"DO I *REALLY* HAVE TO COMPLETE ANOTHER EVALUATION?"

EXPLORING RELATIONSHIPS AMONG PHYSICIANS' EVALUATIVE LOAD, EVALUATIVE STRAIN, AND THE QUALITY OF CLINICAL CLERKSHIP EVALUATIONS

Courtney Jo Traser

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 2017

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

_____

James J. Brokaw, Ph.D., M.P.H., Chair

Doctoral Committee

_____

Adam B. Wilson, Ph.D.

_____

Gary R. Pike, Ph.D.

_____

Ronald L. Shew, Ph.D.

April 14, 2017

_____

Melissa R. Alexander, Ed.D.

DEDICATION

I dedicate this work to the four most cherished people in my life: my parents, Erin and Shane; my brother, Dallas; and my fiancé, Nicholas James.

To my mom, my tiny gran: Words cannot begin to express my gratitude for your enduring love, encouragement, and friendship. With a celebration for each milestone and a soothing word for each hurdle, you've been my constant comfort on this journey. The realization of this dream is as much your accomplishment as mine, as it would not have been attainable without you. You are the most influential, loving, and wonderful woman I know, and I am so grateful to call you, among many things, 'Mom.'

To my dad, my pepaw: Thank you for your guiding wisdom, your German tenacity, and your unwavering, unconditional love and support. You taught me the meaning of determination and perseverance, and showed me that every problem can be solved with a conscious effort and a good attitude; this work (and our arrival at the Olympic Winter Park in Oslo!) is a true testament to that. I hope that I've made you proud.

To my brother, my piddis: You are the personification of joy. Your unmatched optimism and your zest for life encourage me to look at life through a positive lens, and have made me smile on even my lowest of days. Thank you for always being my own personal little sunshine; I admire you more than you could ever know.

Finally, to my (soon to be) husband, Nicholas, the love of my life: Your arrival into my life has brought me indescribable happiness. Your never-ending patience, your constant reassurance, and your enduring faith in my abilities have made this dream a reality. I am incredibly grateful to have held your hand on this arduous journey, and I am so excited to see where our dreams will take us next. Together, there is nothing we cannot accomplish. I love you, always.

ACKNOWLEDGEMENTS

First, and foremost, I would like to thank my *de facto* chair, Dr. Adam Wilson, for his mentorship, encouragement, and reassurance throughout this journey. Among many things, his guidance, insight, and (perhaps, most importantly) his patience were invaluable to this work. I could not have asked for a finer mentor to help me achieve this dream and prepare me for the trials of academia. Thank you, Adam.

I would also like to thank my committee members, Dr. Jim Brokaw, Dr. Gary Pike, Dr. Ron Shew, and Dr. Melissa Alexander. The advice and unique experience (and in Dr. Shew's case, humor) offered by each of them has helped me grow as a researcher and an educator. I am honored to have worked with such talented and sagacious individuals.

Finally, I want, and need, to thank my dear friend, Jessica Byram. A true wonder woman, her friendship, company, and support throughout these four long years has made this journey enjoyable on even the most trying of weeks. I am deeply indebted to her, not only for her indispensable help with this work, but also for her endearing camaraderie. Thank you, thank you, thank you!

Courtney Jo Traser

"DO I *REALLY* HAVE TO COMPLETE ANOTHER EVALUATION?"

EXPLORING RELATIONSHIPS AMONG PHYSICIANS' EVALUATIVE LOAD, EVALUATIVE STRAIN, AND THE QUALITY OF CLINICAL CLERKSHIP EVALUATIONS

Background. Despite widespread criticism of physician-performed evaluations of medical students' clinical skills, clinical clerkship evaluations (CCEs) remain the foremost means by which to assess trainees' clinical prowess. Efforts undertaken to improve the quality of feedback students receive have ostensibly led to higher assessment demands on physician faculty; the consequences of which remain unknown. Accordingly, this study investigated the extent to which physicians' evaluative responsibilities influenced the quality of CCEs and qualitatively explored physicians' perceptions of these evaluations.

Methods. A questionnaire was delivered to physicians (n = 93) at Indiana University School of Medicine to gauge their perceived evaluative responsibilities. Evaluation records of each participant were obtained and were used to calculate one's measurable quantity of CCEs, the timeliness of CCE submissions, and the quality of the Likert-scale and written feedback data included in each evaluation. A path analysis estimated the extent to which one's evaluative responsibilities affected the timeliness of CCE submissions and CCE quality. Semi-structured interviews with a subset of participants (n = 8) gathered perceptions of the evaluations and the evaluative process.

Results. One's measurable quantity of evaluations did not influence one's perceptions of the evaluative task, but did directly influence the quality of the Likert-scale items. Moreover, one's perceptions of the evaluative task directly influenced the timeliness of CCE submissions and indirectly influenced the quality of the closed-ended CCE items. Tardiness in the submission of CCEs had a positive effect on the amount of score differentiation among the Likert-scale data. Neither evaluative responsibilities nor the

timeliness of CCE submissions influenced the quality of written feedback. Qualitative analysis revealed mixed opinions on the utility of CCEs and highlighted the temporal burden and practical limitations of completing CCEs.

Conclusions. These findings suggest physicians' perceptions of CCEs are independent of their assigned evaluative quantity, yet influence both the timeliness of evaluation submissions and evaluative quality. Further elucidation of the mechanisms underlying the positive influence of evaluation quantity and timely CCE submissions on CCE quality are needed to fully rationalize these findings and improve the evaluative process. Continued research is needed to pinpoint which factors influence the quality of written feedback.

James J. Brokaw, Ph.D., M.P.H., Chair

TABLE OF CONTENTS

List of Tables x

List of Figures xi

List of Abbreviations xii

Chapter One: Introduction 1

Introduction 2

Purpose of the Study 6

Presentation of the Research Questions 7

Overview of the Dissertation 7

Chapter Two: Review of the Literature 8

Introduction 9

A Brief History on the Evolution of Clinical Education of Medical Students 10

Physician-Performed Evaluations of Medical Students' Clinical Conduct 13

Overburdening Physicians with Evaluative Tasks 14

Concerns Regarding the 'Quality' of Clinical Clerkship Evaluations 15

Chapter Three: Methodology 17

Introduction 18

Rationale for a Mixed-Methods Approach 19

Overview of the Research Design 19

Phase I: Collection and Analysis of Quantitative Data 22

Phase II: Collection and Analysis of Qualitative Data 54

Ethical Considerations 56

Chapter Four: Quantitative and Qualitative Results 57

Introduction 58

Part I: Quantitative Results 59

Part II: Qualitative Results 79

Chapter Summary 99

Chapter Five: Discussion and Conclusions 101

Introduction 102

Integration and Interpretation of the Findings 104

Limitations 117

Suggestions for Future Research 121

Conclusions 124

Appendices 126

Appendix A: Evaluation of Student Clerkship Performance ('Medium' Form) 127

Appendix B: Evaluation of Student Clerkship Performance ('Long' Form) 130

Appendix C: The Electronic Study Survey 136

Appendix D: Quality of Clinical Clerkship Evaluation (CCE) Rubric 139

Appendix E: Interview Protocol 140

Appendix F: Histogram Illustrating Participants' Scores Across the Closed-ended CCE Items 141

References 142

Curriculum Vitae

LIST OF TABLES

Table 3-1. Description of study timeline and progression 21

Table 3-2. Summary of study variables and data collection methods 25

Table 3-3. Phraseology differences between the SURG- and EVAL-TLX 28

Table 3-4. Example calculation of one's 'aggregated EVAL-TLX score' 30

Table 3-5. Example CCE records for a single physician 32

Table 3-6. Example calculation of a single open-ended comment using the 'Quality of CCE Rubric' 38

Table 3-7. Summary of the direct and indirect effects of each variable 45

Table 3-8. Dummy-coded reference groups for the demographic variables 46

Table 4-1. Response rate by medical department 59

Table 4-2. Description of phase I participants 60

Table 4-3. Descriptive statistics of the five evaluative variables 63

Table 4-4. Correlation matrix of the exogenous and endogenous variables 66

Table 4-5. Estimated variance of the residuals 67

Table 4-6. Standardized effects of the demographic variables on the evaluative variables 69

Table 4-7. Standardized effects of the evaluative variables on other evaluative variables 72

Table 4-8. Description of phase II participants 80

Table 4-9. Themes, descriptions, and representative quotes for 'Evaluative Utility' 85

Table 4-10. Themes, descriptions, and representative quotes for 'Evaluative Quality' 91

Table 4-11. Themes, descriptions, and representative quotes for 'Evaluative Cost' 95

Table 4-12. Themes, descriptions, and representative quotes for 'Evaluative Practicability' 98

Table D-1. Quality of CCE Rubric 139

LIST OF FIGURES

Figure 3-1. Example of a path model 41

Figure 3-2. The conceptual model for the study 44

Figure 3-3. Analysis for Research Question 1 48

Figure 3-4. Analysis for Research Question 2 49

Figure 3-5. Analysis for Research Question 3 50

Figure 3-6. Analysis for Research Question 4 51

Figure 3-7. Analysis for Research Question 5A 52

Figure 3-8. Analysis for Research Question 5B 53

Figure 4-1. Results for Research Question 1 73

Figure 4-2. Results for Research Question 2 74

Figure 4-3. Results for Research Question 3 75

Figure 4-4. Results for Research Question 4 76

Figure 4-5. Results for Research Question 5A 77

Figure 4-6. Results for Research Question 5B 78

Figure A-1. Evaluation of student clerkship performance ('medium' form) 127

Figure B-1. Evaluation of student clerkship performance ('long' form) 130

Figure C-1. Electronic study survey 136

Figure F-1. Histogram illustrating participants' scores across the closed-ended items 141

LIST OF ABBREVIATIONS

AAMC	Association of American Medical Colleges
ABS	American Board of Surgery
AMA	American Medical Association
CCE(s)	Clinical clerkship evaluation(s)
DALI	Driving Activity Load Index
Delay	Time delay
Department	Medical department
Diff	Score differentiation
ETE	Evaluation-to-Evaluative Variance
EVAL-TLX	Evaluative Task Load Index
FM	(Department of) Family Medicine
GPEP	General Professional Education of Physicians and College Preparation for Medicine Panel
GS	(Department of) General Surgery
ICC	Intra-class correlation
IM	(Department of) Internal Medicine
IRB	Institutional Review Board
IRR	Inter-rater reliability
ITI	Item-to-Item Variance
IUSM	Indiana University School of Medicine
Ln	Logarithmic transformation
Load	Evaluative load
MSE	Medical Student Education
MSPEs	Medical Student Performance Evaluations
NASA-TLX	National Aeronautics and Space Administration Task Load Index
NBME	National Board of Medical Examiners
OB/GYN	(Department of) Obstetrics and Gynecology
OPR(s)	Operative Performance Rating(s)
PE	(Department of) Pediatrics
PY	(Department of) Psychiatry
QOF	Quality of Feedback
Rank	Academic rank
REDCap	Research Electronic Data Capture
RVUs	Relative value units
SD	Standard deviation
SE	Standard error
Strain	Evaluative strain
SURG-TLX	Surgery Task Load Index
UME	Undergraduate medical education

CHAPTER ONE

Introduction

Introduction

Purpose of the Study

Presentation of the Research Questions

Overview of the Dissertation

INTRODUCTION

Within the last quarter-century, the Association of American Medical Colleges (AAMC) and the American Medical Association (AMA) have ardently advocated for a national reform of the clinical curriculum of both undergraduate and graduate medical education.[1] In particular, the lack of nationally defined standards for clinical skills instruction coupled with inconsistent and poorly defined assessment measures has resulted in a strong criticism of physician-performed evaluations of medical students' and residents' clinical skills.[1-3] As these clinical evaluations remain the foremost means by which to assess one's clinical prowess in medical academia,[4,5] efforts have been undertaken to revise the evaluative process and enhance the quality of these subjective evaluations. For instance, the American Board of Surgery (ABS) recently imposed a requirement that each general surgery resident be directly evaluated a minimum of 12 times during their residency training, with six evaluations each in the areas of clinical aptitude and operative performance.[6] Because this new requisite requires a specific number of physician-performed observations and judgments of trainees' performances, the implementation and realization of such evaluative protocols may place greater demands on evaluators, perhaps even more than may be expected.

The consequences of implementing higher assessment demands on physician faculty, such as those imposed by the ABS, have not been well documented in the medical literature. While the literature is rich with articles detailing students' experiences during clerkship rotations,[7-11] criticizing the evaluative process,[12-19] or demonstrating the significance of these evaluations for medical students,[20-23] relatively few articles have assessed the evaluative process from the *physician* perspective.[24,25] Moreover, physician evaluators (or 'raters') have only recently become an object of study for researchers,[26] though the focus of published literature consists primarily of proposed theoretical

constructs and frameworks raters use to inform their judgments of trainees;[27,28] or raters' categorization of trainees based on subjective interpretations,[29] rather than physicians' perceptions of the evaluative process. The demands for increased evaluation of trainees in clinical settings[30] and the development of evaluation management systems designed to simplify the creation and distribution of evaluations[31] have clearly prompted an increase in workplace-based assessments.[27] In the absence of sufficient literature describing physicians' evaluative responsibilities, it is challenging to fully appreciate how physicians' current assessment demands compare with demands of the past few decades. If the medical education community is to continue the implementation of more stringent guidelines and robust mandates, such as those established by the ABS, it would be beneficial to understand how the extent of physicians' evaluative responsibilities influences one's perceptions of the task and alters the quality of the clinical evaluations one completes.

For the purposes of this research, one's actual or imposed evaluative responsibilities, defined as the *measurable quantity* of evaluations, ratings, or surveys completed by a physician during a specific period-of-time, will be referred to as one's '**evaluative load**.' While this research will specifically use clinical clerkship evaluations (CCEs) as a surrogate for measuring physicians' evaluative load, any additional surveys or questionnaires issued by a physician's medical department (e.g., evaluations of residents), workplace, or professional societies/organizations should also be considered as factors that may contribute to one's evaluative load. Similarly, one's perceived evaluative responsibilities will be referred to as one's '**evaluative strain**.' Defined as a physician's *perceptions* of the evaluative task, evaluative strain includes one's perceptions of his or her evaluative load and one's conceptualizations of the cognitive demands needed to complete the evaluative task. This construct does not attempt to quantify the extent to which actual cognitive processes are utilized during the completion of an evaluation; such work has been

extensively investigated elsewhere and is thoroughly reviewed by Gauthier et al.[32] Among other things, physicians' evaluative strain is likely to fluctuate as a function of the perceived mental activity involved, the level of frustration or effort a task elicits, or the time pressures associated with a task.[33,34]

Due to the existing gaps in the medical education literature, the notion of causal relationships among physician evaluative load, evaluative strain, and the quality of clinical evaluations remains largely speculative. However, there is tangential evidence suggesting that increasing physicians' assessment demands may inadvertently lessen the quality of clinical evaluations. For example, Tavares and colleagues[26] investigated how increasing the number of dimension-specific behaviors raters were asked to identify altered the quality of evaluations and influenced raters' mental perceptions of the task. Using generalizability theory, the authors demonstrated that raters asked to identify behaviors related to only two dimensions had noticeably higher inter-rater reliabilities (G = 0.56 for an average of ten raters) than raters asked to identify behaviors related to seven dimensions (G = 0.42, respectively). Moreover, increasing the quantity of items to be assessed notably heightened participants' perceived mental burden.[26] For instance, participants reported feeling overwhelmed by the number of behaviors they were asked to evaluate while trying to differentiate between pertinent and irrelevant audio/visual stimuli. This sense of feeling mentally burdened prompted several participants to employ 'load avoidance strategies' or coping mechanisms to the detriment of evaluative quality, as some raters reported evaluating only "negative behaviors," or those they believed to be the "easiest to identify," rather than evaluating behaviors that were more representative of a trainees' clinical knowledge (e.g., equally noting both positive and negative behaviors).[26] These findings coincide with reports from several studies that note how increases in either imposed or perceived rater requirements may produce rating errors that affect the quality of clinical

evaluations.[35,36] Ultimately, these findings suggest that increasing one's measurable evaluative responsibilities may influence one's perceived cognitive demands. This also supports the proposed notion that one's evaluative load and strain, if high enough, may negatively affect the quality of clinical evaluations.

Further evidence of the causal relationships between physicians' evaluative responsibilities and the quality of clinical evaluations is provided by Williams et al.[37] In a retrospective study, Williams and colleagues investigated the extent to which evaluation quality was impacted by time delays between the observation of a resident's operative performance and the rating of that performance. Their results demonstrated that the duration of the time delay notably affected the quality of the evaluation in both the amount of "item-to-item variation" (i.e., the degree of variability present among the Likert-scale items, such that no item-to-item variability would be represented by all scaled items receiving the same rating) and the specificity of written comments. The gravity of their findings led the authors to suggest that ratings should be recorded as soon as possible, as ratings given more than three days after the initial observation are not reliable representations of trainees' clinical performances.[37] While the results of their study were significant and pose direct implications for clinical evaluations, Williams and colleagues[37] did not speculate as to *why* the majority of evaluations were completed four to fourteen days post-observation. It is reasonable to hypothesize that such time delays may be in part related to a physician's evaluative load and/or evaluative strain. If clinical evaluations are to continue to be a source of comparative information for student performance and if steps are to be taken to improve the time-to-completion of evaluations, it would serve the research community well to better understand the prevalence of evaluative load and strain among physicians and how these factors affect the quality of clinical evaluations through the mediating effects of time delays.[5,37]

PURPOSE OF THE STUDY

While several authors have provided evidence supporting theorized relationships among physicians' measurable and perceived evaluative responsibilities and the quality of CCEs,[26,35-37] a thorough exploration of such associations has yet to be performed. As such, the twofold focus of this work was (1) to investigate the extent to which the quality of CCEs was directly influenced by a physician's evaluative load and evaluative strain, and indirectly mediated by time delays occurring between a clinician's final encounter (or observation) with a medical student and the rating of that medical student's clinical performance; and (2) to understand how physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students.

This study is unique in that it is the first to investigate how physician evaluative load and strain influence the quality of CCEs. As such, it is anticipated that this research may contribute to the literature in several ways. Principally, developing an understanding of how physician evaluative load influences evaluative strain may encourage clerkship directors and assessment administrators to re-examine and modify the clinical evaluation process. A second goal of this study is to determine if and how one's evaluative responsibilities affect the quality of physician-performed evaluations. This may help to rationalize the occurrence of time delays between an observation of a student's clinical performance and the rating of that performance. Finally, heightening our understanding of how imposed and perceived evaluative tasks can affect the quality of evaluations may lead to the operationalization of more sophisticated evaluation monitoring systems designed to recognize the point at which a physician has become overburdened by evaluative responsibilities.

PRESENTATION OF THE RESEARCH QUESTIONS

The six research questions underpinning this study include:

Research Question 1 How does physician evaluative load influence evaluative strain?

Research Question 2 How do evaluative load and evaluative strain directly affect the quality of CCEs, as measured by the degree of score differentiation and the quality of written feedback?

Research Question 3 To what degree are evaluative load and evaluative strain associated with the length of time delay between a clinician's final observation or encounter with a clerk and the rating of that clerk's clinical performance?

Research Question 4 To what extent do time delays directly influence the quality of CCEs?

Research Question 5 To what extent do time delays mediate the influences of (5A) evaluative load and (5B) evaluative strain on the quality of CCEs?

