cardiac arrest. He was found down at home by his family members who called 911 and started CPR immediately. On arrival to the Emergency Department the patient is intubated and CPR is in progress. The medics report PEA as being their last cardiac rhythm.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
68	Cardiac Tamponade	No pericardial fluid on bedside ultrasound	-2 -1 0 +1 +2

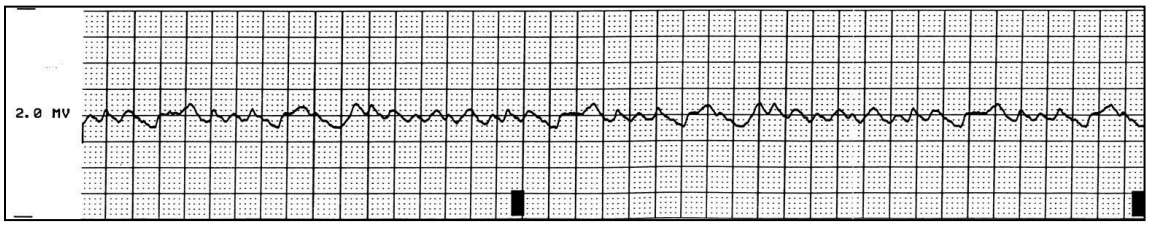
-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
69	Defibrillation	The rhythm strip seen in Figure 1 below	-2 -1 0 +1 +2
70	Additional Epinephrine and Atropine	Total time without pulse of 30 minutes	-2 -1 0 +1 +2
71	Calcium Chloride	AV Fistula in Left Arm	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

Figure 1 Rhythm Strip



CASE 10: A 44 year-old male with a history of asthma presents with 36 hours of cough and progressively worsening dyspnea and wheezing. He denies chest pain or fever.

Given the above case scenario please answer the following questions:

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
72	Chest X-ray	Symmetric wheezing on auscultation	-2 -1 0 +1 +2
73	Arterial Blood Gas	Room air oxygen saturation 89%	-2 -1 0 +1 +2
74	Complete Blood Count	Productive cough	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less Useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
75	IV magnesium	Mild symptoms. Wheezing. No respiratory distress.	-2 -1 0 +1 +2
76	Systemic corticosteroids	Diffuse wheezing, respiratory rate 28	-2 -1 0 +1 +2
77	Azithromycin	History of smoking	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

CASE 11: A 36 year-old female arrives by ambulance after being found unconscious in her home by her boyfriend. Her only past medical history is depression, for which she takes amitriptyline. She was last seen yesterday, at which time she was awake, alert, and asymptomatic.

Given the above case scenario please answer the following questions:

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
78	Head CT	Nuchal rigidity	-2 -1 0 +1 +2
79	Urine drug screen	History of marijuana use	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less Useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
80	IV flumazenil	Empty bottle of diazepam found at scene	-2 -1 0 +1 +2
81	Endotracheal intubation	GCS = 7	-2 -1 0 +1 +2
82	Sodium Bicarbonate	QRS interval 126 ms	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

CASE 12: A 27 year-old female is transported to the emergency department by EMS after a motor vehicle accident. She was a restrained driver involved in a head-on collision at 50 MPH. She was not ejected. She complains of abdominal and chest pain and has a 3 cm forehead laceration.

Given the above case scenario please answer the following questions:

Diagnostic Questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
83	Aortic injury	Normal supine AP CXR	-2 -1 0 +1 +2
84	Intra-abdominal injury	Abdominal seatbelt sign	-2 -1 0 +1 +2
85	Spleen Laceration	Negative ultrasound (FAST)	-2 -1 0 +1 +2

-2-Highly Unlikely
-1-Less likely than before
0-Neither more nor less likely
+1-More likely than before
+2-Very Likely

Investigational Questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
86	Radiographic evaluation of the Cervical Spine	Mild cervical midline tenderness. Non-focal neurologic examination	-2 -1 0 +1 +2
87	Non Contrast Head CT	GCS 14	-2 -1 0 +1 +2

-2-Not useful at all
-1-Less useful
0-Neither more nor less useful
+1-Useful
+2-Absolutely Necessary

Therapeutic Questions

	If you were considering treating with…	…and you find the following evidence….	…that treatment becomes…
88	RhoGAM	First trimester pregnancy (patient's blood type A-)	-2 -1 0 +1 +2
89	Fentanyl for analgesia	BP 98/72	-2 -1 0 +1 +2

-2-Contraindicated totally or almost totally
-1-Not useful; possibly detrimental
0-Neither more nor less useful
+1-Useful
+2-Necessary or absolutely necessary

APPENDIX C

Script Concordance Test Item Survey

The purpose of this survey is to obtain EM physician perceptions on the nature and difficulty of script concordance test items. Participation in this study is voluntary and no benefits or risks are known to be associated with completing this survey.

By entering this survey you are providing consent to participate in this research. Please complete the survey by following the instructions.

Select the month in which you were born to determine which version of the survey you will receive.*

- ☐ January
- ☐ February
- ☐ March
- ☐ April
- ☐ May
- ☐ June
- ☐ July
- ☐ August
- ☐ September
- ☐ October
- ☐ November
- ☐ December

*Participants who selected an odd birth month received Part A of the survey and those who selected an even birth month received Part B of the survey.

SURVEY: PART A

Case 1: A 36 year-old female arrives by ambulance after being found unconscious in her home by her boyfriend. Her only past medical history is depression, for which she takes amitriptyline. She was last seen yesterday, at which time she was awake, alert, and asymptomatic.

Item 1: if you were considering treating with iv flumazenil and you find an empty bottle of diazepam at the scene. The treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 2: if you were considering treating with endotracheal intubation and you find gcs=7, the treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 3: if you were considering treating with sodium bicarbonate and you find QRS interval 126ms, the treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Case 2: a 27 year-old female is transported to the emergency department by EMS after a motor vehicle accident. She was a restrained driver involved in a head-on collision at 50 mph. She was not ejected. She complains of abdominal and chest pain and has a 3 cm forehead laceration.

Item 4: if you were thinking of the following diagnosis - aortic injury - and you find normal supine AP CXT, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 5: if you were thinking of the following diagnosis - spleen laceration - and you find a negative ultrasound (fast), the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 6: if you were considering treating with fentanyl for analgesia and you find a bp of 98/72, the treatment becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Case 3: a 17 yo male patient was brought to the emergency department after being apprehended by the local police. He was driving a stolen car and wrecked the car into a telephone pole, exited the car and fled on foot. He was brought down by the police dog. The officer with him states he thinks he is "on something."

Item 7: if you were thinking of the following diagnosis - cocaine intoxication - and you find tachycardia and dilated pupils bilaterally, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 8: if you were thinking of the following diagnosis - traumatic intracranial hemorrhage - and you find a scalp contusion and gcs=12, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 9: if you were considering treating with amoxicillin/clavulanate and you find superficial wounds from a dog bite, the treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Case 4: 53 yo male patient presents with one day of abdominal pain, nausea and abdominal bloating. He denies fever, vomiting and prior episodes of symptoms. He has had 3-4 episodes of non-bloody diarrhea over the last 12 hrs.

Item 10: if you were thinking of the following diagnosis - gastroenteritis - and you find bilateral lower quadrant tenderness and guarding on exam, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 11: if you were thinking of the following diagnosis - diverticulitis - and you find a history of diverticulosis on colonoscopy, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

Item 12: if you were thinking of the following diagnosis - partial small bowel obstruction - and you find no prior abdominal surgeries, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Easy
- ☐ Moderate
- ☐ Difficult

SURVEY: PART B

Case 1: A 36 year-old female arrives by ambulance after being found unconscious in her home by her boyfriend. Her only past medical history is depression, for which she takes amitriptyline. She was last seen yesterday, at which time she was awake, alert, and asymptomatic.

Item 1: if you were considering treating with iv flumazenil and you find an empty bottle of diazepam at the scene. The treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 2: if you were considering treating with endotracheal intubation and you find gcs=7, the treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 3: if you were considering treating with sodium bicarbonate and you find QRS interval 126ms, the treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Which of the above 3 items is (are) most ambiguous or ill-defined? (Check all that apply.)

- ☐ Item 1
- ☐ Item 2
- ☐ Item 3
- ☐ None are ambiguous or ill-defined

Case 2: a 27 year-old female is transported to the emergency department by EMS after a motor vehicle accident. She was a restrained driver involved in a head-on collision at 50 mph. She was not ejected. She complains of abdominal and chest pain and has a 3 cm forehead laceration.

Item 4: if you were thinking of the following diagnosis - aortic injury - and you find normal supine AP CXT, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 5: if you were thinking of the following diagnosis - spleen laceration - and you find a negative ultrasound (fast), the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 6: if you were considering treating with fentanyl for analgesia and you find a bp of 98/72, the treatment becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Which of the above 3 items is (are) most ambiguous or ill-defined? (Check all that apply.)

- ☐ Item 4
- ☐ Item 5
- ☐ Item 6
- ☐ None are ambiguous or ill-defined

Case 3: a 17 yo male patient was brought to the emergency department after being apprehended by the local police. He was driving a stolen car and wrecked the car into a telephone pole, exited the car and fled on foot. He was brought down by the police dog. The officer with him states he thinks he is "on something."

Item 7: if you were thinking of the following diagnosis - cocaine intoxication - and you find tachycardia and dilated pupils bilaterally, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 8: if you were thinking of the following diagnosis - traumatic intracranial hemorrhage - and you find a scalp contusion and gcs=12, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 9: if you were considering treating with amoxicillin/clavulanate and you find superficial wounds from a dog bite, the treatment becomes...

-2=contraindicated totally or almost totally
-1=not useful, possibly detrimental
0=neither more or less useful
+1=useful
+2=necessary or absolutely necessary

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Which of the above 3 items is (are) most ambiguous or ill-defined? (Check all that apply.)

- ☐ Item 7
- ☐ Item 8
- ☐ Item 9
- ☐ None are ambiguous or ill-defined

Case 4: 53 yo male patient presents with one day of abdominal pain, nausea and abdominal bloating. He denies fever, vomiting and prior episodes of symptoms. He has had 3-4 episodes of non-bloody diarrhea over the last 12 hrs.

Item 10: if you were thinking of the following diagnosis - gastroenteritis - and you find bilateral lower quadrant tenderness and guarding on exam, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 11: if you were thinking of the following diagnosis - diverticulitis - and you find a history of diverticulosis on colonoscopy, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Item 12: if you were thinking of the following diagnosis - partial small bowel obstruction - and you find no prior abdominal surgeries, the hypothesis becomes...

-2=highly unlikely
-1=less likely than before
0=neither more nor less likely
+1=more likely than before
+2=very likely

This item is:

- ☐ Straight forward or non-ambiguous.
- ☐ Somewhat ambiguous or somewhat ill-defined.
- ☐ Very ambiguous or very ill-defined.

Which of the above 3 items is (are) most ambiguous or ill-defined? (Check all that apply.)

- ☐ Item 10
- ☐ Item 11
- ☐ Item 12
- ☐ None are ambiguous or ill-defined

APPENDIX D.1

Table D.1: Schmid-Leiman solution for scoring method A (SCT-PS exam)

		First-order Factors		
ITEM	SECOND	I	II	III
Q01C01T1	.297	.164	.055	.046
Q02C01T1	.356	.198	.039	.096
Q04C01T1	.079	-.001	.206	-.213
Q06C01T1	.368	.054	.202	.114
Q10C02T1	.230	-.082	.298	.004
Q13C03T1	.400	-.040	.320	.141
Q14C03T1	.170	.036	.059	.086
Q15C03T1	.244	.156	.036	.012
Q16C04T1	.361	.119	.206	-.021
Q17C04T1	.370	.163	-.023	.276
Q19C04T1	.160	.067	.034	.056
Q21C05T1	.292	.104	.212	-.104
Q22C05T1	.291	.096	.142	.021
Q23C05T1	.320	.108	.095	.116
Q24C05T1	.040	.058	.033	-.098
Q25C05T1	.352	.261	.033	-.016
Q26C06T1	.216	-.013	.034	.280
Q27C06T1	.408	.029	.357	-.030
Q28C06T1	.266	.083	.046	.163
Q29C06T1	.313	.224	-.089	.189
Q30C06T1	.368	-.089	.380	.088
Q31C07T1	.462	.106	.248	.082
Q32C07T1	.126	.053	.166	-.178
Q33C07T1	.139	.156	-.030	-.034
Q34C07T1	.262	.229	.036	-.091
Q35C08T1	.285	-.006	.053	.337
Q36C08T1	.359	.257	.092	-.091
Q37C08T1	.244	.033	-.034	.349
Q38C08T1	.286	.244	-.032	.025
Q39C09T1	.329	.014	.329	-.072
Q40C09T1	.455	.171	.144	.120
Q41C09T1	.376	-.056	.317	.140
Q42C09T1	.260	.129	-.045	.214
Q43C10T1	.006	.026	.073	-.155
Q44C10T1	.347	.381	-.169	.085
Q45C10T1	.181	.159	-.024	.012
Q46C10T1	.309	.013	.097	.270
Q47C10T1	.405	.090	.298	-.052
Q50C11T1	.080	-.055	.217	-.131
Q52C11T1	.035	.036	.092	-.162
Q53C12T1	.189	.081	.065	.024
Q54C12T1	.189	-.213	.358	.088
Q55C12T1	.338	.188	.125	-.048
Q56C12T1	.365	.213	.030	.095
Q57C12T1	.333	.194	.151	-.109
Q58C13T1	.256	.181	-.026	.085
Q59C13T1	.378	.263	.103	-.092
Q61C13T1	.284	.244	.036	-.088
Q62C13T1	.437	.170	.173	.049
Q65C14T1	.317	.136	.124	.016
Q66C14T1	.235	.013	-.015	.339
Q67C14T1	.375	.171	.008	.219
Q68C15T1	-.020	.202	-.075	-.274
Q70C15T1	.159	.214	-.107	.014
Q71C15T1	.280	.134	.159	-.090
Q73C16T1	.186	.031	.181	-.077
Q74C16T1	.364	.298	-.019	.020
Q75C16T1	.405	.084	.167	.168