Research Question 6 How do physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students?

OVERVIEW OF THE DISSERTATION

The ensuing chapters of this work will include a review of the relevant literature (Chapter Two), a description of the methods employed to answer the six research questions (Chapter Three), a reporting of the study's quantitative and qualitative findings (Chapter Four), and a discussion of the study's results, implications, and limitations, as well as suggestions for future research (Chapter Five).

CHAPTER TWO

Review of the Literature

Introduction

A Brief History on the Evolution of Clinical Education of Medical Students

Physician-Performed Evaluations of Medical Students' Clinical Conduct

Overburdening Physicians with Evaluative Tasks

Concerns Regarding the 'Quality' of Clinical Clerkship Evaluations

INTRODUCTION

This chapter provides the scholarly foundation necessary for understanding physician evaluative load and strain as concepts that may directly and/or indirectly relate to the quality of physician-performed evaluations of third-year medical students' clinical aptitude. This chapter begins with a brief history of the evolution of clinical education within medical institutions; starting with the introduction of the first clinical curricular model in the late 1800s and continuing until present day. Transitioning from the clinical model itself to the role of the physician in clinical education, the second section describes the clinical evaluation process in terms of physicians' evaluative responsibilities. As little research has examined physicians' perspectives of the evaluative process, it is difficult to characterize the extent to which physicians feel overwhelmed by their imposed or perceived evaluative responsibilities, or how this may affect the quality of the evaluations they complete. Consequently, the third section of this chapter will broadly introduce survey fatigue and response burden as impetuses for rationalizing how physicians' evaluative burdens may be contributing to a reduction in the quality of clinical evaluations. Finally, the concerns surrounding these subjective evaluations will be presented and addressed in the context of the significance that evaluations hold for medical students.

A BRIEF HISTORY ON THE EVOLUTION OF CLINICAL EDUCATION OF MEDICAL STUDENTS

Although formal clinical training is currently regarded as an integral component of medical students' undergraduate education, this has not always been the case. Prior to the 1900s, formal clinical instruction (as we currently know it) was absent from nearly all medical institutions within the United States. At this time, trainees were expected to acquire an understanding of clinical medicine by serving as apprentices to community-based physicians or by entering independent practice. The notion of including formal clinical instruction in medical curricula did not occur until Sir William Osler implemented a new curricular model at Johns Hopkins Hospital in the late 1800s. Under his guise, third-year medical students began to attend clinical demonstrations within the hospital clinic while fourth year medical students were held responsible for the care of patients in rotating medical disciplines. While the first two years of students' medical education consisted primarily of coursework deemed necessary to develop a foundation in medical science, the final two years became known as the "clinical clerkship" years, in which students were required to complete rotations in internal medicine, obstetrics and gynecology, pediatrics, psychiatry, and surgery.[1] Thanks in part to The Carnegie Foundation's publication of "The Flexner Report" in 1910,[38] Osler's curricular model came to be implemented in nearly all U.S. medical institutions during the first half of the twentieth century.

These "Oslerian objectives"[39] remained constant in medical institutions until the 1950s. During this time, a restructuring of the non-clinical years took place, in which some institutions moved away from a discipline-specific block structure in favor of systems-based instruction. The clinical years were also reorganized, such that the core of the clinical clerkship emphasis moved from the fourth year to the third year, leaving the fourth year open for students to take electives in courses of interest to them.[1] During these clinical

years, medical students ('clerks') were typically assigned to teams of residents and attending physicians and were expected to learn through observation and interactions with their clerkship team. This curricular design created a large amount of variability in students' formal clinical instruction, as one student's clerkship experience may have been vastly different from another's, even within the same medical discipline.[1,40] According to Nutter and Whitcomb:[1]

> The variability was inevitable, because of the varied nature of the clinical sites to which the students were assigned over the course of any given year, the variable spectrum of the conditions encountered at those sites, and the variable quality of the supervision and teaching provided by resident physicians and attending physicians at those sites.

This variability prompted a number of organizations, such as the Association of American Medical Colleges' (AAMC) General Professional Education of Physicians and College Preparation for Medicine Panel (collectively referred to as the GPEP Panel), to investigate the quality of the clinical clerkship experience. In their 1984 report, "Physicians for the Twenty-First Century," the authors noted that clinical clerkships did a poor job of delineating clear learning objectives for students; leaving students unable to adequately meet their preceptors' expectations or allowing for an accurate evaluation of students' clinical accomplishments.[40] These sentiments were echoed in a seminar presented by the Macy Foundation in 1988 that described the clinical education of medical students as outdated and inappropriate for modern times.[41] A report published by the AAMC in 1998 noted that only a few medical institutions had begun to modify their clinical curricula to reflect these concerns,[42] though a more widespread change began to occur towards the end of the decade.[1] The prevalence of curriculum reform grants offered by the Robert Wood Johnson Foundation and the Bureau of Health Professions encouraged several medical institutions to restructure their medical curricula in accordance with the suggestions put forth by the GPEP Panel. However, further inquiry into the restructuring of medical

curricula in the late 1990s found that most institutions were only making changes to the non-clinical years, claiming that a restructuring of the clinical curricula was too difficult to achieve.[1]

The early 2000s witnessed little change in curricular models employed at most U.S. allopathic institutions. The majority of medical institutions continued to adhere to the two-phase model of undergraduate medical education (UME) that had been utilized for the past 100 years. Within the first phase (still referred to as the pre-clinical/pre-clerkship years), students were didactically taught the foundational principles of medicine (e.g., human gross anatomy, human physiology, etc.), though many curricula had expanded to reflect advances made in the fields of genetics and biochemistry. Phase two continued to constitute the greater part of the clinical curriculum, or the clinical-years. Within the last decade, however, a greater number of U.S. medical schools have begun to shift away from this bi-phasic structure, favoring instead a more seamless integration of foundational and clinical content across all four years of the UME curriculum.[1] Clinical content has begun to appear in the foundational courses, with educators frequently incorporating 'clinical correlates' and medical images into their lectures. Moreover, courses focusing on doctor-patient relationships, the development of adequate communication skills, and instruction on obtaining patient histories are now commonly required during the non-clinical years. As of late, students are more frequently exposed to various clinical settings early within their first year of medical training.[1]

Despite these modern curricular changes, the preponderance of clinical curricula remain in the final two years of medical school at most U.S. institutions. Although the specific structural organization of these years varies by institution,[43] students are commonly required to complete a series of clinical clerkship rotations in a core group of medical specialties similar to that of years past, including family medicine, internal

medicine, neurology, obstetrics and gynecology, pediatrics, psychiatry, and general surgery (with some schools also requiring rotations in ambulatory care and emergency medicine).[1,44,45] Designed to provide students with an opportunity to clinically apply the foundational science and medical knowledge obtained during their pre-clinical years and to improve their interpersonal skills with patients and other healthcare professionals, these clinical clerkship rotations typically occupy the entirety of the students' third year and occur at an assortment of clinical venues/sites. Although there continues to be a great deal of variability in clinical rotations, the role of physician preceptors and the need to evaluate clerks' clinical skills are considered more stable aspects of the clerkship experience.

PHYSICIAN-PERFORMED EVALUATIONS OF MEDICAL STUDENTS' CLINICAL CONDUCT

During each clerkship rotation, medical students are often assigned to either a faculty or resident preceptor who is responsible for mentoring students and observing their behaviors within the clinical setting. Although clerks are customarily assigned to a specific preceptor during each rotation, it is not uncommon for one preceptor to be simultaneously accountable for multiple medical students. At the culmination of each clerkship rotation (or shift, in some instances), these clinical-educators are required to evaluate the performance of their assigned medical student(s) based on overall demeanor, clinical reasoning skills, foundational knowledge, technical skills, level of care, and depth of patient interactions via a clinical clerkship evaluation form (or CCE).[20,24] The length and specificity of each medical department's CCE can differ by specialty or medical institution; however, most forms routinely include a checklist of clinical performance markers and personality characteristics for faculty to use to evaluative students' clinical aptitude using numeric scales and additionally provide several open-ended questions so that faculty can justify their overall ratings of student performance or add additional comments.[24] As each medical student requires at least one clinical evaluation per rotation, this evaluative responsibility may

quickly become unmanageable for faculty preceptors and may ultimately affect the quality of the evaluations;[34] particularly if the culture of some departments dictates that a sole physician assumes responsibility for the completion of all clerkship evaluations per rotation.

OVERBURDENING PHYSICIANS WITH EVALUATIVE TASKS

The frequency with which CCEs need to be completed may pose challenges for faculty preceptors. Though little to no research has investigated the extent to which physicians feel overwhelmed by their imposed or perceived evaluative responsibilities, or how this may affect the quality of the evaluations they complete, related research on 'respondent burden' or 'survey fatigue' may provide insights into this aspect of the profession. Respondent burden is typically defined as "the time required for the completion of various forms," including surveys, or questionnaires.[46] Survey fatigue is a constituent of respondent burden that refers to not only the temporal demands needed to complete a survey, but also the mental effort or exhaustion that accompanies the completion of such forms.[47] Although studies investigating survey fatigue are limited, the findings seem to be relatively consistent. For example, it has been demonstrated that the completion of surveys back-to-back decreases one's response rate, such that participants are less likely to complete a second survey if they have recently completed another.[47,48] Moreover, it has been reported that the number of surveys participants are willing to complete decreases as the time it takes to complete a survey increases.[49] Furthermore, the content of the survey and its relevance to participants is believed to affect one's willingness to complete the form. For example, students at the United States Air Force Academy reported feeling overburdened by their institution's survey requests as they deemed the content to be unrelated or immaterial to their student roles.[50] Finally, survey fatigue has been linked to

'satisficing tendencies,' or the selection of a response deemed 'good enough' in order to expedite the survey or evaluative process; typically, at the cost of survey quality.[51,52]

While such implications of survey fatigue are directly related to survey research, it can be argued that these conclusions may be applicable to physician-performed evaluations of clerks, as well. As some physicians may be expected to complete multiple CCEs per shift or rotation, it is possible that the back-to-back completion of these evaluations may decrease their willingness to thoroughly complete the next. Moreover, physicians have reported the teaching of medical students to be a "low priority task," as department chairs and deans rarely and/or minimally reward faculty for excellence in education.[1] Consequently, the salience of CCEs may be of little personal interest to physician faculty who feel obliged to complete the forms only because it is a requisite of their faculty position. Ultimately, these notions suggest that CCEs may not be of top priority, thereby affecting their quality.

CONCERNS REGARDING THE 'QUALITY' OF CLINICAL CLERKSHIP EVALUATIONS

The quality of CCEs has long been debated in the medical education literature. Despite claims that CCEs "represent a reliable and valid method of evaluating both cognitive and non-cognitive aspects" of clinical performance, the subjective nature of clinical evaluations has prompted some scholars to scrutinize these evaluations.[12-19] One such criticism is the time that physician faculty spend in direct observation of their medical trainees. Although the majority of U.S. medical institutions use direct observation of trainees to assess clinical competency,[53] the observation of clerks by physician faculty is generally considered scarce, inconsistent, and insufficient for physicians to accurately evaluate clinical knowledge.[4,54-56] Several studies have reported that direct observation of bedside behaviors, such as interviewing a patient or conducting a physical evaluation, improves one's clinical skills.[57,58] However, other studies have reported that trainees were

never observed performing such tasks during their clinical experiences.[55,59,60] Findings from a multi-year study conducted at the University of Virginia School of Medicine in 2004 reported that more than half of their participating clerk population (n = 344) had never been observed by a faculty member during a patient interview or a physical examination in one of their six required rotations.[55] These results parallel those of earlier studies conducted by both the AAMC and the National Board of Medical Examiners (NBME), whose findings also indicated a low incidence of direct observations by physician faculty in the clinical setting.[59,60] This infrequency of direct observation by physicians has also been reported in graduate medical education, with one study reporting that a significant proportion of emergency medicine residents had never been observed performing basic bedside skills, conducting a physical examination, or taking a patient's history. [56,61]

Such concerns regarding the quality of clinical evaluations necessitate that each clinical evaluation provide an honest and accurate summary of trainees' clinical skills if trainees are to benefit from these clinical evaluations. Moreover, this sentiment is perhaps most pressing in regards to preceptors' written comments on the clinical evaluation forms, as they can be used to "differentiate 'A'-level performance from 'B'-level performance"[24] and pose implications for residency matching, as their contributions to clerkship rotation grades are reflected within Medical Student Performance Evaluations (i.e., Dean's letters).[21] Preceptors' written comments regarding students' knowledge base have also been used to predict low performance on NBME subject examinations[23] and highlight student lapses in professionalism.[22] Furthermore, CCEs are often the sole means of providing clerks with formal feedback regarding their clerkship experience.[20] As CCEs are of great significance to medical students, this research project serves to improve our understanding of how the quality of these evaluations may be affected by physicians' evaluative load and/or strain.

CHAPTER THREE

Methodology

Introduction

Rationale for a Mixed-Methods Approach

Overview of the Research Design

Phase I: Collection and Analysis of Quantitative Data

Phase II: Collection and Analysis of Qualitative Data

Ethical Considerations

INTRODUCTION

The principal focus of this work was twofold: (1) to investigate the extent to which the quality of CCEs was directly influenced by a physician's evaluative load and evaluative strain, and indirectly mediated by time delays occurring between the observation of a clerk's performance and the rating of that performance; and (2) to understand how physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students. A two-phase, sequential explanatory mixed-methods approach was utilized to address the study's central aims. The first, quantitative phase permitted an examination of the direct and indirect factors that influence the quality of CCEs. These results informed the second phase of the study, which qualitatively explored physicians' perceptions of evaluative utility, quality, cost, and practicability.

Data for this study was collected from a variety of sources. The primary sources of quantitative data were survey results acquired through the administration of a modified version of the Surgery Task Load Index (SURG-TLX), referred to as the Evaluative Task Load Index (EVAL-TLX), and participants' records of CCEs completed between January 2015 and August 2016 obtained from the E*Value Data Management System (Advanced Informatics Solutions, 2016) maintained by Indiana University School of Medicine's (IUSM) Office of Medical Student Education (MSE). The central source of qualitative data was collected through one-on-one, semi-structured interviews conducted with a sample of physicians who had completed the study survey.

RATIONALE FOR A MIXED-METHODS APPROACH

A mixed-methods approach was used to address the two aims of this work. Mixed-methods research is defined as "a procedure for collecting, analyzing, and 'mixing' or integrating both quantitative and qualitative data at some stage of the research process within a single study for the purpose of gaining a better understanding of the research problem."[62] In general, mixed-methods research is believed to be a more robust methodology for understanding complex problems or phenomena than purely quantitative or qualitative studies alone, as neither approach can truly illustrate the details of a situation without the other.[62-64] In this study, the adoption of a mixed-methods approach was imperative for interpreting the results of the quantitative analyses within the context of participants' qualitative admissions.

OVERVIEW OF THE RESEARCH DESIGN

Of the forty mixed-methods approaches generally accepted in the literature,[65] a sequential explanatory mixed-methods design was adopted for this study. According to Creswell,[63] a sequential explanatory mixed-methods approach consists of two distinct phases: a quantitative phase and a qualitative phase. Using this approach, a researcher commonly collects and analyzes the quantitative data before collecting and analyzing the qualitative data, which are then used to "refine and explain those statistical results [acquired in the first phase] by exploring participants' views in more depth."[62] Integration, or the period in the research process during which the 'mixing' of the quantitative and qualitative methods occurs,[64] can occur in two places during the research process. Commonly, a researcher will use the quantitative results to guide the selection of participants for the qualitative phase of the study.[66] Moreover, a researcher can also choose to develop the qualitative data collection protocols using the results from the quantitative phase. In this way, the researcher is able to formulate a specific protocol that will allow for a

more thorough analysis of the quantitative data during the qualitative phase.[62] As the quantitative phase precedes the qualitative phase, the principal intent of the research design (i.e., the priority) is to test one or more hypotheses using a fairly large sample and follow-up those results with a small, purposeful sample of participants to enrich or enhance the researcher's understanding of the phenomenon or question(s) of interest.[66,67]

Accordingly, this study began with a quantitative phase aimed at providing numerical estimates of the variables of interest (to be defined later in this chapter) and investigated the hypothesized causal relationships among these variables using a path analysis. The results of the quantitative analyses laid the foundation for the second phase of this study, which permitted a more in-depth exploration of physicians' perceptions of the utility, quality, cost, and practicability of CCEs via semi-structured interviews. The selection of participants for the qualitative phase was guided by the quantitative phase, as was the interview protocol. Finally, the results of the quantitative phase were interpreted within the context of participants' qualitative admissions. A timeline of this research project (Table 3-1) reinforces this phasic approach.

Table 3-1. Description of study timeline and progression

Phase	Procedures	Timeline
Quantitative Data Collection	- Administration of EVAL-TLX - Collection of E*Value records - Pairing of survey and E*Value data	June – September, 2016
Quantitative Data Analysis	- Description of demographics - Calculation of the five evaluative variables - Conduction of the path analysis	September – December, 2016
Connecting Quantitative & Qualitative Phases	- Purposeful selection of participants for phase II from phase I participants	August – October, 2016
Qualitative Data Collection	- Phase II participants interviewed on their perceptions of the utility, quality, cost, and practicability of CCEs	September – October, 2016
Qualitative Data Analysis	- Coding and thematic analysis of transcribed interviews	September – December, 2016
Integration of Quantitative & Qualitative Results	- Simultaneous interpretation and comparison of quantitative and qualitative results	January – February, 2017

Note: Adopted from Ivankova et al.[62]

PHASE I: COLLECTION AND ANALYSIS OF QUANTITATIVE DATA

The quantitative portion of this study was designed to estimate the extent to which the quality of CCEs is directly affected by evaluative load and evaluative strain, and indirectly mediated by time delays occurring between the observation of a clerk's performance and the rating of that performance. This section begins with a description of the participants and sampling strategy used in this phase of the study. A description of the study's variables, data collection techniques, and instruments pertinent to each variable will follow. This section will conclude with a description of the path analysis used to estimate the magnitude of the hypothesized causal relationships among the variables of interest.

Participants

The participants in this study (n = 93/1518) were fulltime physicians (residents, fellows, or attendings) who met three general requirements: (1) participants must have been associated with the Indianapolis campus of Indiana University School of Medicine (IUSM); (2) participants must have been affiliated with one of six required third-year clerkships (i.e., Family Medicine (FM), General Surgery (GS), Internal Medicine (IM), Obstetrics and Gynecology (OB/GYN), Pediatrics (PE), and Psychiatry (PY)); and (3) participants must have evaluated clerks (i.e., IUSM student trainees) between January 2015 and August 2016 using a 'medium' or 'long' clinical evaluation form. Formally named the "Evaluation of Student Clerkship Performance," this CCE form is consistently used to evaluate clerks at IUSM, but the *length* of the form (i.e., the number of questions) differs by medical department and may be classified as 'short,' 'medium,' or 'long.' The 'medium' CCE (Appendix A, Figure A-1), employed by the Department of General Surgery, consisted of 11 mandatory closed-ended items and two-open ended items (one mandatory, one voluntary); while the 'long' CCE form (Appendix B, Figure B-1), employed by the five remaining departments of interest, consisted of 18 mandatory closed-ended items and nine open-

ended items (one mandatory, eight voluntary). As a multidimensional instrument, both versions of the form inquire into aspects of a clerk's professional behaviors (e.g., professional attributes, professional knowledge, and demeanor) and clinical performance (e.g., data taking skills (e.g., performing a physical examination), data reporting skills (e.g., writing clinical notes), knowledge base, ability to interpret and integrate data, and capacity to make clinical judgments). Any and all residents, fellows, and attending physicians who met these three requirements were eligible for the study and invited to participate.

Sampling Protocol

The sampling protocol for the first phase of this study included all physicians affiliated with the six clerkships of interest at the Indianapolis campus of IUSM (n = 1518). To maximize physician participation in the study, the research team enlisted the aid of the Clerkship Coordinators, Clerkship Directors, and departmental Vice Chairs of Education from the Indianapolis campus. The study's aims and potential significance for medical education were presented to these individuals on three separate occasions by the primary researcher (C.J.T.) and one additional member of the research team (J.J.B). Following the third and final meeting, the Clerkship Coordinators and Directors of the included medical departments were emailed by one member of the research team (M.R.A.) using a directory maintained by the Office of MSE. This email asked recipients to forward the attached study information sheet and link to the electronic survey to the physicians in their departments, and to encourage participation in the study. A second email with identical content was sent to the departmental Vice Chairs of Education at the Indianapolis campus by a second member of the research team (J.J.B), encouraging the Vice Chairs to promote participation within their respective departments.

Physicians who met all of the aforementioned eligibility criteria were allowed to complete the survey in its entirety. If physicians who did not meet the entrance criteria (e.g.,

physicians affiliated with the Neurology Clerkship that use the 'short' form) happened to receive a request to complete the survey, they were directed to a 'stop' point, upon answering certain demographic questions, which thanked them for their participation and prevented them from continuing with the survey. Two weeks following the initial email request for participation, a second and final email was delivered to the Clerkship Coordinators, Clerkship Directors, and departmental Vice Chairs of Education as a reminder to promote survey participation among physicians within their respective departments. Submission of the survey containing the EVAL-TLX served as confirmation that participants had consented to the data-pairing procedures necessary to conduct this study. Participating physicians were additionally informed that their completion and submission of the survey qualified them for a raffle drawing, in which one 'winner' would receive a $100 gift card to Amazon (Amazon.com, Inc., 2016).

Variables

This study included two sets of variables. The first set of variables pertained to a physician's evaluative characteristics and included evaluative load, evaluative strain, time delay, score differentiation, and quality of feedback. The second set of variables described a physician's departmental demographics and included medical department, academic rank, and gender. Table 3-2 offers a brief summary of each variable. Outlined below is a detailed description of the study variables, their associated data collection instrument/measure, and how they were used to answer the research questions.

Table 3-2. Summary of study variables and data collection methods

Variable	Abbreviation	Definition	Data Collection Method
Evaluative Variables			
Evaluative Load	Load	The measurable quantity of CCEs completed by a physician between Jan. 2015- Aug. 2016.	Participants' E*Value records
Evaluative Strain	Strain	A physician's perceptions of his or her evaluative load and the perceived cognitive demands needed to complete an evaluation.	EVAL-TLX on survey instrument
Time Delay	Delay	The amount of time between a clinician's final observance or encounter with a clerk and the rating of that clerk's clinical performance.	Participants' E*Value records
Score Differentiation	Diff	A measure of one's evaluative variation within and across CCEs one completes; represents the average quality of a physician's closed-ended CCE items.	Participants' E*Value records
Quality of Feedback	QOF	A measure of the 'clarity' or 'resolution' of the required open-ended item on the CCEs one completes; represents the average quality of a physician's open-ended items.	Quality of CCE Rubric
Physician Demographic Variables			
Medical Department	Department	Participant's medical department affiliation.	Survey instrument
Academic Rank	Rank	Participant's academic rank.	Survey instrument
Gender	Gender	Participant's gender.	Survey instrument

Demographic Variables and Evaluative Strain

Participants' evaluative strain and demographic characteristics (i.e., medical department affiliation, academic rank, and gender) were determined through the administration of an electronic survey (Appendix C, Figure C-1) to all physicians from all medical departments associated with third-year clerkships on the Indianapolis campus. Only physicians who met the aforementioned entrance criteria were able to complete the survey.

Overview of the Survey Instrument

The survey was developed, administered, and managed through REDCap (Research Electronic Data Capture).[68] The introductory portion of the survey thanked participants for their voluntary participation and informed them that completion and submission of the survey granted the research team permission to access records of the CCEs they completed between January 2015 and August 2016.

The body of the survey consisted of two sections: (1) a demographics section was used to acquire information on participants' demographic characteristics and (2) the Evaluative Task Load Index (EVAL-TLX) was used to estimate participants' numeric evaluative strain. At the end of the survey, participants were asked a dichotomous yes/no item on whether they would be willing to participate in a brief 10-15 minute follow-up interview on their perceptions of CCEs. More specifically, the interview allowed participants to elaborate on their survey responses, orally describe their perceptions of the utility, quality, cost, and practicability of the CCEs they complete, and comment on the evaluative process as a whole.

Demographic Portion. The demographic portion of the survey asked participants to provide their names (for data pairing purposes only) and to indicate their departmental affiliation (i.e., FM, GS, IM, OB/GYN, PE, PY, or other); academic rank (i.e., resident, fellow,

instructor/lecturer, assistant professor, associate professor, or full professor); and gender (male, female, or elect not to answer). Each participant's departmental and gender characteristics were recorded and incorporated into a path analysis designed to investigate the first five research questions underpinning this study. The path analysis originally incorporated participants' academic rank characteristics in addition to their departmental and gender demographics, but was removed following preliminary analysis of model fit (Chapter 4).

Evaluative Task Load Index (EVAL-TLX). The second portion of the survey contained the EVAL-TLX and was used to calculate one's numeric evaluative strain ('Strain'), or one's perceived evaluative responsibilities; including both one's perceptions of his or her evaluative load and one's conceptualizations of the cognitive demands needed to complete the evaluative task. This portion of the survey was a modified version of the Surgery Task Load Index (SURG-TLX). First developed and validated in 2011 by Wilson et al.,[69] the SURG-TLX is a multi-dimensional scale designed to estimate an individual surgeon's subjective experience during or immediately following the completion of a surgical task. Since the purpose of this quantitative survey was to numerically calculate participants' evaluative strain, the EVAL-TLX estimated an individual physician's subjective experiences affiliated with the completion of CCEs.

Both the SURG-TLX and the EVAL-TLX instruments retain components from the National Aeronautics and Space Administration Task Load Index (NASA-TLX), the most ubiquitous measure of human subjective experience during the performance of a task,[33,34] and the Driving Activity Load Index (DALI), an instrument derived from the NASA-TLX designed to measure the subjective experience perceived while driving a car.[70] Drawing from both of these instruments, the SURG-TLX includes six dimensions, or subscales, that are thought to represent prevalent sources of stress within the operating room.[71] Similarly,

the EVAL-TLX has been modified to represent prevalent sources of stress within an *evaluative* setting. Table 3-3 shows the differences in phraseology between the SURG-TLX and the EVAL-TLX items.

Table 3-3. Phraseology differences between the SURG- and EVAL-TLX

SURG-TLX	EVAL-TLX
Mental demands: How mentally demanding was the procedure?	***Mental demands:*** How mentally demanding is it to complete clerkship evaluations?
Physical demands: How physically fatiguing was the procedure?	***Physical demands:*** How physically fatiguing is it to complete clerkship evaluations?
Temporal demands: How hurried or rushed was the pace of the procedure?	***Temporal demands:*** How hurried or rushed do you feel while completing clerkship evaluations?
Task complexity: How complex was the procedure?	***Task complexity:*** How complex is the task of completing clerkship evaluations?
Situational stress: How anxious did you feel while performing the procedure?	***Situational stress:*** How anxious do you feel when completing clerkship evaluations?
Distractions: How distracting was the operating environment?	***Distractions:*** How distracting is the evaluative environment when you complete clerkship evaluations?

The first three dimensions (i.e., mental, physical, and temporal demands) correspond to the demands the task imposes on the subject, and were adopted directly from the NASA-TLX. The fourth dimension (i.e., task complexity), was created by Wilson et al[69] and corresponds to the complicatedness of the task experienced by the subject. The final dimensions (i.e., situational stress and distractions) correspond to environmental demands imposed on the subject, and were adopted from the DALI instrument.

Owing to the complexity of measuring one's unique subjective experiences, the EVAL-TLX had two distinct sections. The first section asked participants to indicate the strength or magnitude with which they experienced each of the six dimensions during their completion of CCEs. Much like a physician may ask a patient to rate their pain level on a scale of one to ten, this portion of the instrument asked participants to rate how strongly they experienced each of the six dimensions associated with the completion of CCEs using six separate one hundred-point analog sliding scales (as established by the original format

used in the NASA-TLX). Anchored with "very low" and "very high" descriptors, the scales were used to gauge participants' perceived importance of each dimension. According to Hart and Staveland,[34] "scales of this sort are extremely useful, but their utility suffers from the tendency people have to interpret them in individual ways. For example, some people feel that mental or temporal demands are the essential aspects…regardless of the effort they expended on a given task or the level of performance they achieved." As such, the second section of the EVAL-TLX was designed to assign weights to each dimension according to the importance with which they were perceived. This is achieved by asking participants to directly 'rank-order' the six dimensions from '0' (i.e., least influential in the individual's subjective experience) to '5' (i.e., most influential in the individual's subjective experience). Each number can only be selected once, ensuring that a ranking system was established. In this way, the dimensions were rank-ordered based on the extent to which the respondent perceived them to have influenced his or her subjective experience of completing CCEs.

After both portions of the EVAL-TLX were completed, the researcher created an 'aggregated EVAL-TLX score' for each participant. This score was calculated by first multiplying a participant's scaled rating score for each dimension (determined by the number they selected on the corresponding 100-point sliding scale) by the ranked weighting for that dimension; this created an 'adjusted weighting score' for each dimension. The adjusted weighting scores were then summed across dimensions (e.g., mental demands + temporal demands + physical demands, etc.) and divided by 15 (i.e., the total number of comparisons that could be made amongst the six dimensions (e.g., physical demands vs. mental demands; physical demands vs. temporal demands; for example, see Table 3-4)). The resulting number was a participant's aggregated EVAL-TLX score (with the highest

possible score of 100); this represented an individual's numerical evaluative strain, which was incorporated into the study's path analysis.

Table 3-4. Example calculation of one's 'aggregated EVAL-TLX score'

Scale Title	Scale Rating (0-100)	Weighting	Adjusted Weighting Score (Scale Rating * Weighting)
Mental Demands	75	5	(75*5) = 375
Physical Demands	10	0	(10*0) = 0
Temporal Demands	85	3	(85*3) = 255
Situational Stress	50	4	(50*4) = 200
Distractions	45	2	(45*2) = 90
Task Complexity	30	1	(30*1) = 30
		Sum of 'Adjusted Weighting Scores:'	(375 + 0 + 255 + 200 +90 + 30) = 950
		Aggregated EVAL-TLX Score (i.e., Sum of Adj. Ratings/15) =	(950/15) = **63.3**

Note: In this example, physician X's experience with each of the six dimensions has been rated using six separate 100-point rating scales. Physician X's weighting was determined by the order in which he ranked his experience with each dimension. From highest to lowest, his ranked-order was: (5) mental demands, (4) situational stress, (3) temporal demands, (2) distractions, (1) task complexity, and (0) physical demands.

Validity and Reliability of the EVAL-TLX

Wilson and colleagues found the SURG-TLX to be reliable and have a high degree of validity for evaluating resident surgeons' subjective experiences of completing a surgical task.[69] Following an approach used to validate a different version of the original NASA-TLX survey,[70] the researchers experimentally manipulated operative task demands to test the instrument's sensitivity for commonly experienced stressors.[69] After demonstrating a high degree of validity, the SURG-TLX was adopted by several studies[72-76] and is believed to provide researchers with a glimpse of how one perceives a particular surgical experience. As the SURG-TLX has already been modified from the original NASA-TLX and confirmed to be well suited for surgeons, the researchers of this project contend that the task load index is transferable to the physician population as a whole. Moreover, the EVAL-TLX differed from the SURG-TLX in minimal aspects of phraseology (i.e., changing the past tense of 'was' to 'is' and using 'evaluative' rather than 'operating'). The investigators considered the EVAL-

TLX to be a reasonable extrapolation of previously validated instruments conducive for measuring the perceived evaluation strain of participants.

Data Pairing

After the time window for survey submission closed, participants' names were used to pair survey data with participants' E*Value records (i.e., CCE records). Data were 'married' by pairing a participant's numeric evaluative strain score (i.e., one's 'aggregated EVAL-TLX score' derived from the EVAL-TLX) with his or her E*Value records in preparation for analysis to understand how one's evaluative load (a purely objective measure) directly influenced strain (as surmised from the EVAL-TLX). Following this pairing, all E*Value records were stripped of identifiers and each evaluating physician was assigned a random identifier (e.g., GS1) to anonymize information for data analysis.

Evaluative Load

Evaluative load ('Load') was defined as the *measurable quantity* of evaluations, ratings, or surveys completed by a physician during a specific period-of-time. As this research specifically used CCEs as a surrogate for measuring physicians' evaluative load, this variable was calculated by counting the number of CCEs a physician completed between January 2015 and August 2016.

All medical departments affiliated with the Indianapolis campus of IUSM use E*Value Data Management (Advanced Informatics Solutions, 2016) to manage records of completed CCEs. As a cloud-based system, E*Value offers evaluation services to a variety of different organizations, including both undergraduate and graduate medical education offices throughout the country. At IUSM, E*Value is used by both faculty and students as an evaluation platform to provide feedback on the learning experience, suggest curricular modifications, and to store both faculty-completed evaluations of medical students' elective courses and clinical clerkships, as well as student-completed evaluations of preceptor

performance. Within the E*Value system, residents, fellows, and attending physicians submit department-specific clerkship evaluation forms for each student performance they observe. These datasets serve as a cumulative library for all of the CCEs completed by a particular IUSM medical department or physician for a given academic period. Included in the records are the names of the evaluating physician as well as the student being evaluated; the time frame during which the observation took place; the location of the observation (i.e., the clinical site); the date the evaluation was administratively requested; the date of submission; the evaluator's responses to each of the closed-ended questions posed on the CCE; and responses to any required or optional open-ended questions. Table 3-5 provides an abbreviated example of one's CCE records.

Table 3-5. Example CCE records for a single physician

Evaluator	Student	Begin Time Frame	End Time Frame	Clinical Site	Date Requested	Date Submitted
A	1	4/27/15	5/23/15	Methodist	5/25/15	6/3/15
A	2	4/27/15	5/23/15	Methodist	5/25/15	6/8/15
A	3	4/27/15	5/23/15	Methodist	5/25/15	6/9/15
A	4	4/27/15	5/23/15	Methodist	5/25/15	6/2/15
A	5	4/27/15	5/23/15	Methodist	5/25/15	5/25/15

As each CCE must be submitted to this system, this dataset was used to calculate a physician's evaluative load. E*Value records for each included medical department were exported to and organized within Microsoft Excel (Microsoft Corporation, 2016) in order to compute participants' evaluative load. Using Table 3-5 as an example, Physician A's CCE records indicate that he or she completed evaluations for five students within the study's timeframe (i.e., January 2015 to August 2016). As evaluative load was the measurable quantity of CCEs a physician completed during that timeframe, physician A had an evaluative load of five evaluations. This value was then incorporated into the path analysis to represent the individual's evaluative load. So as to protect the privacy of IUSM's

physicians, only physicians who completed the survey and consented to participate in the study had their E*Value records accessed and married to their Strain data.

Time Delay

In this study, time delay ('Delay') represented the time lapse between a physician's last observation of (or encounter with) a clerk and the rating of that clerk's clinical performance using a CCE. E*Value records of physicians' completed CCEs include information on the 'end time frame,' or the period during which the observation period ended, as well as the 'date submitted,' or the date and time the evaluation was submitted to the E*Value system. Calculating the time difference between the 'date submitted' and the date listed on 'end time frame' allowed the researcher to measure the time delay between the observational period of a clerk's clinical performance and the evaluation of that performance (assuming the CCE was completed the same day as the submission). Again, using Table 3-5 as an example, Physician A's last date of observation (denoted by 'end time frame') for Student 1 was subtracted from the date the evaluation was submitted (denoted by 'date submitted'). Such a calculation (e.g., 6/3/15 – 5/23/15) resulted in a delay of 11 days and represented Physician A's time delay for that particular evaluation. Time delays were calculated for each evaluation completed by a physician and an 'aggregated time delay' score was created by averaging all of an individual's separate 'time delays' together. This aggregate score represented the average amount of time that elapsed between observing and submitting an evaluation of a clerk and was used to represent 'Delay' in the path analysis.

Quality of CCEs

Within this study, 'quality of CCEs' represented the collective quality measure of both the closed-ended and open-ended CCE items on the evaluation. Because the closed-ended and open-ended items allow a physician to rate different aspects of a clerk's

performance in distinct ways, a separate outcome variable was warranted for each portion of the evaluation. 'Score differentiation' was defined as a measure of evaluative variation and was used to represent the quality of the closed-ended CCE items. 'Quality of feedback' was defined as an estimate of the 'clarity' of written feedback and was used to represent the quality of the single required open-ended CCE item.

Because 'quality' is a theoretical construct that is not readily observed,[51] the operationalization of this construct was challenging. Though there is no single, perfect measure of quality, the utilization of several 'quality indicators' as a proxy of this construct were used.[51] The quality indicators chosen drew from published measures of satisficing[51,77,78] and aspects of constructive feedback[79] and were used to calculate a 'quality score' for each of the two quality variables. Each quality score was then incorporated into the path analysis.

Score Differentiation

Score differentiation ('Diff') measured one's evaluative variation within and across CCEs completed by a single physician. Within evaluations and surveys, non-differentiation (or a lack of variation among items) can manifest as 'straight-lining,' which occurs when respondents or evaluators select the same choice for all (or nearly all) responses, such that a 'straight-line' of responses is visible.[77] Straight-lining is suggestive of poor-quality as it implies that the respondent is 'satisficing,' or selecting an answer deemed 'good enough' in order to expedite the survey or evaluative process.[51,52] A measure of score differentiation was assessed for each physician by calculating the physician's mean variance across items by student (i.e., item-to-item variance (abbreviated ITI variance)) for five randomly selected evaluations he or she had completed. As variance describes the spread of a set of numbers or observations,[80] a low ITI variance (i.e., close to zero) was considered indicative of

straight-lining, while a high ITI variance (i.e., close to one or above) indicated a low degree of non-differentiation, implying a higher degree of rating quality.[51,81]

Although straight-lining is accepted as an indicator of (poor) quality,[51,77,81] it should be noted that the presence of straight-lining is not *always* evidence of satisficing. It is possible for a respondent or evaluator to consciously and purposefully respond to each item and still produce a visible 'straight-line.'[77] To account for this, a second quality measure was necessary. As such, one's evaluation-to-evaluation (ETE) variance was assessed by calculating the participant's mean variance across evaluations by item for the same five evaluations. In other words, a physician's ETE variance indicated the extent to which the participant habitually graded CCEs in a similar manner across students, relative to his or her peers. As this measure represented how well physicians express performance differences across evaluations by item, a low ETE variance (i.e., close to zero) indicated that a physician was not likely to note individual differences among students relative to his or her peers. A high ETE variance (i.e., close to or above one) indicated that a physician was more likely to detect individual strengths and weaknesses between evaluated students. An 'aggregated differentiation score' was calculated for each physician by summing his or her ITI and ETE variances. This aggregate score represented the average quality of a physician's closed-ended CCE items and was used to represent 'Diff' in the path analysis.