APPENDIX D.2

Table D.2: Schmid-Leiman solution for scoring method B (SCT-PS exam)

		First-order Factors		
ITEM	SECOND	I	II	III
Q01C01T1	.148	-.005	.100	.043
Q02C01T1	.345	.020	.214	.095
Q04C01T1	.208	-.002	-.082	.351
Q06C01T1	.253	-.029	.230	.017
Q10C02T1	.031	-.022	.087	-.058
Q13C03T1	.101	.041	-.025	.107
Q14C03T1	.337	.007	.068	.293
Q15C03T1	.139	.068	-.005	.098
Q16C04T1	.364	.019	.317	-.019
Q17C04T1	.106	-.054	.115	.025
Q19C04T1	.218	.031	.135	.041
Q21C05T1	.296	.032	.283	-.067
Q22C05T1	.046	-.074	.169	-.095
Q23C05T1	.433	-.005	.185	.259
Q24C05T1	.003	-.011	-.007	.025
Q25C05T1	.351	.029	.148	.179
Q26C06T1	.213	.005	-.063	.324
Q27C06T1	.248	.017	.051	.201
Q28C06T1	.270	.023	.171	.063
Q29C06T1	.273	.044	.097	.142
Q30C06T1	-.046	-.028	-.142	.163
Q31C07T1	.418	.024	.318	.036
Q32C07T1	-.131	-.003	.100	-.280
Q33C07T1	.210	-.032	.066	.186
Q34C07T1	.367	.022	.304	-.002
Q35C08T1	-.077	.001	-.140	.095
Q36C08T1	.257	.038	.199	-.006
Q37C08T1	.088	-.059	.112	.014
Q38C08T1	.211	-.017	.081	.152
Q39C09T1	.341	.050	.217	.056
Q40C09T1	.356	.014	.182	.156
Q41C09T1	.276	.003	.041	.262
Q42C09T1	-.045	.015	-.055	.005
Q43C10T1	.091	.062	.112	-.105
Q44C10T1	.216	-.055	.275	-.058
Q45C10T1	.183	-.024	.132	.060
Q46C10T1	.294	-.017	.052	.287
Q47C10T1	.410	-.007	.307	.074
Q50C11T1	.174	-.004	.009	.193
Q52C11T1	.017	-.018	.102	-.098
Q53C12T1	-.105	-.008	.065	-.200
Q54C12T1	.211	-.071	.003	.311
Q55C12T1	.334	.006	.301	-.020
Q56C12T1	.212	.044	.046	.140
Q57C12T1	.359	.000	.371	-.077
Q58C13T1	.419	.500	.055	-.098
Q59C13T1	.421	.480	-.023	.028
Q61C13T1	.516	.482	.103	-.031
Q62C13T1	.528	.534	-.042	.122
Q65C14T1	.497	.507	.074	-.040
Q66C14T1	.380	.497	-.133	.110
Q67C14T1	.474	.509	.070	-.064
Q68C15T1	.398	.527	-.018	-.053
Q70C15T1	.401	.522	-.051	.000
Q71C15T1	.461	.504	.067	-.070
Q73C16T1	.411	.671	-.121	-.048
Q74C16T1	.512	.491	.022	.063
Q75C16T1	.494	.486	-.016	.098

APPENDIX D.3

Table D.3: Schmid-Leiman solution for scoring method C (SCT-PS exam)

		First-order Factors		
ITEM	SECOND	I	II	III
Q01C01T1	.274	.174	.140	.001
Q02C01T1	.277	.178	.006	.158
Q04C01T1	.128	.266	.090	-.141
Q06C01T1	.354	.149	.171	.061
Q10C02T1	.129	.046	.159	-.085
Q13C03T1	.498	.065	.360	.035
Q14C03T1	.218	.246	.149	-.110
Q15C03T1	-.135	.207	-.275	.059
Q16C04T1	.345	.286	.142	.001
Q17C04T1	.330	.109	-.003	.264
Q19C04T1	.080	.255	-.049	-.021
Q21C05T1	.314	.300	.132	-.027
Q22C05T1	.272	.180	.122	.017
Q23C05T1	.208	.295	.023	-.002
Q24C05T1	.084	.122	.043	-.042
Q25C05T1	.268	.335	-.010	.071
Q26C06T1	.221	-.060	.125	.112
Q27C06T1	.403	.244	.224	-.011
Q28C06T1	.246	.052	.016	.195
Q29C06T1	.361	.025	.073	.259
Q30C06T1	.479	.043	.302	.097
Q31C07T1	.443	.233	.164	.105
Q32C07T1	-.009	.245	.006	-.168
Q33C07T1	.144	-.031	-.022	.188
Q34C07T1	.166	.305	-.034	.016
Q35C08T1	.317	-.166	.032	.381
Q36C08T1	.302	.351	.035	.042
Q37C08T1	.034	-.031	-.025	.082
Q38C08T1	.240	.227	-.130	.250
Q39C09T1	.348	.240	.256	-.102
Q40C09T1	.296	.274	.077	.035
Q41C09T1	.432	-.045	.292	.117
Q42C09T1	.197	.106	.016	.111
Q43C10T1	.065	.071	.089	-.083
Q44C10T1	.217	.362	-.164	.183
Q45C10T1	.125	.019	-.080	.207
Q46C10T1	.367	-.078	.129	.263
Q47C10T1	.363	.335	.158	-.031
Q50C11T1	.208	.088	.144	-.016
Q52C11T1	.093	.221	.014	-.061
Q53C12T1	.062	.293	-.034	-.080
Q54C12T1	.365	-.104	.389	-.026
Q55C12T1	.288	.329	.076	-.006
Q56C12T1	.204	.260	.002	.039
Q57C12T1	.254	.431	-.013	.001
Q58C13T1	.234	.109	-.075	.252
Q59C13T1	.330	.396	.015	.066
Q61C13T1	.154	.395	-.084	.006
Q62C13T1	.348	.245	.115	.061
Q65C14T1	.209	.186	.019	.071
Q66C14T1	.303	-.200	.136	.267
Q67C14T1	.186	.082	.010	.123
Q68C15T1	-.098	.316	-.174	-.090
Q70C15T1	.315	-.076	.026	.330
Q71C15T1	.243	.245	.075	.002
Q73C16T1	.150	.220	.042	-.036
Q74C16T1	.251	.377	-.067	.095
Q75C16T1	.209	.059	.008	.162