Quality of Feedback

'Quality of Feedback' ('QOF') was a measure of the 'clarity' or 'resolution' of the one required open-ended item included on the CCE. Using the definition provided by Williams et al.,[37] quality of the open-ended item was "analogous to high-resolution or high-definition television displays," such that a 'high-resolution' CCE implied that the written feedback on the evaluation contained an appreciable amount of detail, and a clear, precise description of the clerk's clinical prowess. The QOF for each individual physician was assessed using a

'Quality of CCE Rubric' (Appendix D, Table D-1) that consisted of four quality measures. These quality measures were designed to assess the substance of the physicians' responses to the open-ended question. This was done by judging the frequency with which a physician included:

- A diagnostic observation. This measure required a physician's written feedback to comment on at least one observable behavior and/or skill (e.g., "X did a great job. He was calm, thoughtful, and polite. He arrived on time and wore proper attire");
- A formative comment. This was a two-part measure that required a physician's written feedback to suggest at least one specific area of strength and/or weakness and provide a clear explanation of what was done well or how to improve (e.g., "It was clear he really cared for his patients' well-being by being an excellent advocate for his patients");
- A specific remark. This was also a two-part measure that required a physician's written feedback to be both specific (i.e., not globally descriptive, like "did well") and uniquely formulated for the evaluated student (e.g., "Y gave a very detailed presentation on kidney stones"); and
- A practical suggestion. Practical suggestions were thorough, useful, and clearly actionable; the classification of a physician's written feedback to the student as 'practical' was directly related to the presence of a 'specific' comment, such that a physician could not receive credit for having included a practical comment if they did not also include a specific comment. For example, the comment, "Y gave a very detailed presentation on kidney stones. Though her presentation skills are above average among her peers, she could work on improving eye contact with her audience during future presentations" includes both a specific comment and a practical comment. Had the physician instead written, "Y gave a very detailed presentation on kidney stones. She should keep reading and learning," the physician would have received points for providing a specific comment but

no points for a practical comment, as the physician provided no guidance as to what the student should read to improve her performance.

Each physician was judged on the frequency with which he or she included a diagnostic observation, a formative comment, a specific remark, and a practical suggestion within each of the same five randomly scored CCEs used in the closed-ended items analysis.[79,82] The presence of each of these comments was considered a positive indicator of feedback quality, and the inclusion of each comment type in an evaluation garnered a physician 1/5th of a point (or 0.2 points). A physician who included a diagnostic, formative, specific, and practical comment in each of his or her five evaluations received a total of one point for each type of comment (for a maximum possible total of four points). A physician's total 'quality of feedback' score was tallied and used in the path analysis to represent the quality of the written feedback provided in response to the open-ended item.

Although the examples provided with the definition of each comment type are independent of one another, it was not necessary for a physician to include four separate sentences within each evaluation to garner full credit. Rather, the researcher utilizing the Quality of CCE Rubric to assess the quality of a physician's feedback on the open-ended item needed only to observe the essence of each quality measure within the text of the physician's written feedback. To further illustrate this point, Table 3-6 provides a calculation of the QOF score for an example comment (portrayed as 'Evaluation 1' in the table).

Table 3-6. Example calculation of a single open-ended comment using the 'Quality of CCE Rubric'

	Evaluations Yes (0.2)/ No (0)					
'Quality of Feedback' Quality Measures	**1**	**2**	**3**	**4**	**5**	**Total**
Diagnostic observation: a comment based on observable behaviors and/or skills	0.2	0	0.2	0.2	0.2	0.8
Formative comment: a two-part comment suggesting specific areas of strengths and/or weaknesses *with* a clear explanation of how to improve or what was done well	0.2	0.2	0.2	0.2	0.2	1.0
Specific remark: a two-part comment that was both not globally descriptive, like "did well," and uniquely formulated for the evaluated student	0.2	0	0.2	0.2	0.2	0.8
Practical suggestion: a comment that was thorough, useful, and clearly actionable	0.2	0	0.2	0.2	0	0.6
					Total	3.2/4

Note: The example comment graded using the Quality of CCE Rubric for Evaluation 1 stated, "Y gave a very detailed presentation on kidney stones. Though her presentation skills are above average among her peers, she could work on improving eye contact with her audience during future presentations."

This physician's written feedback for 'evaluation 1' received the total maximum points possible for a single evaluation (i.e., 0.2 x 4 = 0.8). The physician received credit for having a 'diagnostic observation,' as he noted that the student did not make eye contact with her audience. Full points were awarded for having a 'formative comment,' as he noted one of the student's strengths (e.g., "her presentation skills are above average among her peers") and suggested that she could improve on this strength by "making eye contact with her audience during future presentations." Additionally, the physician's feedback was considered to be specific, because it was clear that he was referencing an individual student and was able to address her attributes directly; and practical, as he provided an actionable suggestion to improve her presentation performance.

Inter-Rater Reliability and Percent Agreement

Three members of the research team (C.J.T., R.L.S., and A.B.W.) were responsible for calculating one's QOF using the Quality of CCE Rubric. To guarantee that the rubric was used accurately and consistently among the 'graders,' each grader was initially responsible for

scoring the written feedback for the same 20 randomly-selected physicians. All three graders were provided definitions of each comment type along with five examples of accurately graded written feedback. After reviewing the definitions and examples, each grader scored the written feedback and calculated a QOF score for each physician using the Quality of CCE Rubric independently. After each rater had scored the written feedback for the same twenty physicians, percent agreement and inter-rater reliability were calculated using Excel (Microsoft Office, 2016) and SPSS software (IBM Analytics, 2016, Version 24), respectively, to ensure consistency among the graders and their usage of the rubric.

Percent agreement is commonly utilized in quantitative research as a measure of rater agreement. For studies involving interval, ratio, or ordinal data, percentages of agreement can be "expressed as the percentage of ratings that are in agreement within a particular interval."[83] Accordingly, adequate percent agreement in this study was defined as having ≤0.6 points of difference among the scores produced by the three graders, as this represented 15% of the maximum score (4) each participant could receive for their five randomly selected evaluations. As 75% agreement is considered the minimum value that should be achieved when calculating percent agreement,[84] this was considered the cut-off for this measure.

Inter-rater reliability (IRR) is another commonly utilized measure of rater agreement in quantitative data. IRR was calculated using a two-way mixed model, single measures intra-class correlation statistic (ICC). ICC was selected as a measure of IRR because this statistic is commonly employed in studies that use two or more raters assessing the same subset of participants.[83] ICC values range from -1 to 1, where '-1' indicates an absolute lack of agreement between raters and '1' is demonstrative of perfect agreement between raters. ICC values close to zero indicate that the raters are in random agreement with each other.[83] As ICC values of 0.75-1.00 are considered an excellent

measure of inter-rater reliability,[83] 0.75 was considered the group cut-off value for this measure.

Percent agreement calculations less than 75% and/or ICC values less than 0.75 prompted graders to discuss scoring discrepancies and rubric guidelines. Once each rater felt comfortable with their usage of the rubric, each rater re-evaluated the written feedback for the same twenty physicians. This process of evaluating written feedback and meeting to discuss rater discrepancies was repeated until an ICC value of ≥0.75 and a percent agreement of ≥ 75% were achieved among graders for the set of 20 randomly selected physicians that were used for rubric training. Once grading consistency had been established, each grader scored the quality of feedback for another 24 or 25 physicians using the Quality of CCE Rubric independently.

Statistical Analyses

Software

The statistical analyses conducted in this study were performed using two statistical software packages. Descriptive statistics and examination of relevant statistical assumptions were performed using SPSS (IBM Analytics, 2016, Version 24), while *LISREL* (Scientific Software International, 2016, Version 9.2) was utilized to conduct a path analysis aimed at answering the five research questions underpinning the quantitative portion of this study.

Path Analysis

A path analysis was used to examine the relationships among two demographic variables (i.e., Department and Gender) and the five evaluative variables (i.e., Strain, Load, Delay, Diff, and QOF) of interest. A path analysis is a form of causal modeling that allows one to estimate the strength or magnitude of hypothetical causal relationships amongst any number of independent and dependent variables using correlational data.[80,85] These

hypothetical relationships are customarily displayed by a path model (Figure 3-1), defined as "a pictorial representation of the theoretical explanations of cause-and-effect relationships among a set of variables." A path model includes a number of causal paths (denoted as arrows) that represent the direction of the predictive or causal flow and can be uni- (i.e., recursive) or bi-directional (i.e., non-recursive).[80] Each causal path has a corresponding path coefficient (abbreviated '*p*') that represents the degree of standard deviation change in each dependent variable that is explained or predicted by a one-unit standard deviation change in the independent variable.[85] Each path coefficient is denoted with a subscript consisting of two numbers (e.g., p_{41}). The first number ("4" in the example) indicates the variable affected, while the second number ("1" in the example) indicates the source of the cause (Figure 3-1).[85] Path models also include 'disturbances' (abbreviated 'd') or error terms that represents "all other influences on the outcome variables other than those shown in the model."[80]

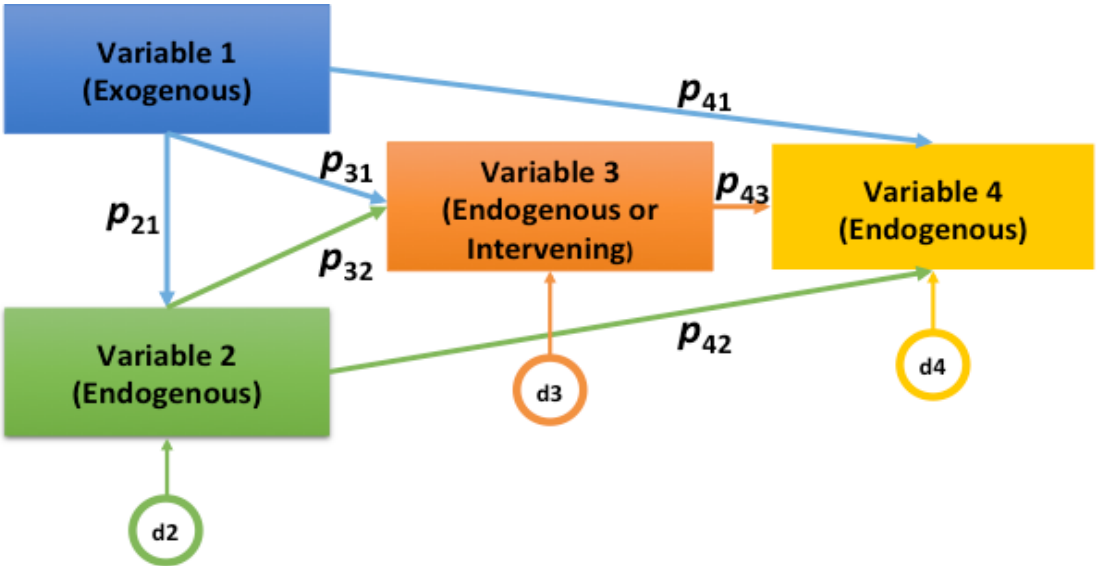


Figure 3-1. Example of a path model
Note: The model illustrates path coefficients (*p*), endogenous, exogenous, and intervening variables, and disturbances (d).

Within causal modeling, variables are further classified as exogenous (i.e., not influenced or predicted by a variable included in the causal model) or endogenous (i.e., affected, influenced, or predicted by at least one variable in the model).[85] Exogenous variables are positioned causally prior to the endogenous variables included in the model and are always considered independent variables. Endogenous variables are neither definitively classified as dependent or independent variables. Rather, the determination of endogenous variables as either independent or dependent variables depends on the path being examined and the research question.[80,85] For example, Variable 2 in Figure 3-1 is an endogenous, dependent variable in p_{21}, as it is directly affected by Variable 1; however, Variable 2 is an endogenous, independent variable in p_{42} because it directly affects Variable 4. The relationships between exogenous and endogenous variables are described as 'total effects,' which can be further broken down into 'direct' and 'indirect' effects. A direct effect occurs when a variable affects, predicts, or causes an effect on another variable *directly* (as in p_{41}, where Variable 1 affects Variable 4). Conversely, an indirect effect occurs when a variable affects another variable through a third, intervening variable (e.g., Variable 1 affects Variable 4 through Variable 3); in other words, the effects of the first variable on the second variable are mediated by an intervening variable (e.g., Variable 3).[86]

Model Specification

According to Mertler and Vannatta,[85] "the specification of the [causal or conceptual] model is a formal declaration of the researcher's beliefs regarding the causal links among the variables." The path model is ultimately formed by a variety of influences, including the researcher's experiences and observations with the variables, findings presented in the literature, and logic/intuition. The conceptual model used in this study is shown in Figure 3-2. For simplicity, the abbreviated variable notation introduced in Table 3-2 is used when

referencing the variable's roles within the path model and during the discussion of the results in Chapter 4.

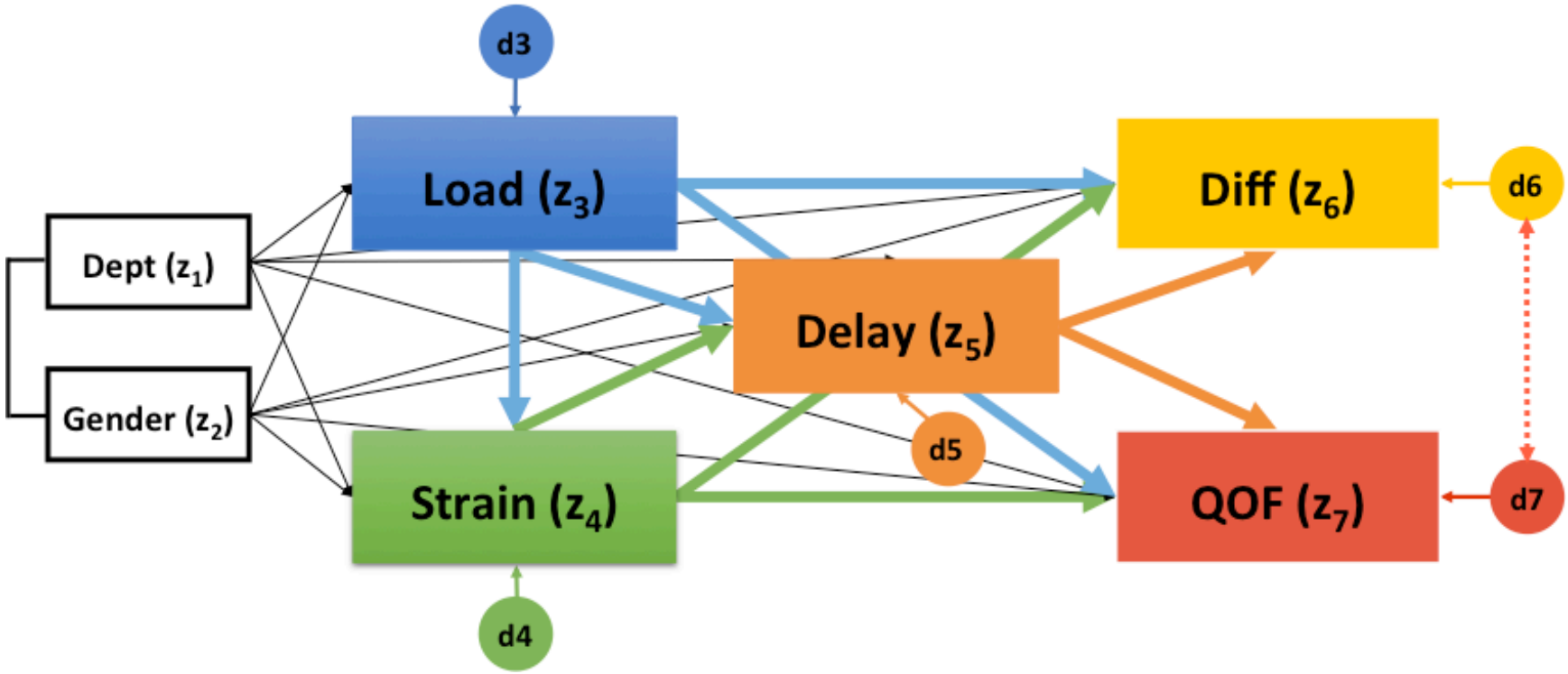


Figure 3-2. The conceptual model for the study

Note: The model details the hypothesized relationships among participants' demographics, evaluative responsibilities, and the quality of CCEs. Within the model, 'z#' denotes the ordering of the specified variable within the model and 'd#' refers to the error associated with that same endogenous variable.

The study's conceptual model included two demographic variables (i.e., Department and Gender) and five evaluative variables (i.e., Strain, Load, Delay, Diff, and QOF) and was considered fully saturated ($df = 0$) with perfect goodness of fit ($\chi^2 = 0$). Rank was originally included in the conceptual model alongside Department and Gender, but was removed following preliminary analysis of model fit when it was discovered that rank had no significant associations with any other variable within the model. The demographic variables were positioned causally prior to the evaluative variables, and thus served as exogenous variables in the model. Both demographic variables were proposed to have a direct relationship with each of the five evaluative variables. Moreover, Department and Gender were allowed to intercorrelate. Using path analysis, it was also possible to calculate the indirect and total effects of the demographic variables on Strain, Delay, Diff, and QOF. The existence of an indirect relationship between the demographic variables and Load was not possible, as there was no intervening variable positioned between Load and the demographic variables. A detailed summary of the direct and indirect relationships proposed among the variables in the model is presented in Table 3-7.

Table 3-7. Summary of the direct and indirect effects of each variable

Variable	Direct Effect On	Indirect Effect On
Department	Load, Strain, Delay, Diff, QOF	Strain, Delay, Diff, QOF
Gender	Load, Strain, Delay, Diff, QOF	Strain, Delay, Diff, QOF
Load	Strain, Delay, Diff, QOF	Delay, Diff, QOF
Strain	Delay, Diff, QOF	Diff, QOF
Delay	Diff, QOF	-

Given that the demographic variables were categorical, it was necessary to dummy-code these variables to include them in the analysis. A description of the dummy-coded variables, their reference groups, and comparison group(s) are included in Table 3-8. The selection of reference groups was primarily based on group sample size, with the largest group chosen to represent the reference group.