APPENDIX D.4

Table D.4: Schmid-Leiman solution for scoring method D (SCT-PS exam)

		First-order Factors		
ITEM	SECOND	I	II	III
Q01C01T1	.230	.225	.055	.009
Q02C01T1	.388	.072	.166	.130
Q04C01T1	.047	.286	.014	-.158
Q06C01T1	.321	.276	.018	.109
Q10C02T1	.058	.316	-.109	-.021
Q13C03T1	.304	.431	-.096	.125
Q14C03T1	.174	.265	.051	-.066
Q15C03T1	.176	-.107	.224	-.027
Q16C04T1	.304	.320	.064	.007
Q17C04T1	.445	-.030	.161	.257
Q19C04T1	.169	.091	.093	-.006
Q21C05T1	.263	.297	.087	-.044
Q22C05T1	.242	.222	.047	.032
Q23C05T1	.298	.158	.111	.052
Q24C05T1	.065	.088	.051	-.056
Q25C05T1	.377	.135	.221	.013
Q26C06T1	.241	-.012	.019	.218
Q27C06T1	.306	.457	-.007	.005
Q28C06T1	.304	.030	.087	.171
Q29C06T1	.417	-.039	.181	.213
Q30C06T1	.281	.412	-.086	.104
Q31C07T1	.417	.310	.077	.108
Q32C07T1	-.004	.240	.018	-.182
Q33C07T1	.251	-.035	.136	.103
Q34C07T1	.274	.121	.195	-.047
Q35C08T1	.346	-.033	.025	.326
Q36C08T1	.366	.205	.210	-.031
Q37C08T1	.225	-.083	.039	.225
Q38C08T1	.377	-.032	.268	.065
Q39C09T1	.193	.473	-.049	-.065
Q40C09T1	.387	.229	.130	.069
Q41C09T1	.301	.346	-.081	.160
Q42C09T1	.302	-.026	.120	.166
Q43C10T1	-.009	.125	.018	-.112
Q44C10T1	.462	-.090	.350	.088
Q45C10T1	.226	-.037	.158	.054
Q46C10T1	.358	.052	.036	.269
Q47C10T1	.284	.399	.058	-.056
Q50C11T1	.092	.236	-.028	-.033
Q52C11T1	.082	.127	.066	-.083
Q53C12T1	.125	.140	.097	-.087
Q54C12T1	.071	.408	-.238	.085
Q55C12T1	.287	.246	.131	-.040
Q56C12T1	.339	.098	.166	.065
Q57C12T1	.323	.214	.206	-.074
Q58C13T1	.330	-.033	.195	.108
Q59C13T1	.380	.233	.203	-.027
Q61C13T1	.274	.142	.204	-.071
Q62C13T1	.377	.245	.132	.046
Q65C14T1	.308	.120	.140	.053
Q66C14T1	.330	-.045	.015	.330
Q67C14T1	.380	.002	.161	.173
Q68C15T1	-.012	-.007	.188	-.230
Q70C15T1	.294	-.079	.185	.115
Q71C15T1	.210	.255	.095	-.077
Q73C16T1	.130	.231	.026	-.056
Q74C16T1	.408	.060	.275	.028
Q75C16T1	.314	.048	.110	.142

APPENDIX D.5

Table D.5: Schmid-Leiman solution for scoring method E (SCT-PS exam)

		First-order Factors		
ITEM	SECOND	I	II	III
Q01C01T1	.223	-.014	.069	.058
Q02C01T1	.298	.017	.075	.064
Q04C01T1	.486	.054	.145	.006
Q06C01T1	.233	.110	.030	-.032
Q10C02T1	-.059	-.023	-.070	.158
Q13C03T1	.239	.013	.096	-.037
Q14C03T1	.495	.101	.090	.077
Q15C03T1	.109	-.034	.028	.086
Q16C04T1	.262	.107	-.042	.182
Q17C04T1	.349	.013	.044	.195
Q19C04T1	.326	.001	.113	.020
Q21C05T1	.096	-.021	-.003	.131
Q22C05T1	.129	-.023	.049	.035
Q23C05T1	.574	.172	.075	.075
Q24C05T1	.112	-.011	-.003	.131
Q25C05T1	.242	-.061	.053	.191
Q26C06T1	.458	.041	.100	.114
Q27C06T1	.170	-.019	.077	-.004
Q28C06T1	.360	.075	.007	.200
Q29C06T1	.316	.100	.014	.102
Q30C06T1	.242	-.006	.122	-.072
Q31C07T1	.557	.409	-.064	.035
Q32C07T1	-.058	-.087	-.058	.230
Q33C07T1	.423	.262	.009	-.043
Q34C07T1	.403	.301	-.024	-.041
Q35C08T1	.261	.126	.034	-.042
Q36C08T1	.069	-.002	.044	-.042
Q37C08T1	.098	-.051	-.009	.195
Q38C08T1	.413	.245	.014	-.037
Q39C09T1	.286	.005	.104	.000
Q40C09T1	.296	-.001	.064	.119
Q41C09T1	.223	.066	.060	-.046
Q42C09T1	-.026	.051	.028	-.176
Q43C10T1	-.053	.016	-.055	.062
Q44C10T1	.159	.012	-.003	.138
Q45C10T1	.516	.357	-.030	-.007
Q46C10T1	.302	.083	.118	-.145
Q47C10T1	.522	.307	.018	-.042
Q50C11T1	.281	.079	.051	.011
Q52C11T1	-.129	.074	-.083	-.028
Q53C12T1	-.236	.000	-.049	-.099
Q54C12T1	.383	-.005	.142	.010
Q55C12T1	.429	-.024	.072	.260
Q56C12T1	.347	.020	.136	-.048
Q57C12T1	.251	.041	-.061	.323
Q58C13T1	.170	.027	-.051	.244
Q59C13T1	.373	-.029	.134	.060
Q61C13T1	.471	-.037	.135	.159
Q62C13T1	.464	.052	.050	.227
Q65C14T1	.135	.016	.016	.060
Q66C14T1	.410	.067	.058	.132
Q67C14T1	.343	.072	.037	.114
Q68C15T1	-.012	.003	-.081	.188
Q70C15T1	.205	.113	-.034	.097
Q71C15T1	.163	-.031	.039	.104
Q73C16T1	.090	.020	.018	.008
Q74C16T1	.264	.021	.104	-.046
Q75C16T1	.371	.073	.101	-.022

APPENDIX D.6

Table D.6: Schmid-Leiman solution for scoring method F (SCT-PS exam)

		First-order Factors		
ITEM	SECOND	I	II	III
Q01C01T1	.168	-.028	.151	.101
Q02C01T1	.222	.061	.189	.053
Q04C01T1	.329	.035	.274	.130
Q06C01T1	.186	.150	.043	.014
Q10C02T1	-.044	-.046	-.288	.147
Q13C03T1	.124	.006	.246	-.018
Q14C03T1	.397	.196	.134	.120
Q15C03T1	.151	-.027	.075	.124
Q16C04T1	.249	.153	-.212	.199
Q17C04T1	.284	.049	.044	.193
Q19C04T1	.237	-.037	.180	.159
Q21C05T1	.099	-.044	-.081	.171
Q22C05T1	.114	-.030	.096	.082
Q23C05T1	.466	.291	.095	.119
Q24C05T1	-.062	-.128	.020	.047
Q25C05T1	.195	-.057	.149	.153
Q26C06T1	.350	.095	.232	.118
Q27C06T1	.073	-.023	.232	-.031
Q28C06T1	.322	.110	-.067	.231
Q29C06T1	.229	.205	.063	-.005
Q30C06T1	.134	-.016	.321	-.026
Q31C07T1	.436	.556	-.207	.009
Q32C07T1	-.144	-.118	-.345	.149
Q33C07T1	.281	.236	.124	-.016
Q34C07T1	.328	.460	-.112	-.053
Q35C08T1	.156	.152	.139	-.063
Q36C08T1	.059	.042	.150	-.060
Q37C08T1	.034	-.100	-.008	.124
Q38C08T1	.283	.314	.114	-.079
Q39C09T1	.194	-.017	.224	.079
Q40C09T1	.242	-.014	.100	.184
Q41C09T1	.335	.170	.157	.075
Q42C09T1	-.076	.077	.096	-.188
Q43C10T1	-.055	.048	-.101	-.042
Q44C10T1	.137	.034	-.023	.107
Q45C10T1	.223	.269	-.037	-.017
Q46C10T1	.186	.110	.280	-.070
Q47C10T1	.372	.410	.069	-.060
Q50C11T1	.216	.102	.086	.063
Q52C11T1	-.179	.004	-.169	-.082
Q53C12T1	-.209	.003	-.045	-.171
Q54C12T1	.228	-.016	.365	.037
Q55C12T1	.365	-.011	.094	.296
Q56C12T1	.225	.011	.288	.049
Q57C12T1	.279	.077	-.205	.290
Q58C13T1	.208	.155	-.057	.082
Q59C13T1	.260	.015	.346	.049
Q61C13T1	.368	-.064	.228	.277
Q62C13T1	.369	.030	.031	.294
Q65C14T1	.133	.006	-.065	.149
Q66C14T1	.334	.081	.170	.147
Q67C14T1	.271	.087	-.016	.178
Q68C15T1	-.078	-.073	-.244	.118
Q70C15T1	.182	.107	.047	.047
Q71C15T1	.116	-.063	-.052	.189
Q73C16T1	.074	.056	.029	.003
Q74C16T1	.137	.035	.295	-.055
Q75C16T1	.257	.129	.226	.005

REFERENCES

1. Charlin B, Brailovsky C, Leduc C, Blouin D. The diagnosis script questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education.* 1998;3(1):51-58.

2. Lubarsky S, Charlin B, Cook DA, Chalk C, Van Der Vleuten CPM. Script concordance testing: A review of published validity evidence. *Medical Education.* 2011;45(4):329-338.

3. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: Validity evidence for a script concordance test. *Annals of Emergency Medicine.* 2008;53(5):647-652.

4. Charlin B, Brailovsky C, Brazeau-Lamontagne L, Samson L, Leduc C, Van Der Vleuten C. Script questionnaires: Their use for assessment of diagnostic knowledge in radiology. *Medical Teacher.* 1998;20(6):567-571.

5. Humbert A. Assessing the clinical reasoning skills of emergency medicine clerkship students using a script concordance test. *Academic Emergency Medicine.* 2008;15:S230-S231.

6. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: A script concordance test designed for pre-clinical medical students. *Medical Teacher.* 2011;33(6):472-477.

7. Khonputsa P, Besinque K, Fisher D, Gong WC. Use of script concordance test to assess pharmaceutical diabetic care: A pilot study in thailand. *Medical Teacher.* 2006;28(6):570-573.

8. Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: Validation study of a new tool to assess clinical reasoning. *Radiation Oncology.* 2009;4(7).

9. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: A new tool assessing clinical judgement in neurology. *The Canadian Journal of Neurological Sciences.* 2009;36(3):326-331.

10. Ruiz JG, Tunuguntla R, Charlin B, et al. The script concordance test as a measure of clinical reasoning skills in geriatric urinary incontinence. *Journal of the American Geriatrics Society.* 2010;58(11):2178-2184.

11. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. *Medical Teacher.* 2002;24(5):522-527.

12. Sibert L, Darmoni S, Dahamna B, Hellot MF, Weber J, Charlin B. On line clinical reasoning assessment with script concordance test in urology: Results of a french pilot study. *BMC Medical Education.* 2006;6(1):45.

13. Deschênes M, Charlin B, Gagnon R, Goudreau J. Use of a script concordance test to assess development of clinical reasoning in nursing students. *The Journal of nursing education.* 2011:1.

14. Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. Assessment of clinical reasoning competence in urology with the script concordance test: An exploratory study across two sites from different countries. *European Urology.* 2002;41(3):227-233.

15. Fournier J, Demeester A, Charlin B. Script concordance tests: Guidelines for construction. *BMC Medical Informatics and Decision Making.* 2008;8(1):18.

16. Humbert AJ, Besinger B, Miech EJ. Assessing clinical reasoning skills in scenarios of uncertainty: Convergent validity for a script concordance test in an emergency medicine clerkship and residency. *Academic Emergency Medicine.* 2011;18(6):627-634.

17. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: Insights from a systematic review. *Medical Education.* 2012;46(6):552-563.

18. Charlin B, Desaulniers M, Gagnon R, Blouin D, Van Der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine.* 2002;14(3):150-156.

19. Charlin B, Gagnon R, Pelletier J, et al. Assessment of clinical reasoning in the context of uncertainty: The effect of variability within the reference panel. *Medical Education.* 2006;40(9):848-854.

20. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine.* 2005;80(4):395.

21. Groves M, O'rourke P, Alexander H. Clinical reasoning: The relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. *Medical Teacher.* 2003;25(6):621-625.

22. Chimowitz MI, Logigian EL, Caplan LR. The accuracy of bedside neurological diagnoses. *Annals of Neurology.* 1990;28(1):78-85.

23. Brailovsky C, Charlin B, Beausoleil S, Cote S, Van Der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: An experimental study on the script concordance test. *Medical Education.* 2001;35(5):430-436.

24. Bordage G, Brailovsky C, Cohen T, Page G. Maintaining and enhancing key decision-making skills from graduation into practice: An exploratory study. *Advances in Medical EducationI. Dordrecht, The Netherlands: Kluwer Academic.* 1996:128-130.

25. Feltovich P, Barrows H. Issues of generality in medical problem solving. *Tutorials in problem-based learning.* 1984:128-142.