Table 3-8. Dummy-coded reference groups for the demographic variables

Variable	Reference Group	Comparison Group(s)
Department	Internal Medicine (IM)	Family Medicine (FM), General Surgery (GS), Obstetrics and Gynecology (OB/GYN), Pediatrics (PE), Psychiatry (PY)
Gender	Men	Women

Model Estimation and Interpretation

Statistical analysis of the path model generated unstandardized and standardized path coefficients (direct, indirect, and total effects) and significance levels (in t-values) that were interpreted to estimate the relative influence(s) of the exogenous variables on the endogenous variables (e.g., Gender on Load) and the endogenous variables on other endogenous variables (e.g., Load on Strain). Because the path model included continuous variables (i.e., Strain, Load, Delay, Diff, and QOF) and categorical variables (i.e., Department and Gender), the standardized path coefficients were favored and reported in Chapter 4. Standardized coefficients remove the specific scaling information pertinent to each variable and are interpreted in standard deviation (SD) units. This allowed for easy comparison among the variables. Moreover, standardized coefficients can be directly interpreted as effect sizes. Using Cohen's *d*, small effects are classified as 0.20-0.49, medium as 0.50-0.79, and large as ≤0.80.[87]

Path Analysis Assumptions

Prior to conducting the path analysis presented in Figure 3-2 using *LISREL* (Scientific Software International, 2016, Version 9.2), the data were organized, cleaned, and thoroughly examined to ensure that the assumptions of the statistical analyses had been met. For path analysis, the eight underlying assumptions of multiple regression were examined prior to evaluating the assumptions of path analysis. Three of the assumptions of multiple regression are concerned with the independence of the independent variables (i.e., the independent variables are fixed, measured without error, and observations are

independent of one another) and are considered factors of the research design. The five remaining assumptions of multiple regression address the normality of the data distribution, linearity (i.e., the requisite that the relationship between the independent variables and the dependent variable is linear), and homoscedasticity (i.e., the requisite that the variance of the residuals across all levels of the independent variables is constant).

In addition to the eight assumptions associated with multiple regression, there are an additional five assumptions unique to path analysis. Path analysis assumes: (1) the proposed model is an accurate representation of the actual causal sequence; (2) all variables considered to be direct causes of each endogenous variable are included in that variable's structural equation; (3) the model contains no reciprocal causation (i.e., the model contains only one-way causal flow); (4) the variables display additive, causal, and linear relationships; and (5) all exogenous variables contained within the model are measured without error.[85]

Use of the Path Model to Address the Study's Research Questions

The next series of figures (Figures 3-3 through 3-8) demonstrates how each of the five research questions was addressed via the path analysis. Highlighted areas are meant to demonstrate how each specific research question fit within the scope of the conceptual model. Additionally, path coefficients (e.g., p_{53}) have been included in the figures to indicate which specific path(s) was/were interpreted to answer each research question. For simplicity, the paths from the exogenous variables to the endogenous variables have been removed, but the proposed relationships established in Figure 3-2 remain.

Analysis for Research Question 1

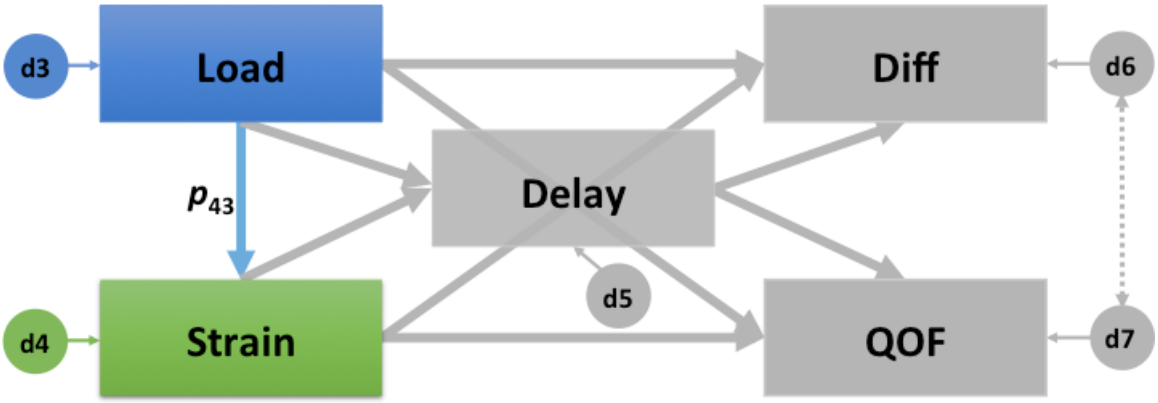


Figure 3-3. Analysis for Research Question 1
Note: The highlighted portion of the conceptual model addresses how the path analysis will answer the question, "How does physician evaluative load influence evaluative strain?"

Interpretation of direct path coefficient p_{43} was used to determine the relative influence and significance of physicians' evaluative load on evaluative strain. As this study only utilized physicians' CCEs as a surrogate measure of their evaluative load, d3 represented measurement error in Load and any additional sources of error derived from physicians' additional evaluative tasks (e.g., clinical evaluations of residents) not included in the model that likely contributed to their evaluative load. Similarly, d4 represented measurement error in Strain and any additional sources of error derived from physicians' perceptions of their additional evaluative tasks (e.g., perceptions of their completed number of clinical evaluations for residents) that were not included in the model and were likely to influence a physician's overall evaluative strain.

Analysis for Research Question 2

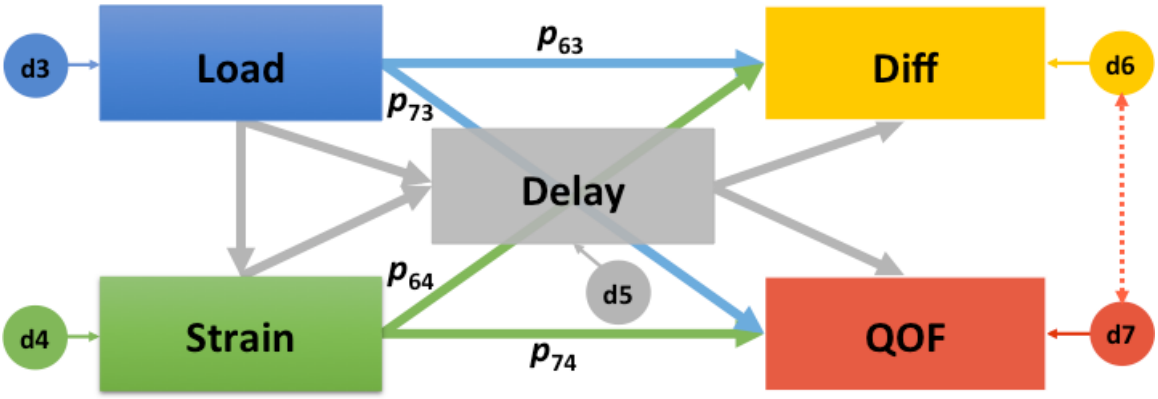


Figure 3-4. Analysis for Research Question 2
Note: The highlighted portion of the conceptual model addresses how the path analysis will answer the question, "How do evaluative load and evaluative strain directly affect the quality of CCEs, as measured by the degree of score differentiation and the quality of feedback?"

The interpretation of four direct path coefficients was needed to determine the extent to which evaluative load and evaluative strain directly influenced the quality of CCEs. The relative contributions of Load on Diff and Load on QOF were determined through analysis of the direct path coefficients p_{63} and p_{73}, respectively; while the estimated influences of Strain on Diff and Strain on QOF were determined through interpretation of the direct path coefficients p_{64} and p_{74}, respectively. Within the model, d6 and d7 represented measurement error in the variables, as well as other sources of error that may have affected the degree of score differentiation among closed ended-items and quality of feedback among the open-ended item, respectively.

Analysis for Research Question 3

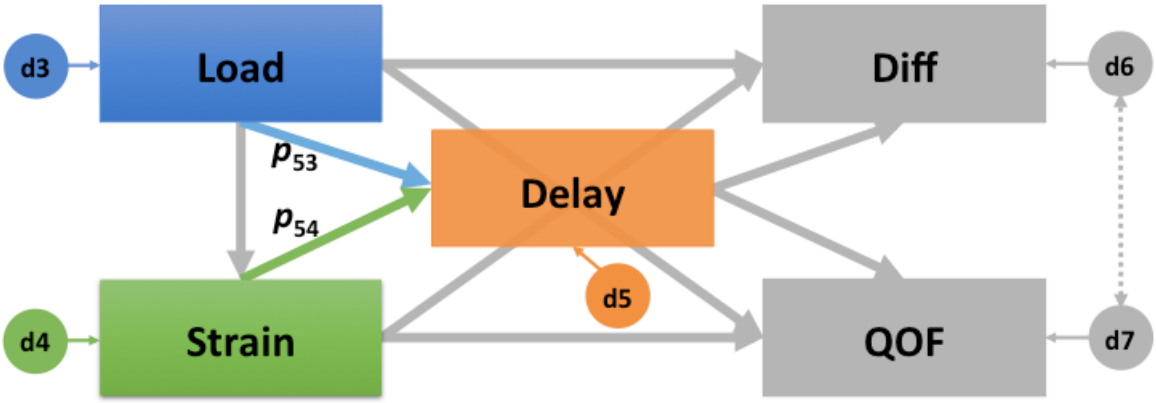


Figure 3-5. Analysis for Research Question 3
Note: The highlighted portion of the conceptual model addresses how the path analysis will answer the question, "To what degree are evaluative load and evaluative strain associated with the length of time delay between a clinician's final observation of (or encounter with) a clerk and the rating of that clerk's clinical performance?"

Analysis of direct path coefficients p_{53} and p_{54} were used to determine the relative contributions of evaluative load and evaluative strain on the timeliness of CCE submissions, respectively. Within the model, d5 represented measurement error in Delay and other sources of error that may have contributed to the length of these time delays (e.g., a physician's patient load) that were not included in the model.

Analysis for Research Question 4

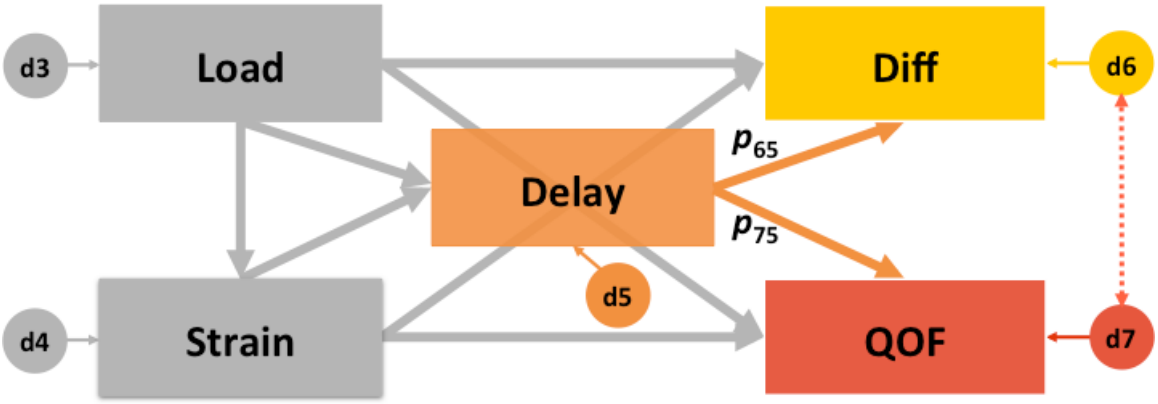


Figure 3-6. Analysis for Research Question 4
Note: The highlighted portion of the conceptual model addresses how the path analysis will answer the question, "To what extent do time delays directly influence the quality of CCEs?"

Interpretation of direct path coefficient p_{65} was necessary to determine the relative influence of the length of time delay in CCE submission on the degree of score differentiation present among the closed-ended CCE items. As score differentiation was a measure of the variation present among the closed-ended items, high degrees of differentiation (i.e., large variances) were considered a measure of good quality among this portion of the evaluation. Thus, an examination of p_{65} permitted an understanding of the relative contribution of the length of time delay in CCE submission on the quality of the closed-ended items on the evaluation. Similarly, analysis of direct path coefficient p_{75} permitted an interpretation of the relative contribution of the length of time delay in CCE submission on the quality of feedback present among the open-ended items. As QOF was a composite measure of the 'clarity' or 'resolution' of the open-ended items, examination of p_{75} provided an understanding of how the length of time delay in CCE submission influenced the quality of the open-ended items on the evaluations.

Analysis for Research Question 5

An examination of several path coefficients was necessary to understand how physician evaluative load (research question 5A) and evaluative strain (research question 5B) indirectly influenced the quality of CCEs through the mediating effects of time delay. To streamline the interpretation of these path coefficients, a separate figure has been provided for the indirect effects of Load on Diff and QOF (Figure 3-7) and Strain on Diff and QOF (Figure 3-8).

Analysis for Research Question 5A

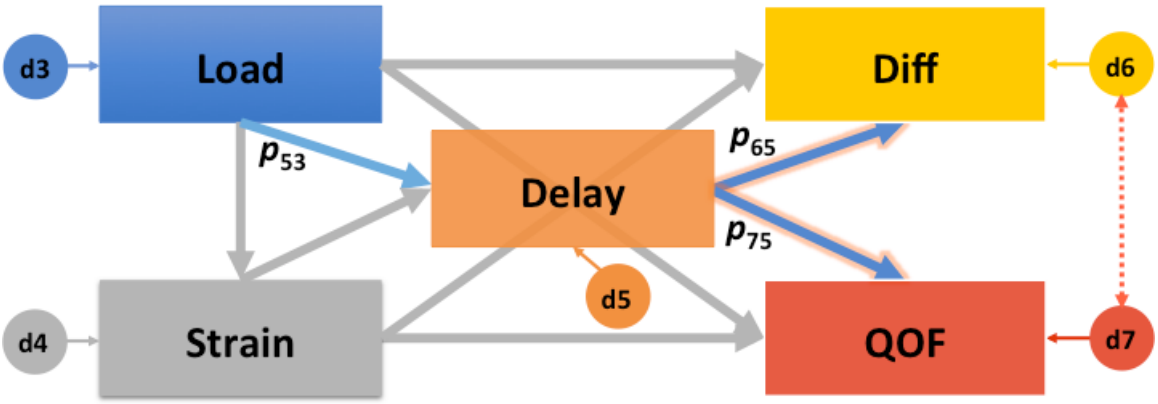


Figure 3-7. Analysis for Research Question 5A
Note: The highlighted portion of the conceptual model addresses how the path analysis will answer the question, "To what extent do time delays mediate the influences of evaluative load on the quality of CCEs?" The two-toned arrows represent the indirect effects of Load on Diff and QOF, respectively.

Investigation of the relative contribution of evaluative load on score differentiation through time delay required interpretation of the indirect path coefficient of Load on Diff. Indirect path coefficients are the product of the direct path coefficients of interest. As such, the indirect path coefficient for Load on Diff is the product of the direct path coefficient for Load on Delay (i.e., p_{53}) and the direct path coefficient for Delay on Diff (i.e., p_{65}), or p_{53}*p_{65}. Interpretation of p_{53}*p_{65} (provided by *LISREL*) permitted an understanding of how physician evaluative load indirectly contributed to the quality of the closed-ended items on the evaluations when mediated by the length of time delay in CCE submission. Similarly, the

indirect effect of Load on QOF through Delay was interpreted by p_{53}*p_{75}. Examination of the indirect effect of Load on QOF through Delay allowed for understanding of how physician evaluative load indirectly contributed to the quality of the open-ended items on the evaluations when mediated by the length of time delay in CCE submission.

Analysis for Research Question 5B

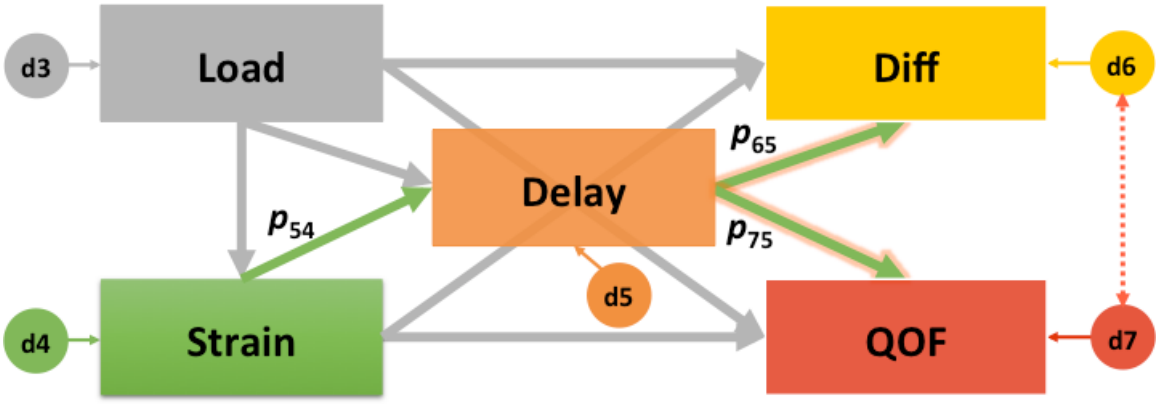


Figure 3-8. Analysis for Research Question 5B
Note: The highlighted portion of the conceptual model addresses how the path analysis will answer the question, "To what extent do time delays mediate the influences of evaluative strain on the quality of CCEs?" The two-toned arrows represent the indirect effects of Strain on Diff and QOF, respectively.

To determine the extent to which time delays mediated the influences of evaluative strain on the quality of CCEs, it was necessary to interpret the relative indirect contributions of Strain on Diff (i.e., p_{54}*p_{65}) and Strain on QOF (i.e., p_{54}*p_{75}). Interpretation of p_{54}*p_{65} permitted an understanding of how physician evaluative strain contributed to the quality of the closed-ended items on the evaluations when mediated by the length of time delay in CCE submission; while analysis of p_{54}*p_{75} explained the extent to which the quality of the closed-ended items on the evaluations was indirectly influenced by physician evaluative strain through the length of time delays in CCE submission.

PHASE II: COLLECTION AND ANALYSIS OF QUALITATIVE DATA

While the quantitative phase of the study sought to numerically estimate the effects of physicians' evaluative responsibilities on the quality of the clinical evaluations they produce, the qualitative phase of the study was designed to explore how physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students. This section begins with a description of the participants and sampling strategy used. A depiction of the data collection techniques and data analyses will follow. The chapter concludes with a discussion of the study's ethical considerations.

Participants and Sampling Strategy

Participants for phase II were selected using a stratified, purposeful sampling technique. In general, purposeful sampling allows the researcher to intentionally select individuals, thus promoting a better, in-depth understanding of the research question(s).[63] As is customary with a two-phase, sequential explanatory mixed-methods approach, the participants in phase II must have participated in phase I.[62] Individuals selected for phase II must have answered "yes" to the dichotomous yes/no survey question, which queried participants on their willingness to provide information in the form of a follow-up interview. From the pool of respondents willing to be interviewed, the investigator purposefully contacted participants with variable background demographics and evaluative responsibilities from each medical department surveyed. In the case where a selected individual declined the researcher's invitation to participate in phase II, the selection process was repeated until at least one participant from each medical department consented to participate or all participants from a given medical department had declined the invitation.

Data Collection

Phase II data were collected via semi-structured, one-on-one interviews (n = 8). According to Merriam,[88] semi-structured interviews fall in the middle of an interview continuum, between 'highly-structured' and 'unstructured' interviews. Semi-structured interviews utilize an interview guide or protocol to direct the topic of conversation, while allowing the researcher to follow any tangible thoughts or comments that arise during the course of the conversation. Moreover, this guide is used flexibly, such that there is no pre-defined order or specific wording to the questions.[88]

Phase II participants selected the time and location of their interview. An interview protocol (Appendix E) was developed after the collection and analysis of the quantitative data, as is common in a two-phase, sequential explanatory mixed-methods approach. This protocol was used to prompt and engage participants, but the interviews were not limited to questions included on the protocol. Moreover, not all questions included on the protocol were discussed in each interview. The questions included on the interview protocol were designed to stimulate discussion about the participants' survey responses, the evaluative process as a whole, and their views on the utility, quality, cost, and practicability of CCEs.

Each interview was conducted by the principal member of the research team (C.J.T.) and lasted 20-30 minutes. During each interview, the researcher took notes regarding participants' expressions and non-verbal gestures. These notes served to remind the researcher of important observational data obtained during the interview that were used to provide context during transcription.[88] Finally, each interview was audio-recorded using the Voice Memos application (Mobile Nations, 2016), with permission of the participants. After being paired to each participants' previous data points (i.e., the survey and E*Value records), any identifying information was stripped and recordings remained confidential. Each interview was transcribed verbatim using Transcribe (Wreally Studios, 2016).

Data Analysis

Analysis of the qualitative data occurred "simultaneously with data collection."[88] According to Merriam,[88] "To wait until all data are collected is to lose the opportunity to gather more reliable and valid data; it is also to court disaster, as many a qualitative researcher has been overwhelmed and rendered impotent by the sheer amount of data in a qualitative study." Using this approach, the primary researcher transcribed the audio recording of each interview immediately following its completion. The transcription was read and reread, notes were written in the margins, and the researcher documented her thoughts and feelings as they arose, as to detail any personal biases. The researcher also 'memo-ed' her reflections, recorded possible themes, and noted ideas that were utilized in the subsequent interview. This repetitive cycle allowed the researcher to continuously inform the next round of data collection while keeping the researcher aware of emerging themes.[88] Themes were generated through a thematic analysis that utilized coding, a process that groups observations based on similarities and denotes them with a meaningful 'code,' or name.[89] This thematic analysis was performed in duplicate by the researcher and a second qualitative researcher (outside of the research team). In the event of coding discrepancies, the researchers reanalyzed the codes until a consensus was established. Coding continued until both researchers were satisfied that coding saturation had been reached, or the point at which analyzing data yielded no new findings or meanings.[63]

ETHICAL CONSIDERATIONS

Ethical approval to conduct the study was provided by the Institutional Review Board (IRB) at Indiana University (Study number 1604657202). All quantitative and qualitative data was electronically secured on a password-protected server, also maintained by Indiana University.

CHAPTER FOUR

Quantitative and Qualitative Results

Introduction

Part I: Quantitative Results

Part II: Qualitative Results

Chapter Summary

INTRODUCTION

The purpose of this two-phase, sequential explanatory mixed-methods study was twofold. The first, quantitative phase was devised to numerically estimate the extent to which the quality of clinical clerkship evaluations (CCEs) was directly affected by physician evaluative load and evaluative strain and indirectly mediated by time delays in clerkship evaluation submissions. To achieve this aim, this phase of the study included the calculation of the five evaluative variables (i.e., evaluative strain, evaluative load, time delays, differentiation, and quality of feedback) and the three demographic variables (i.e., gender, medical department affiliation, and academic rank). A path analysis was then used to estimate the magnitude of the hypothesized, causal relationships among the first seven variables. The second, qualitative phase explored how physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students. These perceptions were gathered through semi-structured, one-on-one interviews and were analyzed using thematic analysis.

This chapter presents the quantitative and qualitative results of this study. Presented first, the quantitative findings will begin with a description of the demographic characteristics of the participants for phase I. Next, examination of the necessary statistical assumptions will lead to the results of the path analysis used to answer the five research questions underpinning this phase of the study. The second portion of this chapter will detail the qualitative findings. After describing the demographic characteristics of the participants for phase II, the results of the thematic analysis used to understand how participants perceive the CCEs they complete will be presented by sub-research question.

PART I: QUANTITATIVE RESULTS

Study Participants

Of the 1518 surveys that were potentially administered, 160 (11%) were completed. Of this value, 67 were excluded from the study due to incomplete survey responses or missing E*Value records (pulled from January 2015 to August 2016) that prevented the calculation of participants' evaluative strain and/or evaluative load. This resulted in a total of 93 surveys viable for analysis (6.1%). Participation rate by medical department is displayed in Table 4-1.

Table 4-1. Response rate by medical department

Medical Department	Participating Physicians (*n*)	Physicians Eligible to Participate (*n*)	Participation Rate (%)
FM	1	213	0.5%
GS	11	187	5.9%
IM	34	501	6.8%
OB/GYN	6	110	5.5%
PE	34	428	7.9%
PY	7	79	8.9%
Total	93	1518	6.1%

Fifty-five percent (51/93) of the surveyed participants were male. The majority of participants cited departmental affiliation with either IM (*n* = 34) or PE (*n* = 34), but all six medical departments surveyed had a least one physician respond. Thirty-four percent (32/93) of participants identified as residents, 25.8% (24/93) identified as assistant professors, 26.9% (25/93) identified as associate professors, and 12.9% (12/93) identified as full professors. No fellows or instructors/lecturers completed the survey.

When asked about the extent to which one feels overwhelmed by the number of CCEs he or she is asked to complete, 54.8% (51/93) of respondents reported feeling "sometimes overwhelmed," while only 9.7% (9/93) of participants reported feeling "very often overwhelmed." Respondents were also polled on their additional evaluative

responsibilities. The overwhelming majority of participants reported also having regular evaluative duties related to resident evaluations (96%), professional society surveys/questionnaires (70%), and institution-specific surveys/questionnaires (85%). A detailed description of the participants from the quantitative phase of this study is summarized in Table 4-2.

Table 4-2. Description of phase I participants

Variable		*n* (%) [n_{Total} = 93]
Gender		
	Male	51 (55.0%)
	Female	42 (45.0%)
Medical Department		
	FM	1 (1.1%)
	GS	11 (11.8%)
	IM	34 (36.6%)
	OB/GYN	6 (6.4%)
	PE	34 (36.6%)
	PY	7 (7.5%)
Academic Rank		
	Resident	32 (34.4%)
	Fellow	0 (0%)
	Instructor/Lecturer	0 (0%)
	Assistant Professor	24 (25.8%)
	Associate Professor	25 (26.9%)
	Full Professor	12 (12.9%)
Extent to which one feels overwhelmed by one's evaluative load		
	I never feel overwhelmed.	21 (22.6%)
	I sometimes feel overwhelmed.	51 (54.8%)
	I often feel overwhelmed.	12 (12.9%)
	I very often feel overwhelmed.	9 (9.7%)
Other surveys/evaluations regularly completed		
	Resident evaluations	89 (96.0%)
	Professional society surveys	65 (70.0%)
	Institution-related surveys	79 (85.0%)

Path Analysis Results

The aim of the quantitative phase of this study was to numerically estimate the extent to which the quality of CCEs was directly affected by physicians' evaluative load and strain and indirectly mediated by time delays in CCE submission. To achieve this aim, a path analysis examined the hypothesized causal relationships (Figure 3-2) among two demographic variables (i.e., Department and Gender) and five evaluative variables (i.e., Load, Strain, Delay, Diff, and QOF). The results of the statistical assumptions necessary to conduct the path analysis will be presented first, followed by a summary of model fit. Results are organized by research question and are presented alongside figures that demonstrate how the path model specifically addressed each respective research question. As previously mentioned, presentation of the path analysis results will consist of the standardized parameter estimates, as such estimates remove all individual scaling information pertinent to each variable and adjust variables to have the same standard deviation. Such an adjustment allows for easy interpretation of the data regardless of the original classification of the variable (e.g., continuous or categorical).

Statistical Assumptions

Prior to conducting the path analysis (Figure 3-2) using *LISREL* (Scientific Software International, 2016, Version 9.2), the data were thoroughly examined to ensure that the assumptions of the statistical analysis had been met. In the instance where a particular assumption was not met, the literature was consulted to determine the severity of the violation and whether it was appropriate to continue with the analysis.

Conducting a path analysis requires an examination of thirteen statistical assumptions. Eight of these assumptions pertain specifically to multiple regression analyses and must be examined prior to evaluating the remaining five assumptions that pertain specifically to path analysis. Three of the assumptions of multiple regression are concerned

with the independence of the independent variables and are factors of the research design. The five remaining assumptions of multiple regression address the normality of the data distribution, linearity (i.e., the requisite that the relationship between the independent variables and the dependent variable is linear), and homoscedasticity (i.e., the requisite that the variance of the residuals across all levels of the independent variables is constant).

The normality of the data distribution was assessed through visual inspection of both histograms and normal probability plots of the five evaluative variables and calculation of several descriptive statistics (i.e., skewness and kurtosis). Preliminary examination revealed an approximately normal distribution with appropriate descriptive statistics for Strain and QOF, while Load, Delay, and Diff all displayed a strong positive skew and leptikurtosis (Table 4-3). To correct for these departures from normality, a logarithmic (ln) transformation was conducted for the latter three variables using SPSS (IBM Analytics, 2016, Version 24). Descriptive statistics, histograms, and normal probability plots for the transformed variables revealed that the log transformation was successful in eliminating problems of skewness and kurtosis for each variable. The resulting variables (ln Load, ln Delay, and ln Diff) were used in all subsequent data analyses directly related to the path analysis. To achieve normalization of the data, the log transformations substantially changed the scale of the variables, reduced the standard deviations, and resulted in a desirable platikurtotic increase of the variables. The descriptive statistics for all five evaluative variables, including the pre- and post-transformations of Load, Delay, and Diff are displayed in Table 4-3. It was not appropriate to test the normality of the data distribution for the demographic variables (i.e., Department and Gender), as the variables were categorical in nature.

Table 4-3. Descriptive statistics of the five evaluative variables

Variable	Mean	SD	Skewness[a]	Kurtosis[b]
Strain	50.14	18.58	-0.28	-0.29
Load	21.34	16.68	1.67	2.92
ln Load	2.79	0.74	0.10	-0.76
Delay	8.30	10.69	2.77	9.37
ln Delay	1.41	1.32	-0.41	-0.29
Diff	3.03	1.24	1.34	3.26
ln Diff	3.18	0.67	-0.24	-0.22
QOF	2.48	0.90	-0.70	0.61

Note: Strain = Evaluative Strain; Load = Evaluative Load; ln Load = Transformed Evaluative Load; Delay = Time Delay; ln Delay = Transformed Time Delay; Diff = Score Differentiation; ln Diff = Transformed Score Differentiation; QOF = Quality of Feedback
[a]SE = 0.3; [b]SE = 0.5

Linearity, or the assumption that the relationship between the independent variables and the dependent variable is linear, was examined through visual inspection of bivariate scatterplots of the variables of interest. The scatterplots showing the linearity between ln Load and ln Diff suggested a strong linear relationship; other indications of linearity were present between ln Load and ln Delay; ln Diff and ln Delay; ln Strain and ln Diff; and Strain and ln Delay. Finally, the assumption of homoscedasticity, or the requisite that the variance of the residuals be constant across all values of the independent variables, was assessed through an examination of residual plots. Visual examination of both the residual plots for the outcome variables ln Diff and QOF provided unclear evidence as to whether the assumption of homoscedasticity was violated. However, modest violations of the assumptions of linearity and homoscedasticity do not invalidate the results of the regression analysis.[85,90] As such, the assumptions for the regression analysis were considered satisfactorily met.

Once the eight assumptions of multiple regression were met, it was necessary to examine the five additional assumptions unique to path analysis. Path analysis assumes (1) the proposed model is an accurate representation of the actual causal sequence; (2) all variables considered to be direct causes of each endogenous variable are included in that

variable's structural equation; (3) the model contains no reciprocal causation (i.e., the model contains only one-way causal flow); (4) the variables display additive, causal, and linear relationships; and (5) all exogenous variables contained within the model are measured without error.[85] The first four assumptions of path analysis are directly related to the model specification, and consequently are not able to be examined through statistical methods. Rather, assumptions concerning model fit are best examined through a subjective analysis of the model's "credibility, reasonableness, and utility."[85] The path model unique to this study was formed through the researcher's experiences and observations with the variables, logic/intuition, and findings presented in the literature that support the existence of the hypothesized causal relationships among the variables. The combination of these influences implies *credibility*, and lends itself to the *reasonableness* of the results being interpreted within the context of this literature. Finally, the model's *utility* is evidenced by its relevance to medical academia, as was previously described in Chapter 1. The fifth assumption of path analysis is associated with data collection. The exogenous variables in this model (i.e., Department and Gender) were collected using the study's electronic survey and are believed to be honest representations of the participants' demographic characteristics.

Path Analysis Findings

The path model used was fully saturated ($df = 0$) and produced perfect goodness of fit ($\chi^2 = 0$). Table 4-4 demonstrates the correlation matrix, means, and standard deviations for the two demographic variables (dummy-coded into a total of seven variables) and five evaluative variables included in the model. The squared multiple correlations (R^2) of the structural equations (i.e., the percentage of variance of the endogenous variables explained by the model), are also presented in Table 4-4. The independent variables accounted for 28% of the variability in evaluative load. Likewise, the model accounted for 19%, 36%, 88%,

and 15% of the variance in evaluative strain, time delay, differentiation, and quality of feedback, respectively.

Table 4-4. Correlation matrix of the exogenous and endogenous variables

	Load	Strain	Delay	Diff	QOF	FM	GS	OB/GYN	PE	PY	Gender
Load	1.00										
Strain	0.07	1.00									
Delay	-0.02	0.35	1.00								
Diff	0.80	0.27	0.47	1.00							
QOF	-0.07	-0.09	-0.04	-0.09	1.00						
FM	-0.04	-0.06	-0.11	-0.11	0.18	1.00					
GS	0.10	-0.01	0.19	0.15	-0.19	-0.04	1.00				
OB/GYN	0.34	0.38	0.20	0.37	-0.20	-0.03	-0.10	1.00			
PE	0.08	-0.16	0.13	0.12	0.01	-0.08	-0.28	-0.20	1.00		
PS	0.18	0.00	-0.05	0.19	-0.03	-0.03	-0.10	-0.07	-0.22	1.00	
Gender	-0.08	0.22	0.38	0.11	0.02	-0.09	-0.27	0.20	0.25	-0.10	1.00
Mean	2.79	50.14	1.41	3.18	2.48	-	-	-	-	-	-
SD	0.74	18.58	1.32	0.67	0.90	-	-	-	-	-	-
R^2	0.28	0.19	0.36	0.88	0.15						

Note: Load = Evaluative Load; Strain = Evaluative Strain; Delay = Time Delay; Diff = Score Differentiation; QOF = Quality of Feedback; FM = Family Medicine; GS = General Surgery; OB/GYN = Obstetrics and Gynecology; PE = Pediatrics; PY = Psychiatry; Gender = Female.

Examination of the ψ matrix (abbreviated in Table 4-5) revealed the error variance estimates, standard errors (SE), and t-values of the residuals. Importantly, the matrix indicated no correlation between Diff and QOF, net the effects of other variables in the model (β = -0.01, SE = 0.03, $p > 0.05$), meaning that Diff and QOF measured different evaluation behaviors.

Table 4-5. Estimated variance of the residuals

Variable	Load	Strain	Delay	Diff	QOF
Load	0.72				
Strain		0.81			
Delay			0.64		
Diff				0.12	
QOF				-0.01	0.85

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$
Note: Load = Evaluative Load; Strain = Evaluative Strain; Delay = Time Delay; Diff = Score Differentiation; QOF = Quality of Feedback.

Demographic Variables

The rationale for including the demographic variables (i.e., Department and Gender) in the model was threefold. Firstly, the demographic variables provide a familiar context within which to interpret the results of the analysis. One's medical department affiliation and gender are often presented in literature pertaining to medical faculty. Including these variables in the model helps to make the results of the study more relatable/applicable to the end users of this work. Secondly, the inclusion of these variables permitted an understanding of how one's departmental culture/climate may influence one's evaluative responsibilities. As little literature exists on physician evaluative load and strain, theoretically one's medical department affiliation and gender could be influential determinants of the evaluative variables. As such, it was deemed necessary to understand which, if any, demographic factors may ultimately affect the quality of CCEs. Finally, inclusion of these variables allowed the model to control for medical department affiliation

and gender while interpreting the results of the path model that pertain specifically to the evaluative variables of interest.

Despite the inclusion of these demographic variables in the model, the interpretation of the relative influences of the demographic variables on the evaluative variables was not the principal focus of the study; accordingly, results that pertain specifically to the demographic variables are presented separately from the discussion of the results related to each specific research question. Moreover, the discussion of the effects of the demographic variables on the evaluative variables details how one's evaluative responsibilities are affected by one's medical department affiliation and gender. A summary of the standardized direct, indirect, and total effects of the demographic variables on the evaluative variables is shown in Table 4-6. Unless otherwise stated, the significant effects referenced in the text refer to the total effects.

Table 4-6. Standardized effects of the demographic variables on the evaluative variables

Variables		FM	GS	OB/GYN	PE	PY	Gender
Load							
	Direct	0	0.22*	0.49***	0.35***	0.30**	-0.17
	Indirect	-	-	-	-	-	-
	Total	0	0.22*	0.49***	0.35***	0.30**	-0.17
Strain							
	Direct	-0.04	0.05	0.34**	-0.12	0.03	0.19
	Indirect	0	-0.01	-0.02	-0.01	-0.01	0.01
	Total	-0.04	0.04	0.32**	-0.13	0.02	0.19
Delay							
	Direct	-0.03	0.40***	0.19	0.28**	0.13	0.31**
	Indirect	-0.01	-0.03	0.01	-0.09	-0.04	0.08
	Total	-0.04	0.37***	0.20*	0.19	0.08	0.39***
Diff							
	Direct	-0.02	-0.02	0	0.02	0.07	-0.03
	Indirect	-0.02	0.36***	0.50***	0.36***	0.28**	0.06
	Total	-0.04	0.34***	0.50***	0.37***	0.34***	0.03
QOF							
	Direct	0.16	-0.32**	-0.34**	-0.2	-0.14	0.02
	Indirect	0	0.08	0.08	0.08	0.05	0.02
	Total	0.15	-0.24*	-0.26*	-0.13	-0.09	0.04

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Note: Load = Evaluative Load; Strain = Evaluative Strain; Delay = Time Delay; Diff = Score Differentiation; QOF = Quality of Feedback; FM = Family Medicine; GS = General Surgery; OB/GYN = Obstetrics and Gynecology; PE = Pediatrics; PY = Psychiatry; Gender = Female.

Medical Department. Findings revealed that medical department affiliation had a small but statistically significant effect on evaluative load for physicians from the departments of GS ($\beta = 0.22$, $p < 0.05$), OB/GYN ($\beta = 0.49$, $p < 0.001$), PE ($\beta = 0.35$, $p < 0.001$), and PY ($\beta = 0.30$, $p < 0.01$) when compared to those from IM, suggesting that physicians from these departments had more assigned evaluative tasks. Additionally, affiliation with OB/GYN had a positive and significant effect on evaluative strain, as compared to an affiliation with IM ($\beta = 0.32$, $p < 0.01$). Medical department affiliation also had a small and statistically significant effect on time delay for physicians from the departments of GS ($\beta = 0.37$, $p < 0.001$) and OB/GYN ($\beta = 0.20$, $p < 0.05$), and a small, significant direct effect on time delay for those from PE ($\beta = 0.28$, $p < 0.01$) when compared to physicians from the department of IM. These findings suggest that physicians from GS, OB/GYN, and PE have lengthier time delays than those from IM. Moreover, one's medical departmental affiliation was significantly correlated with the quality of one's closed-ended CCE items and open-ended comments. Departmental association had a small positive effect on score differentiation for physicians from GS ($\beta = 0.34$, $p < 0.001$), PE ($\beta = 0.37$, $p < 0.001$), and PY ($\beta = 0.34$, $p < 0.001$), and a medium yet significant effect for physicians from OB/GYN ($\beta = 0.50$, $p < 0.001$) in relation to those from IM. Stated otherwise, the data suggest that physicians from GS, PE, PY, and OB/GYN vary their scores more widely within and across learners than physicians from IM. However, a significant effect on differentiation among items did not equate to a significant effect on the quality of free response comments. Despite good utilization of the CCE scales on the closed-ended items, departmental affiliation with GS ($\beta = -0.32$, $p < 0.01$) and OB/GYN ($\beta = -0.34$, $p < 0.01$) had a negative effect on QOF scores, as compared to those produced by physicians from IM, suggesting that physicians from GS and OB/GYN delivered lower quality written feedback to clerks than those from IM.