26. Brailovsky C, Charlin B, Émond C, Maltais P. Script questionnaire as a method of assessing clinical reasoning after educational programs1999.

27. Bruning RH, Schraw GJ, Norby MM. *Cognitive psychology and instruction.* 5 ed. Boston, MA: Pearson Education, Inc.; 2011.

28. Mcclelland JL. Connectionist models and psychological evidence. *Journal of Memory and Language.* 1988;27(2):107-123.

29. Scott IA. Errors in clinical reasoning: Causes and remedial strategies. *BMJ (Clinical Research Ed.).* 2009;338:338.

30. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine.* 2006;355(21):2217-2225.

31. Hatala R, Norman GR, Brooks LR. Influence of a single example on subsequent electrocardiogram interpretation. *Teaching and Learning in Medicine.* 1999;11(2):110-117.

32. Schmidt H, Boshuizen H, Hobus P. Transitory stages in the development of medical expertise: The “intermediate effect” in clinical case representation studies. Paper presented at: Proceedings of the Tenth Annual Conference of the Cognitive Science Society1988.

33. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise - theory and implications. *Academic Medicine.* 1990;65(10):611-621.

34. Norman G. Research in clinical reasoning: Past history and current trends. *Medical Education.* 2005;39(4):418-427.

35. Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ (Clinical Research Ed.).* 2002;324(7339):729-732.

36. Charlin B, Van Der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty. *Evaluation and the Health Professions.* 2004;27(3):304-319.

37. Bordage G. Elaborated knowledge: A key to successful diagnostic thinking. *Academic Medicine.* 1994.

38. Boshuizen H, Schmidt H. The development of clinical reasoning expertise; implications for teaching. *Clinical reasoning in the health professions. 2nd completely revised edition.* 2000:15-22.

39. Boshuizen H, Schmidt HG. On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive science.* 1992;16(2):153-184.

40. Gagnon R, Charlin B, Lambert C, Carriere B, Van Der Vleuten C. Script concordance testing: More cases or more questions? *Advances in Health Sciences Education.* 2009;14(3):367-375.

41. Charlin B, Roy L, Brailovsky C, Goulet F, Van Der Vleuten C. The script concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine.* 2000;12(4):189-195.

42. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: More meaning for scores. *Teaching and Learning in Medicine.* 2010;22(3):180-186.

43. Gagnon R, Charlin B, Coletti M, Sauvé E, Van Der Vleuten C. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education.* 2005;39(3):284-291.

44. Charlin B, Gagnon R, Sauvé E, Coletti M. Composition of the panel of reference for concordance tests: Do teaching functions have an impact on examinees' ranks and absolute scores? *Medical Teacher.* 2007;29(1):49-53.

45. Thorndike RM, Thorndike-Christ T. *Measurement and evaluation in psychology and education.* 8th ed. Boston, MA: Pearson Education, Inc; 2010.

46. Handfield-Jones R, Brown JB, Biehn J, Rainsberry P, Brailovsky CA. Certification examination of the college of family physicians of canada. Part 3: Short-answer management problems. *Canadian Family Physician.* 1996;42:1353.

47. Williams RG, Klamen DL, Hoffman RM. Medical student acquisition of clinical working knowledge. *Teaching and Learning in Medicine.* 2008;20(1):5.

48. Wainer H, Mee J. On assessing the quality of physicians' clinical judgment. *Evaluation and the Health Professions.* 2004;27(4):369-382.

49. Association AER, Association AP, Education NCOMI, Educational JCOSF, Testing P. *Standards for educational and psychological testing*: Amer Educational Research Assn; 1999.

50. Haynes SN, Richard D, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment.* 1995;7(3):238.

51. Kassirer JP. Teaching clinical reasoning: Case-based and coached. *Academic Medicine.* 2010;85(7):1118.

52. Gagnon R, Charlin B, Roy L, et al. The cognitive validity of the script concordance test: A processing time study. *Teaching and Learning in Medicine.* 2006;18(1):22-27.

53. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *The American journal of surgery.* 2007;193(2):248-251.

54. Collard A, Gelaes S, Vanbelle S, et al. Reasoning versus knowledge retention and ascertainment throughout a problem - based learning curriculum. *Medical Education.* 2009;43(9):854-865.

55. Brown JB, Handfield-Jones R, Rainsberry P, Brailovsy CA. Certification examination of the college of family physicians of canada: Part 4: Simulated office orals. *Canadian Family Physician.* 1996;42:1539.

56. Lurie S. Towards greater clarity in the role of ambiguity in clinical reasoning. *Medical Education.* Apr 2011;45(4):326-328.

57. Schuwirth L, Vleuten CPM, Donkers H. A closer look at cueing effects in multiple - choice questions. *Medical Education.* 1996;30(1):44-49.

58. Charlin B, Tardif J, Boshuizen H. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine.* 2000;75(2):182.

59. Eva KW. What every teacher needs to know about clinical reasoning. *Medical Education.* 2005;39(1):98-106.

60. Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Medical education online.* 2011;16.

61. Neufeld V, Norman G, Feightner J, Barrows H. Clinical problem - solving by medical students: A cross-sectional and longitudinal analysis. *Medical Education.* 1981;15(5):315-322.

62. Williams RG, Klamen DL, White CB, et al. Tracking development of clinical reasoning ability across five medical schools using a progress test. *Academic Medicine.* 2011;86(9):1148.

63. Patel VL, Groen GJ. Knowledge based solution strategies in medical reasoning. *Cognitive science.* 1986;10(1):91-116.

64. Norman GR, Brooks LR, Colle CL, Hatala RM. The benefit of diagnostic hypotheses in clinical reasoning: Experimental study of an instructional intervention for forward and backward reasoning. *Cognition and Instruction.* 1999;17(4):433-448.

65. Nendaz MR, Bordage G. Promoting diagnostic problem representation. *Medical Education.* 2002;36(8):760-766.

66. Raidl MA, Wood OB, Lehman JD, Evers WD. Computer-assisted instruction improves clinical reasoning skills of dietetics students. *Journal of the American Dietetic Association.* 1995;95(8):868-873.

67. Bordage G, Grant J, Marsden P. Quantitative assessment of diagnostic ability. *Medical Education.* 1990;24(5):413-425.

68. Friedman MH, Connell KJ, Olthoff AJ, Sinacore JM, Bordage G. Thinking about student thinking: Medical student errors in making a diagnosis. *Academic Medicine.* 1998;73(10):S19.

69. Bordage G. Why did i miss the diagnosis? Some cognitive explanations and educational implications. *Academic Medicine.* 1999;74(10):138.

70. Aalten C, Samson M, Jansen P. Diagnostic errors; the need to have autopsies. *Netherlands Journal of Medicine.* 2006;64(6):186-190.