Gender. Analysis of the data revealed that gender had very little effect on the endogenous variables. Gender had no effect on physicians' evaluative load, or on how male and female faculty perceived their evaluative load (i.e., their evaluative strain). Additionally, gender had no significant effect on the quality of closed- or open-ended items produced by faculty. Gender did have a direct and significant effect, however, on one's length of time delay in CCE submission. Identifying as female had a small, positive and significant effect on time delay ($\beta = 0.39$, $p < 0.001$) when compared to male faculty, suggesting that women had lengthier delays in CCE submission than males.

Evaluative Variables

The following section presents the results of the path analysis as they pertain to each research question. A summary of the standardized direct, indirect, and total effects coefficients generated for the endogenous variables on other endogenous variables is presented in Table 4-7. The specifics of these relationships (illustrated using Figures 4-1 through 4-6) will be discussed alongside the research question they address. Unless otherwise stated, the significant effects referenced in the text refer to the direct effects.

Table 4-7. Standardized effects of the evaluative variables on other evaluative variables

Variables		Load	Strain	Delay	Diff	QOF
Load						
	Direct	-	-	-	-	-
	Indirect	-	-	-	-	-
	Total	-	-	-	-	-
Strain						
	Direct	-0.03	-	-	-	-
	Indirect	-	-	-	-	-
	Total	-0.03	-	-	-	-
Delay						
	Direct	-0.16	0.27**	-	-	-
	Indirect	-0.01	-	-	-	-
	Total	-0.17	0.27**	-	-	-
Diff						
	Direct	0.79***	0.05	0.48***	-	-
	Indirect	-0.08	0.13**	-	-	-
	Total	0.70***	0.18**	0.48***	-	-
QOF						
	Direct	0.13	-0.05	0.14	-	-
	Indirect	-0.02	0.04	-	-	-
	Total	0.11	-0.01	0.14	-	-

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Note: Load = Evaluative Load; Strain = Evaluative Strain; Delay = Time Delay; Diff = Score Differentiation; QOF = Quality of Feedback.

Results for Research Question 1

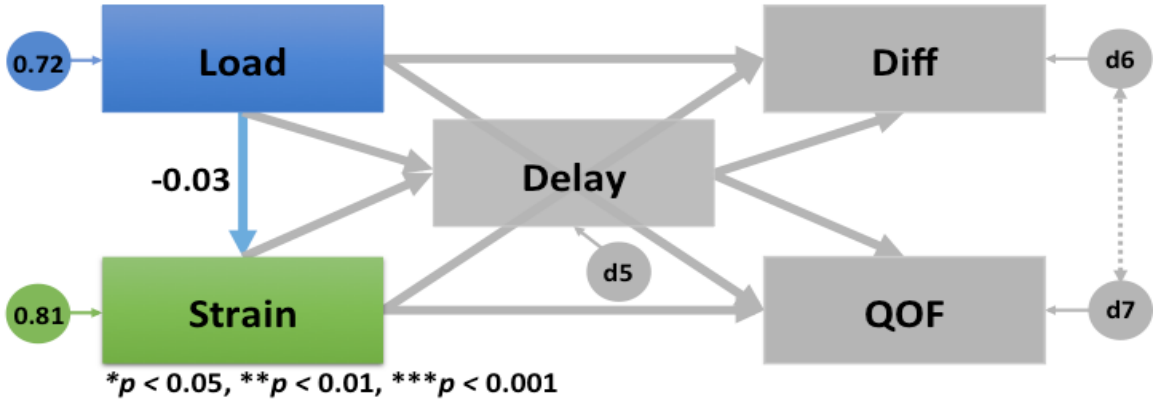


Figure 4-1. Results for Research Question 1
Note: The highlighted portion of the conceptual model addressed how the path analysis answered the research question, "How does physician evaluative load influence evaluative strain?"

Physicians' evaluative load had no direct effect on their evaluative strain (β = -0.03). Stated otherwise, physicians' subjective conceptualizations of the number of evaluations they completed and the perceived cognitive demands needed to complete the assigned evaluative tasks were not significantly influenced by the quantity of evaluations they were assigned.

Results for Research Question 2

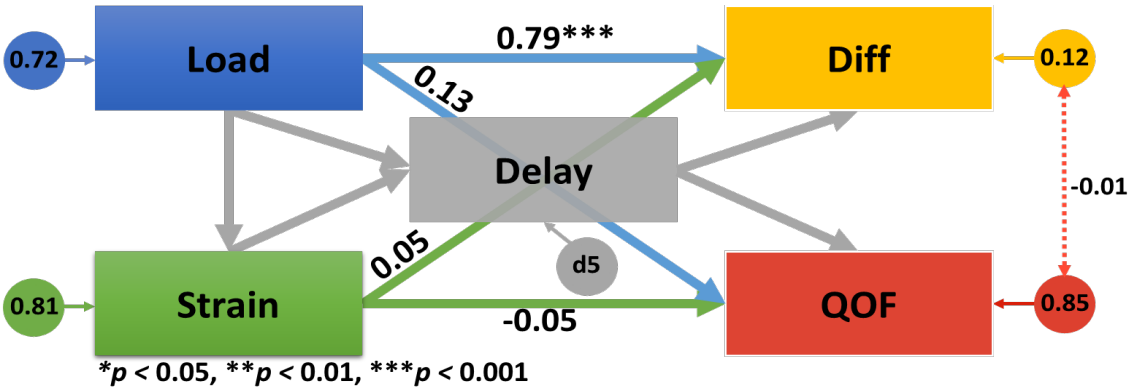


Figure 4-2. Results for Research Question 2
Note: The highlighted portion of the conceptual model addressed how the path analysis answered the research question, "How do evaluative load and evaluative strain directly affect the quality of CCEs, as measured by the degree of differentiation and the quality of feedback?"

Physicians' evaluative load had a positive and large significant effect on the degree of differentiation present among the closed-ended CCE items ($\beta = 0.79$, $p < 0.001$). This suggests that physicians with large evaluative assignments varied their scores more widely within and across learners than physicians with smaller evaluative responsibilities. Despite the significant influence of physicians' evaluative load on the degree of differentiation present among closed-ended items, evaluative load had no effect on the quality of feedback ($\beta = 0.13$) physicians gave to clerks. Similarly, the direct effect of physicians' evaluative strain on differentiation ($\beta = 0.05$) and QOF ($\beta = -0.05$) were non-significant, suggesting that one's perceptions of his or her evaluative load and one's conceptualizations of the cognitive demands needed to complete the task do not influence the quality of the closed-ended CCE items or the clarity of written feedback on the evaluations one produces.

Results for Research Question 3

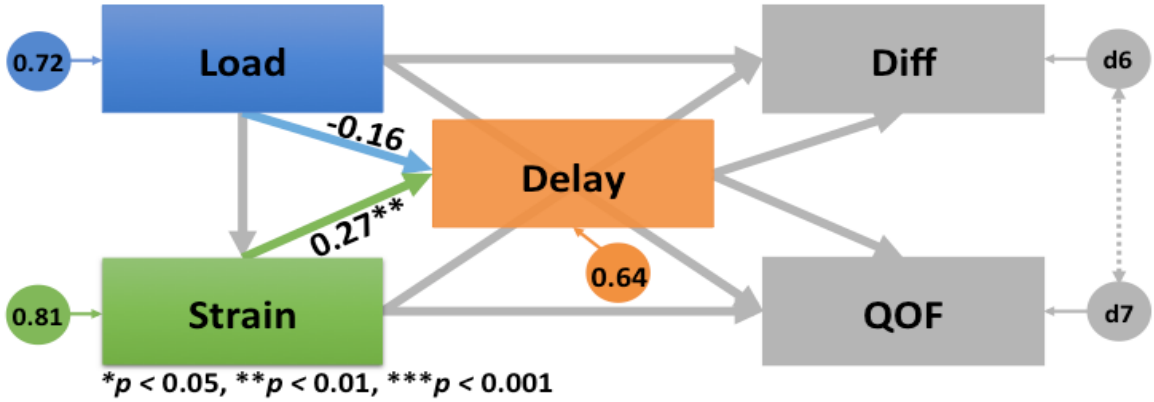


Figure 4-3. Results for Research Question 3
Note: The highlighted portion of the conceptual model addressed how the path analysis answered the research question, "To what degree are evaluative load and evaluative strain associated with the length of time delay between a clinician's final observation of (or encounter with) a clerk and the rating of that clerk's clinical performance?"

Physicians' evaluative load had no effect on time delay (β = -0.16). Stated otherwise, one's assigned number of CCEs did not influence the timeliness of his or her evaluation submissions. One's perceptions of the complexity of the evaluative task, however, did significantly influence the length of time delay in evaluation submissions (β = 0.27, p < 0.01). These results indicated that physicians with higher evaluative strain took longer to submit their CCEs.

Results for Research Question 4

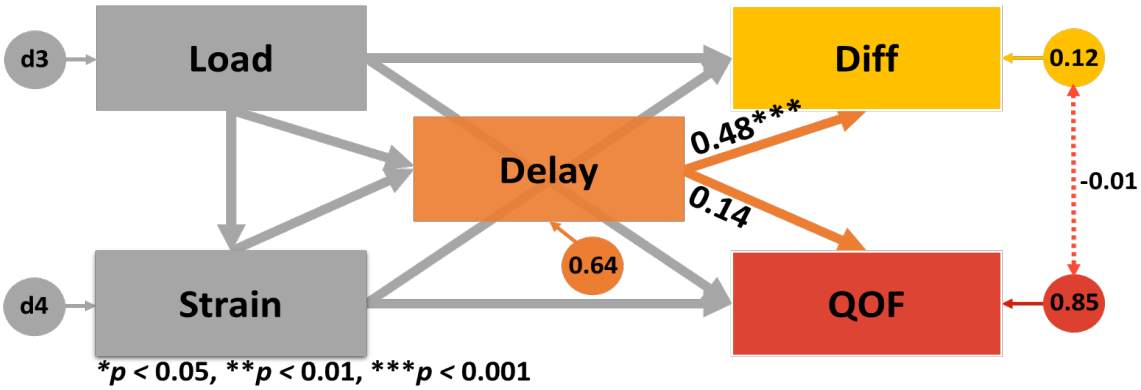


Figure 4-4. Results for Research Question 4
Note: The highlighted portion of the conceptual model addressed how the path analysis answered the research question, "To what extent do time delays directly influence the quality of CCEs?"

Time delay had a direct, positive, and significant effect on degree of differentiation ($\beta = 0.48$, $p < 0.001$). This suggests that physicians with lengthy time delays in CCE submission vary their scores on the closed-ended CCE items within and across learners more than physicians with shorter time delays. Despite the significant positive effect of Delay on Diff, no significant relationship was found for the effects of Delay on QOF ($\beta = 0.14$), suggesting that the timeliness of one's CCE submission did not influence the clarity of one's written feedback.

Results for Research Question 5

As was done in Chapter 3, the results of the path analysis related to the fifth research question are presented using different figures. Figure 4-5 presents the indirect effects of evaluative load on evaluative quality, while Figure 4-6 discusses the indirect effects of evaluative strain on evaluative quality.

Results for Research Question 5A

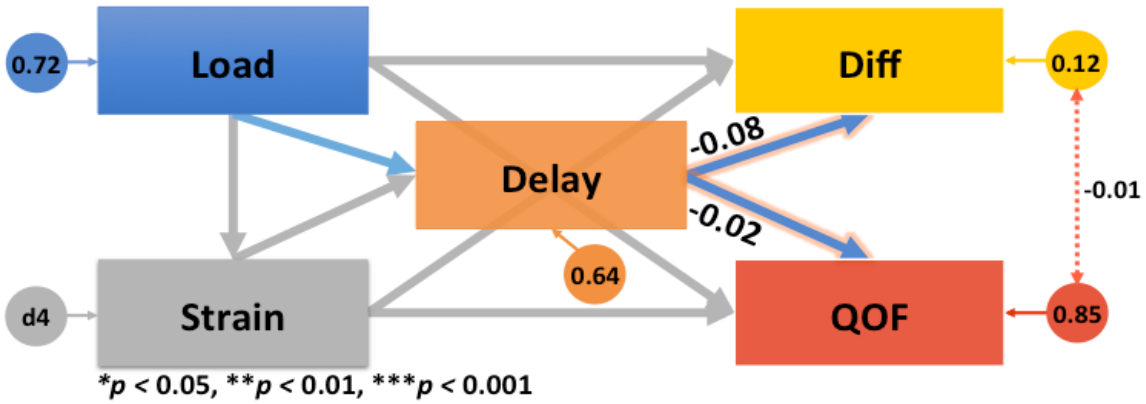


Figure 4-5. Results for Research Question 5A
Note: The highlighted portion of the conceptual model addressed how the path analysis answered the research question, "To what extent do time delays mediate the influences of evaluative load on the quality of CCEs?" The two-toned arrows represent the indirect effects of Load on Diff and QOF, respectively.

Physicians' evaluative load had no significant indirect effect on either the degree of differentiation among the closed-ended evaluation items (β = -0.08) or the quality of feedback (β = -0.02). These findings suggested that the number of CCEs a physician was asked to complete did not significantly influence either the quality of the closed-ended CCE items or the open-ended questions present on the evaluations when mediated by delays in CCE submissions.

Results for Research Question 5B

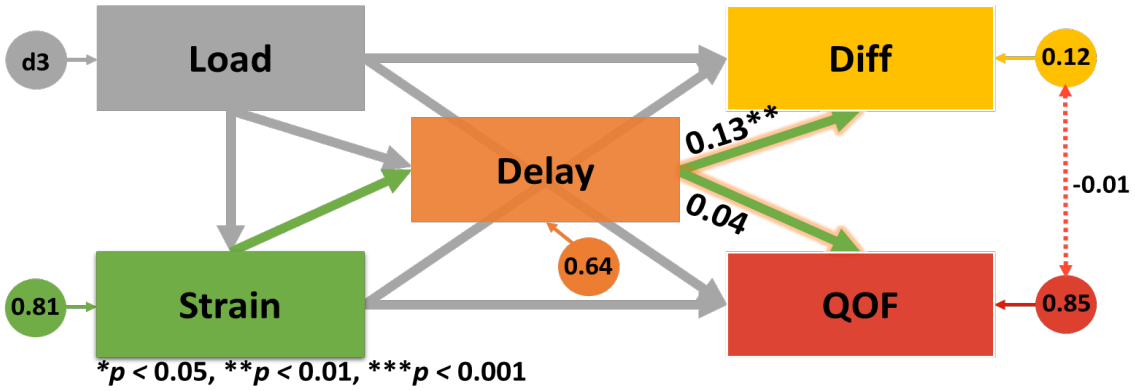


Figure 4-6. Results for Research Question 5B
Note: The highlighted portion of the conceptual model addressed how the path analysis answered the research question, "To what extent do time delays mediate the influences of evaluative strain on the quality of CCEs?" The two-toned arrows represent the indirect effects of Strain on Diff and QOF, respectively.

In contrast to the non-significant indirect effect of evaluative load on CCE quality, physicians' evaluative strain had a positive and significant indirect effect on the degree of differentiation among the closed-ended evaluation items ($\beta = 0.13$, $p < 0.01$). From this, it was interpreted that physicians' perceptions of their evaluative tasks significantly contributed to the amount of variation present in the closed-ended CCE items by influencing the length of time delay in CCE submission. In other words, physicians with high evaluative strain are more likely to have more variation among the closed-ended items they produce when they also have lengthy time delays in evaluation submission. Strain had no indirect effect on QOF ($\beta = 0.04$), suggesting that one's perceptions of their evaluative tasks do not influence the clarity of written feedback one produces, regardless of one's length of time delay in CCE submissions.

PART II: QUALITATIVE RESULTS

Designed to complement and expand upon the quantitative findings, the aim of the qualitative phase of this mixed-methods explanatory study was to explore how physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students. To facilitate an understanding of how physicians think about the evaluations they complete, the overarching research question was deconstructed into four sub-questions:

- Evaluative Utility: How do physicians interpret the 'utility' of CCEs?
- Evaluative Quality: How do physicians perceive the 'quality' of CCEs?
- Evaluative Cost: How do physicians conceptualize the 'cost' of completing CCEs?
- Evaluative Practicability: How do physicians view the 'practicability' of completing CCEs?

This chapter presents the qualitative findings of the study. After describing the participants for phase II, the themes that emerged following transcription and coding of the one-on-one semi-structured interviews will be presented by sub-research question.

Study Participants

A total of eight participants were purposefully recruited to participate in phase II of the study. Efforts were made to recruit participants with variable background demographics and evaluative responsibilities from each medical department surveyed. Five of the eight participants were male. Three participants cited departmental affiliation with IM, but five of the six medical departments surveyed had a least one physician agree to be interviewed. No participants from PY consented to the interview. Two participants identified as residents, three as assistant professor, one as associate professor, and the remaining two identified as full professor. When asked about the extent to which one feels overwhelmed by the number of CCEs he or she is asked to complete, five of the eight participants reported feeling "sometimes overwhelmed." Additionally, two of the eight

reported feeling "very often overwhelmed." Participants also reported having to regularly complete resident evaluations (8 of 8), professional society surveys/questionnaires (6 of 8), and institution-specific surveys/questionnaires (8 of 8). A full description of the participant demographics is provided in Table 4-8.

Table 4-8. Description of phase II participants

Variable		*n* (%) [$n_{Total} = 8$]
Gender		
	Male	5 (62.5%)
	Female	3 (37.5%)
Medical Department		
	Family Medicine	1 (12.5%)
	General Surgery	1 (12.5%)
	Internal Medicine	3 (37.5%)
	OB/GYN	1 (12.5%)
	Pediatrics	2 (25.0%)
	Psychiatry	0 (0.0%)
Academic Rank		
	Resident	2 (25.0%)
	Fellow	0 (0.0%)
	Instructor/Lecturer	0 (0.0%)
	Assistant Professor	3 (37.5%)
	Associate Professor	1 (12.5%)
	Full Professor	2 (25.0%)
Extent to which one feels overwhelmed by one's evaluative load		
	I never feel overwhelmed.	1 (12.5%)
	I sometimes feel overwhelmed.	5 (62.5%)
	I often feel overwhelmed.	0 (0.0%)
	I very often feel overwhelmed.	2 (25.0%)
Other surveys/evaluations regularly completed		
	Resident evaluations	8 (100%)
	Professional society surveys	6 (75.0%)
	Institution-related surveys	8 (100%)

Thematic Analysis Findings

Across the four qualitative sub-research questions, eight themes emerged from participants' transcribed interviews. To facilitate an understanding of the qualitative findings, themes are presented alongside the sub-research question they address. A broad summary of the themes that emerged for each sub-research question are presented in the text, while thematic definitions and representative quotes tagged with an alphanumeric identifier to represent the authoring participant are displayed in summary tables at the end of each section.