71. Brownlee S. *Overtreated: Why too much medicine is making us sicker and poorer*. New York: Bloomsbury USA; 2008.

72. Podbregar M, Voga G, Krivec B, Skale R, Parežnik R, Gabrščček L. Should we confirm our clinical diagnostic certainty by autopsies? *Intensive Care Medicine.* 2001;27(11):1750-1755.

73. Harrison BT, Gibberd RW, Hamilton JD, Wilson RML. An analysis of the causes of adverse events from the quality in australian health care study. *Medical Journal of Australia.* 1999;170(9):411-415.

74. Furr RM, Bacharach VR. *Psychometrics: An introduction*. Thousand Oaks, CA: Sage Publications, Inc; 2008.

75. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin.* 1955;52(4).

76. Brown TA. Confirmatory factor analysis for applied research (methodology in the social sciences). 2006.

77. Harrington D. *Confirmatory factor analysis*. Oxford: Oxford University Press; 2009.

78. Joreskog KG. Structural equation modeling with ordinal variables using lisrel. 2005; http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf. Accessed 08/03/2012.

79. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal.* 1999;6(1):1-55.

80. Cole DA. Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology.* 1987;55(4):584.

81. Dziuban CD, Shirkey EC. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin.* 1974;81(6):358.

82. Brown TA. *Confirmatory factor analysis for applied research*. New York: The Guilford Press; 2006.

83. Thompson B. *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association; 2004.

84. Schmid J, Leiman JM. The development of hierarchical factor solutions. *Psychometrika.* 1957;22(1):53-61.

85. Wolff H-G, Preising K. Exploring item and higher order factor structure with the schmid-leiman solution: Syntax codes for spss and sas. *Behavior Research Methods.* 2005;37(1):48-58.

86. Lomax RG. *Statistical concepts: A second course*: Lawrence Erlbaum Associates; 2007.

87. Tabachnick BG, Fidell LS, Osterlind SJ. Using multivariate statistics. 2001.

88. Marsh HW, Balla JR, Mcdonald RP. Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin.* 1988;103(3):391.

89. Bernstein IH, Teng G. Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin.* 1989;105(3):467.

90. Dowie J, Elstein A. *Professional judgment: A reader in clinical decision making*: Cambridge University Press; 1988.

91. Gorsuch RL. *Factor analysis*. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1983.

92. Kreiter CD. Commentary: The response process validity of a script concordance test item. *Advances in Health Sciences Education.* 2012;17(1):7-9.

93. Linn RL. *Educational measurement*. 3 ed. New York: American Council on Education - Macmillan Publishing Company; 1989.

94. Messick S. Validity. In: Linn RL, ed. *Educational measurement*. 3 ed. New York: Macmillan; 1989:13-103.

95. Nabil NM, Guemei AA, Alsaaid HF, et al. Impact of introducing a diagnostic scheme to medical students in a problem based learning curriculum. *Medical Science Educator.* 2013;23(1):16-26.

96. Fabrigar LR, Wegener DT, Maccallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods.* 1999;4(3):272.

97. Kreiter CD, Bergus GR. Case specificity: Empirical phenomenon or measurement artifact? 2007.

98. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Medical Education.* 2006;40(7):618-623.

99. Rindskopf D, Rose T. Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research.* 1988;23(1):51-67.

100. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational measurement.* 1987;24(3):185-201.

CURRICULUM VITAE

Adam Benjamin Wilson

Education

Doctor of Philosophy, Anatomy and Cell Biology (Minor: *Education) 2013
Indiana University, Indianapolis, Indiana
Dissertation Title: A Psychometric Evaluation of Script Concordance Tests for Measuring Clinical Reasoning
*Education course work: intermediate educational statistics; multivariate analysis in educational research; learning and cognition in education (educational psych); college teaching and learning; instruction in the context of curriculum; action research; assessment in higher education; higher education administration; qualitative inquiry in education; research seminar in higher education

Master of Science, Anatomy and Cell Biology 2009
Rush University, Chicago, Illinois
Thesis Title: Anatomy Education: Analysis of Instructional Methodologies

Bachelor of Science, Biochemistry 2007
University of Illinois at Chicago, Chicago, Illinois

Academic Appointments

Institution	Title/Rank (appointment duration)	Description
Indiana University School of Medicine 635 Barnhill Drive Indianapolis, IN 46202-5120	Graduate Ph.D. Student (2010-2013)	Teaching assistantships: • Gross anatomy • Neuroanatomy • Histology
University of Medicine and Health Sciences – School of Medicine P.O. Box 1218 Basseterre, St. Kitts	Visiting Professor (7/19/2010 – 7/30/2010)	Lecture/laboratory topics: • Gross anatomy of the head and neck.
Robert Morris University College of Nursing and Health Studies 401 South State Street Chicago, IL 60605	Full-Time Faculty (2008-2010)	Teaching assignments in allied health: • Anatomy & Physiology • Pharmacy technology

RESEARCH

Research Presentations

Platform/Oral Presentations
March 21-23, 2013
Title: A closer look at clinical data interpretation fosters script theory ideology and offers practical implications
Event: Central Group on Educational Affairs Conference; Cincinnati, OH

Poster Presentations
October 8-9, 2012
Poster Title: Does student or faculty gender influence emergency medicine clerkship evaluations?
Event: American College of Emergency Physicians Research Forum; Denver, CO

March 2, 2012
Poster Title: Medical Students' Perceptions of Residents as Teachers: Residents are Equally Effective as Faculty in Simulation Debriefings
Event: IUPUI Edward C. Moore Symposium on Teaching Excellence

September 9, 2011
Poster Title: Residents as Teachers: Emergency Medicine Residents are as Effective as Faculty in Medical Student Simulation Debriefings
Event: Indiana University Department of Anatomy and Cell Biology Symposium

July 15-18, 2008
Poster Title: Bringing Anatomy to Life
Event: 25th Annual Meeting of the American Association of Clinical Anatomists; University of Toronto, Toronto, Ontario, Canada

Refereed Publications

1. Cooper DD, **Wilson AB**, Huffman GN, Humbert AJ. Medical students' perception of residents as teachers: Comparing effectiveness of residents and faculty during simulation debriefings. *Journal of Graduate Medical Education.* 2012;4(4):486-489.

2. **Wilson A**, Fegan F, Romence B, Uhe K, Dionne B. Precepting the medical assistant practicum: expectations and rewards: an evaluation of preceptors' opinions. *J. Allied Health.* 2011;40(4):212.

3. **Wilson AB**, Petty M, Williams JM, Thorp LE. An Investigation of Alternating Group Dissections in Medical Gross Anatomy. *Teaching and Learning in Medicine: An International Journal.* 2011;23(1):46 - 52.

4. **Wilson AB**, Ross C, Petty M, Williams JM, Thorp LE. Bridging the transfer gap: laboratory exercise combines clinical exposure and anatomy review. *Med. Educ.* Aug 2009;43(8):790-798.

TEACHING

Academic Presentations/Lectures/Talks

Medical Neuroanatomy Online Module
August 31, 2012 – An online module was designed and delivered to approximately 130 medical students. The aim of this module, covering the spinal cord, was to give students a means by which to review fundamental neuroanatomical concepts before being tasked to apply those concepts in a lecture devoted to teaching clinical neurology problems.

Graduate Human Gross Anatomy Lecture
April 17, 2012 – A 1.5 hour lecture on cranial nerves, the face, and parotid region was given to 28 graduate students. Following lecture, students participated in lab where they dissected the muscles of the face and parotid region. They also learned the foramina of the skull via an interactive station facilitated during the lab session.

Graduate Neuroanatomy Lecture/Lab
February 10, 2012 – A lecture and lab activity on the spinal cord were delivered to 12 graduate Neuroanatomy students over the course of 2 hours. The lecture was divided into 4 main topic sections. After each section, students completed an interactive PowerPoint based, think-pair-share exercise to reinforce the covered topics and formatively assess student learning and progress.

Team Based Learning (TBL) Module
October 5, 2010 – A TBL module consisting of an individual readiness assurance test (iRAT), a group readiness assurance test (gRAT), an application exercise, and a debriefing period was delivered to 35 histology graduate students. The module covered cartilage, bone, and the cardiovascular system.

Teaching Assignments

Indiana University School of Medicine

Course Title	Course Description and *Teaching Responsibilities	Year
Gross Anatomy	Intensive introduction to the gross anatomy of the human body, including a complete dissection. Series of lectures on radiographic anatomy and clinical application of anatomy. Frequent conferences and discussions with members of staff. *For 15 weeks I served as a TA in gross anatomy for medical students. Over the course of the semester I designed and implemented online learning modules, gave one course lecture, prosected cadaveric specimens, assisted students in structure identification in every laboratory, and tutored small groups outside of the assigned class time.	2012
Functionally-Oriented Human Gross Anatomy	Introduction to the concepts, terminology, and basic structure of the human body. Prosection of the body will use a regional approach. Emphasis on providing fundamental knowledge of the structure/function of major organ systems, peripheral nervous system, and vascular supply to the trunk, head and neck, limbs, and back	2012

	*For 15 weeks I served as a TA in anatomy for graduate students. Over the course of the semester I gave one course lecture and assisted students in structure identification during laboratory sessions.	
Neuroscience and Clinical Neurology	A multidisciplinary consideration of structural, functional, and clinical features of the human nervous system. *As a TA I instructed in the neuroanatomy wet labs and constructed and implemented an online spinal cord module for medical students to work through and learn from.	2012
Graduate Neuroanatomy	Introduction to terminology, pathways, organization, and concepts of the human nervous system. Emphasis on providing fundamental knowledge of the structure, neurochemistry, and molecular mechanisms of the central and peripheral nervous systems in health and disease. *As a TA I instructed graduate students in the neuroanatomy wet labs and delivered a 2-hour lecture/lab on the spinal cord.	2012
Basic Histology	Lecture and laboratory instruction on the microscopic structure of the basic tissues and organs of the body. Previous exposure to gross anatomy principles and dissection encouraged. *For 15 weeks I served as a graduate TA in basic histology for graduate students. Over the course of the semester I designed and implemented one Team Based Learning (TBL) module, assisted students in structure identification in every laboratory, tutored small groups outside of the assigned class time, and lead two exam review sessions at the request of the students.	2010

University of Medicine and Health Sciences

Course Title	Course Description and *Teaching Responsibilities	Year
Anatomy	Anatomy focuses on the gross structure of organs and function and, through clinical correlations, relates each to clinical medicine. An Anatomical Learning Resource Center has been established to utilize computer-based instruction, anatomical models, radiographic materials as well as supervised laboratory sessions dissecting various parts of the human body. Students study the structure and function of all organs with some interaction with cellular structure. When this course is complete, each student will have extensive knowledge of the gross anatomy of the entire human body as it relates clinically to the practice of medicine. *During two weeks as a visiting professor, I presented four didactic lectures, lead 1 review session, and guided first year medical students in their assigned laboratory dissections. The content covered focused on head and neck anatomy.	2010

Robert Morris University

Course Title	Description of Courses Taught	Year
Human Anatomy and Physiology I for Nurses	This course covers the fundamental principles of the structure, function, and organization of the human body through the study of histology, including cells and tissues, and through the study of major body systems, such as the 146ntegumentary and skeletal systems.	2009
Human Anatomy and Physiology II for Nurses	This course covers the fundamental principles of the structure, function, and organization of the human body through the study of several major body systems, including the muscular, cardiovascular, respiratory, digestive, and urinary systems.	2010
Human Anatomy and Physiology III for Nurses	This course covers the fundamental principles of the structure, function, and organization of the human body through the study of major body systems including the lymphatic, reproductive, endocrine, nervous and special senses.	2010
Human Body Systems I	This course focuses on the fundamental principles of the structure, function, and organization of the human body through the study of word parts; body positions, planes and directions; cells, tissues and membranes; and major body systems including skeletal and muscular. Medical terminology and pathology for systems is covered. The components of human movement are addressed. Critical thinking based on the academic subject matter is developed and enables the incorporation of cognitive knowledge in the performance of psychomotor and affective domains. This course includes a laboratory component.	2009
Human Body Systems III	This course focuses on the fundamental principles of the structure, function, and organization of the human body through the study of several major body systems including the lymphatic and immune, cardiovascular, and respiratory systems. Medical terminology and pathology for systems is included. Critical thinking based on the academic subject matter is developed and enables the incorporation of cognitive knowledge in the performance of Allied Health Studies professions. This course includes a laboratory component.	2009
Anatomy and Physiology I	This course covers the fundamental principles of the structure, function, and organization of the human body through the study of several major body systems including the respiratory, reproductive, circulatory, cardiovascular, hematic, and endocrine systems.	2008
Anatomy and Physiology II	This course focuses on the fundamental principles of the structure, function, and organization of the human body through the study of several major body systems including renal, digestive, muscular, skeletal, 146ntegumentary, lymphatic, special senses, and nervous systems.	2008

SERVICE & MISC.

Professional Organizations

International Association of Medical Science Educators	2011-Present
American Association of Clinical Anatomists	2011-Present
American Association of Anatomists	2010-Present
National Council on Measurement in Education	2013-Present
Association for Surgical Education	2013-Present

Consulting Experience

September 2011-March 2012

- Volunteer educational consultant for Indiana University's competency planning team charged with redesigning the "problem solving" competency as part of a medical school wide curriculum reform effort.

Conferences Attended as a Registrant

April 23-27, 2013 - Surgical Education Week, Gaylord Palms Hotel, Orland, FL.

May 9, 2012 - Society for Academic Emergency Medicine, Annual Meeting, Sheraton Chicago Hotel & Towers, Chicago, IL.

February 25, 2012 - American Association of Anatomists, Regional Meeting, Rush University Medical Center, Chicago, IL.