Themes in 'Evaluative Utility'

The first sub-research question inquired into participants' interpretations of the usefulness or utility of CCEs. Upon iteratively reviewing the transcribed interviews, three themes emerged. The first two themes revealed several distinct and disparate opinions on the usefulness of the evaluative instrument itself. The last theme highlighted participants' contrasting perceptions of the utility of the feedback provided to students, both verbal and written. A summary of each 'Evaluative Utility' theme will follow, with a more detailed presentation of the three themes, their descriptions, and representative quotes included in Table 4-9.

The first theme, "The Evaluative Instrument Is a Useful Template or Model for Behaviors That Should Be Observed," revealed some participants believed the evaluative instrument was an effective measure for assessing medical students' clinical performances. In particular, this utility referred to the instrument's use as a precedent for recalling which behaviors and/or skills are important to evaluate trainees. Moreover, the utilization of the evaluative instrument lessened the complexity of evaluating medical students by providing evaluators with objective assessment measures needed to satisfy not only personal measures of student success, but also course- and institution- specific learning objectives.

The second theme, "The Evaluative Instrument is an Ineffective Measure for Addressing Relevant Aspects of Students' Clinical Performances," presented contrasting opinions to those highlighted in the first theme. The majority of participants felt the evaluative instrument was an unsuitable measure of students' clinical progress, and thus of little utility to students in terms of professional development. For some, this unsuitability lay in the specific wording of the questions, which prevented evaluators from focusing on global behaviors perceived to be more relevant for developing physicians (e.g., student enthusiasm, work ethic, efficiency as a team member). For others, the descriptive anchors included on some of the closed-ended CCE items (e.g., 'Manager,' or 'Educator') were inappropriate measures of third-year medical students' clinical knowledge and behavior and were instead more representative of interns' educational level. Finally, the instrument was found to be an ineffective measure of students' clinical knowledge as it did not provide evaluators with an opportunity to describe how students had progressed during their rotation. While the evaluation includes an opportunity to suggest 'ways to improve,' the instrument was found to have little utility in commenting on student growth and professional maturation during the rotation.

The third and final theme in 'Evaluative Utility,' "Verbal Feedback Is Useful, but Written Feedback Is Less Useful," accentuated the stark difference in participants' perceptions of the utility of feedback provided to students during and after their clerkship rotations. Unanimously, all participants emphasized the usefulness of providing students with verbal feedback. For some, this utility lay in the timing of feedback delivery, as providing students with verbal feedback throughout the rotation gave students an immediate opportunity to improve or alter their clinical performance. For the majority of participants, however, the real utility of verbal feedback lay in the opportunity to provide students with truthful and uncensored criticisms of their clinical conduct. Speaking with

students one-on-one, evaluators felt free to give students advice and suggestions for improvement without fearing that their words could have a potentially harmful effect on students' futures.

Undocumented, the delivery of verbal feedback lay in stark contrast to the formal, written feedback found on students' CCEs. Recognizing that anything written on students' CCEs could be included in their Medical Student Performance Evaluations (MSPEs) or 'Dean's Letters,' many participants admitted to censoring the written and scaled-data included on students' CCEs. A narrative summary of students' educational and personal background, successes within foundational courses, and clinical knowledge, MSPEs are considered an influential component of students' residency applications. Though participants wanted to help students improve as clinicians, they did not want their written assessment of students' clinical abilities to impact their chances for residency matching. In particular, this evaluative censorship and grade inflation applied specifically to students with minor competency lapses. Despite the desire to not harm students with written evaluations, egregious deficits in clinical knowledge or lapses in judgment were reported on students' evaluations. Though concern for the students' residency matching was the most commonly cited reason for inflating students' ratings or censoring written feedback, participants also reported inflating students' ratings in response to pressure from students wanting to receiver higher grades. Finally, wanting to produce ratings that resemble those given by other evaluators within one's department and feeling "fatigued" by one's evaluative demands and the length of the instrument were also reported as reasons for producing inflated ratings on students' CCEs.

Overall, the desire to provide students with meaningful feedback that could be used to improve their performance, but not hurt their chances for residency matching, created an internal conflict for participants. As a consequence of this conflict, participants' perceptions

of the utility of feedback were based on the type of feedback given. Verbal feedback was perceived to be of great utility to the students, as students receive timely and actionable feedback that is representative of their true clinical knowledge, but not detrimental to residency matching. Contrastingly, written feedback provided on students CCEs was considered to be of little utility to the students in terms of professional development, as students were not provided with an accurate portrayal of their current clinical abilities and were unable to address weaknesses in their performances.

Table 4-9. Themes, descriptions, and representative quotes for 'Evaluative Utility'

Theme	Theme Description	Representative Quotes
The Evaluative Instrument Is a Useful Template or Model for Behaviors That Should Be Observed	The evaluative instrument lessened the complexity of the evaluation by serving as a precedent for recalling which behaviors and/or skills are important to evaluate on students' clerkship rotations.	"The structuring of the questions reminds me about what I should be looking for; what behaviors are important to note. It's easy to evaluate things that I think are important, but there are other things I am supposed to evaluate, too." (P4) "I think it's a model; I think it's a model of the things that we should be assessing; so it's a reminder of, 'Oh, yes, we should be assessing this,' and 'Oh, yes, we should be assessing that.' It's a model of the important stuff that ties back to our institutional learning objectives, our course learning objectives. Those things are built into our evaluation forms, and they're useful from that perspective." (P3)
The Evaluative Instrument Is an Ineffective Measure for Addressing Relevant Aspects of Students' Clinical Performances	The wording and specificity of questions included on the instrument, inappropriate anchors on the closed-ended items, and no opportunity to describe student progress throughout the rotation made the evaluative instrument an ill-suited measure of students' clinical conduct.	"I'm more of a global kind of assessing person. Whether I think the student 'gets it;' whether they're invested in their own education; whether they work hard; whether they're a good team member; whether they're excited to learn; how they interact with patients; those kinds of things – some of the ways the questions are structured – it just doesn't get to those things that I think are important in terms of developing as physicians." (P8) "The anchors on that final question – manager, reporter, whatever; I think the bar is too high. It might work great for an intern but it's not for a third-year student." (P6) "There's nothing in the evaluation to reflect how much they've improved over the course of the rotation. How teachable are they? How much can I take someone who doesn't know what they're doing and get them to learn what they're doing? I think that's more important." (P6)

Table 4-9 Continues

Table 4-9 (Continued). Themes, descriptions, and representative quotes for 'Evaluative Utility'

Theme	Theme Description	Representative Quotes
Verbal Feedback Is Useful, but Written Feedback Is Less Useful	Verbal feedback allows evaluators to provide students with timely, actionable, and personal feedback based on their true behaviors and observed performances, while written feedback is censored, inflated, and not representative of students' current clinical abilities.	"I give students verbal feedback all along the way. I try very hard to give them feedback in the moment; tell them about how they interact with patients and how they communicate, so they can work on these skills while they're on rotation. I like to give students personal feedback, too. And to be a little more causal – like, 'Wow, it was super fun having you on my service.' They appreciate that, but it's not something I would write." (P8) "When I was a medical student, I think the feedback that I found most helpful – and I think it's still true with the people I evaluate – was the one-on-one, sit-down feedback; the kind that's closed-doors, no one else is around. Some of the stuff is good, some of the stuff is bad, but you try to give them a nice game plan moving forward; and that's what I personally feel is the most helpful to me, and I feel like that's probably the most helpful thing that I do as an educator when trying to give them feedback. I think this kind of feedback is a little bit more interactive and meaningful than what's written on the evaluation. It's also more truthful; there's no negative repercussion for the students if the feedback is more critical than congratulatory, as it's delivered orally and kept between us." (P1) "The evaluations are good for writing good things down. Nice comments are essentially copied and pasted into recommendation letters for residency. That's really the purpose of the evaluations, to provide written feedback for Dean's Letters." (P7) "I write specific feedback knowing that it will go in their Dean's Letter. People read these Dean's Letters and if there's anything remotely negative, they ding resident candidates for that. I'm hesitant to write anything not positive. I'm honest, but I'm really careful that if, they're just kind of average, I don't write anything negative. If I wrote, 'Just an average medical student,' that would be negative; so I'm really careful with what I write." (P8)

Table 4-9 Continues

Table 4-9 (Continued). Themes, descriptions, and representative quotes for 'Evaluative Utility'

Theme	Theme Description	Representative Quotes
Verbal Feedback Is Useful, but Written Feedback Is Less Useful (*cont.*)	Verbal feedback allows evaluators to provide students with timely, actionable, and personal feedback based on their true behaviors and observed performances, while written feedback is censored, inflated, and not representative of students' current clinical abilities.	"I am scared of writing formative feedback; not necessarily even writing something negative, more just like, 'You could really work more on this, more on that;' I'm scared to put that in there, because, I mean, I want the student to work on it, but I don't want them not to get the residency they want because of that; I just want them to work on it, you know? What I've been doing, rather, is telling them in person, instead of putting it in writing; like 'Hey, you really need to work on this and here is what you could do to improve,' but I don't write that. I don't know if it's true, but if I write something negative, does that follow somebody? Does that go on a transcript? That's why I hold back a bit from the written evaluation; I just don't want to hurt them. Unless they have egregious errors. If they do, I report that, because they're potentially a danger, but otherwise? I hold back. I just don't' want to hurt them." (P7) "It's the whole 'Lake Wobegon Effect.' Everybody is above average, so if you see anything that's *bad*, then you think they're just terrible. That's how it is; it's the truth." (P4) "Grade inflation runs rampant, through the department and the institution." (P3) "I used to rate students using the descriptive anchors, but then I had students calling me and emailing me, asking why I gave them a '6,' and I would say, 'Well, let's look at the evaluation; what does it say a student at '6' should be doing? Did you do more than that? No? Then that's why you got a six.' But then I realized everyone else gets '8s' or '9s' even though they should be getting '4s' and '5s.' The evaluators aren't reading the little anchor stem at all because, again, survey fatigue. You're answering these long surveys, 27 questions, and you don't read all the questions – you just 'click click click.' So I started changing the way I grade, because I didn't want to be that 'odd duck' who was giving everyone low scores, even though they were more honest, and because I didn't want students calling me asking why they were graded a certain way. So now everything is inflated, and the scores are useless." (P6)

Themes in 'Evaluative Quality'

The second sub-research question investigated participants' perceptions of the 'quality' of CCEs they complete for third-year medical students. Repeated reading of participants' transcribed interviews resulted in the emergence of two themes. Interestingly, participants did not express their perceptions of the quality of the evaluations they produce as much as they shared their opinions on the factors believed to influence evaluative quality. Consequently, the two 'Evaluative Quality' themes summarize participants' perceptions of the determinants of evaluative quality, rather than their opinions on the quality of the evaluations themselves. A discussion of each theme will follow. As was the case for 'Evaluative Utility,' a comprehensive summary of the themes, their descriptions, and representative quotes are presented in Table 4-10.

The first theme in 'Evaluative Quality,' "The Breadth of the Student-Evaluator Interaction," emphasized the perceived influence of the duration and relevance of student-evaluator interactions on the quality of CCEs. For many participants, the amount of time spent in contact with students during clinical rotations was perceived to have a sizeable influence on the quality of CCEs produced. Actual time spent in direct observation with students was abbreviated for most participants, ranging from no time to a few days. Complicating matters, student-evaluator interactions often occurred in group settings consisting of third and fourth year medical students, residents, fellows, and at least one attending physician. Daily changes to scheduling and rotation assignments further lessened the duration of time spent in contact with students. Brief contact with students was perceived as a negative indicator of evaluative quality, as participants felt unable to adequately acquaint themselves with students or observe their clinical conduct. Moreover, the limited amount of time evaluators were able to spend with students also increased the difficulty of recalling the specifics of students' performances and ultimately hindered

participants' abilities to provide students with meaningful feedback. In addition to the *duration* of the student-evaluator interaction, the *relevance* of the interaction was also perceived to influence the quality of CCEs produced. Interactions deemed relevant and meaningful (e.g., encounters in inpatient clinics or the operating room) were perceived to positively influence the quality of CCEs produced, as such interactions allowed the evaluator to assess student behaviors queried on the evaluative instrument. Conversely, interactions deemed irrelevant (e.g., those occurring within an office setting or outside of the clinical venue entirely, such as at a restaurant) were found to negatively influence evaluative quality, as such interactions occurred in environments that were not conducive for assessing students' clinical knowledge and aptitudes.

The second 'Evaluative Quality' theme, "The Evaluative Culture of the Department," revealed the perceived influence of one's departmental expectations on the quality of CCEs produced. For some participants, departmental allocation of evaluative responsibilities was perceived to have a great influence on the quality of evaluations produced. While the sharing of evaluative duties among several physicians within a medical department was perceived to positively influence evaluative quality, assigning a single physician to all medical student evaluations was reported to be a substantial detriment to evaluative quality. Coupled with abbreviated time spent in contact with students, such an evaluative allotment forced one participant to complete evaluations for students that he or she had not directly observed, resulting in evaluations based on hearsay. A second participant also admitted completing evaluations for unfamiliar students, but cited departmental/program hierarchy/expectations, rather than the departmental allocation of evaluations, for this behavior. To the participant, the extent of residents' involvement in teaching and mentoring medical students hindered the participant's (i.e., attending's) ability to interact with students at an appropriate level, and was thus perceived to be a negative indicator of CCE

quality. Finally, the departmental salience of CCEs was additionally believed to influence evaluative quality. Medical departments that emphasized assessment proficiency were believed to produce higher quality evaluations for all evaluative populations (e.g., clerks, residents, etc.), as their physicians were frequently educated on how to improve the quality of written feedback.

Table 4-10. Themes, descriptions, and representative quotes for 'Evaluative Quality'

Theme	Theme Description	Representative Quotes
The Breadth of the Student-Evaluator Interaction	The duration and relevance of the interaction between the student and the evaluator greatly influenced the quality of CCEs produced.	"Often when I'm asked to complete an evaluation, it's difficult because the time I've spent with that learner is abbreviated. We usually spend about one half day with the student in clinic, and it's not typically in a one-on-one setting. How can we give good quality evaluations when we aren't spending time with the students?" (P3) "When an attending physician shows up on an inpatient, teaching service, they'll have a senior resident, a couple of interns, a senior student, and a couple third year students. On any given day, that schedule changes; so it's really hard to form any sort of group dynamic, or even remember who you've worked with." (P4) "We rotate around a lot as faculty, and so I may see a student once or twice in a four-week rotation, and I'm supposed to decide how good they are based on one twenty-minute conversation that we had in the operating room for one case? What can I possibly say that will be accurate or of good quality? I don't remember which student did what; I don't even remember what I had for dinner two nights ago, so how am I supposed to remember what a medical student did 3.5 weeks ago?" (P6) "I think that it's not very relevant for the majority of what we do as faculty. I usually see medical students in the office with my patients, so I don't know how they do on a lot of performance aspects. 'Do they come in after hours?' I don't know; there's no after hours for the office. 'How well do they do on procedures?' I don't know; it's the office, not the operating room. I don't see them in a lot of the different environments that it's asking me about. We're not rounding on people every day, so much of it, for what I'm seeing them for, is irrelevant." (P6) "Sometimes the evaluations aren't even based on observed clinical behaviors. I take students out to lunch sometimes, and I think, 'You were a nice guy, a nice gal,' you know, 'I took you to lunch once and you were a nice person so I'll give you a good evaluation.' That's what happens. That's when you see the 'Good student, good fund of knowledge; this person should be a primary care doctor' comments." (P5)

Table 4-10 Continues

Table 4-10 (Continued). Themes, descriptions, and representative quotes for 'Evaluative Quality'

Theme	Theme Description	Representative Quotes
The Evaluative Culture of the Department	The allocation of evaluative responsibilities, the hierarchy of the department, and salience of evaluation proficiency all contributed to the quality of CCEs produced.	"I'm the only one responsible for medical student evaluations. Even though I may not have worked with them during the clinical period, I have all of their evaluations; so I'm left to write evaluations for people that I don't know about things that I haven't seen. If there was more sharing the responsibility of evaluating them, I would feel like we were doing them a better service." (P2) "When they're here, there's a layer between us. There's me, then there's the resident, then there's the medical student. I don't see them round on patients; so it's just hard to evaluate how well they did because I didn't see them do anything. I'm working with the word-of-mouth from the residents who did." (P6) "I am concerned that we are not doing a good enough job as a school in producing evaluations of quality; I'm not sure we have faculty who are good at assessment. That's a problem." (P3) "We have faculty development workshops. We've got trained facilitators on faculty development on how to conduct active teaching based on widely published frameworks. There are a lot of opportunities around. I think a lot of people would like it if they could do more assessment training. It seems to make them more confident, more assured in their abilities as an educator." (P1)

Themes in 'Evaluative Cost'

The third sub-research question explored participants' conceptualizations of the 'cost' of completing CCEs. Thorough review of participants' interview transcriptions resulted in the emergence of a single theme. Summarized below, this theme portrays participants' perceptions of the temporal cost, both actual and illusory, of completing such evaluations. A descriptive summary of the 'Evaluative Cost' theme, its description, and representative quotes are presented in Table 4-11.

When participants were asked about the 'cost' of completing CCEs, "time" was the unanimous response. "Evaluations are Temporally Costly" summarized the aspects of the evaluation that were considered to be both tangibly and perceptively temporally demanding. In regards to the evaluative instrument itself, participants felt that the number of items included on the evaluation, the amount of time it took to physically answer each closed-ended CCE item (i.e., the "clicking" of responses), and the navigability of the software platform used to complete the evaluation form all contributed to the temporal cost of completing CCEs. In fact, many participants perceived these evaluative aspects to be the most 'costly' component of assessing medical students, and felt that modest changes to the evaluative instrument (e.g., shortening the number of closed-ended items) would greatly reduce the 'cost' of completing CCEs. In addition to time required to complete the evaluative instrument, participants' also described evaluative cost in relation to the time necessitated by their evaluative load. The sheer number of evaluations a participant was assigned influenced his or her perceptions of evaluative cost, as the participant was expected to observe each student, remember each student and the details of the clinical performance, and complete the evaluative instrument for each student.

Unquestionably, the evaluation of medical students is a time demanding task for evaluators. The temporal cost of evaluating students, however, was not restricted to actual

time spent completing evaluations. Rather, one's 'cost' of completing evaluations was also influenced by one's perceptions of the time they spent completing evaluations. Regardless of the actual amount of time spent completing evaluations, participants believed the completion of evaluations to be a time consuming responsibility. Combined with the actual amount of time spent completing CCEs, participants' perceptions of the time required to complete these evaluations amplified the temporal cost associated with evaluating medical students.

Table 4-11. Themes, descriptions, and representative quotes for 'Evaluative Cost'

Theme	Theme Description	Representative Quotes
Evaluations are Temporally Costly	Aspects of the evaluative instrument, one's evaluative load, and one's perceptions of the time required to complete CCEs contributed to the temporal 'cost' of evaluating medical students.	"The student evaluations are ridiculously long. They're so tedious." (P7) "The frustrating part of the evaluations is just the time it takes to fill-in all of the objective bubbles. At some point in time, I think people just start 'click click clicking,' trying to get through all of the questions." (P5) "It's exhausting because the software isn't all that great. It's supposed to auto-scroll but it doesn't, so you're just waiting for the next question." (P6) "There are 27 questions and I have to try to figure out if each question really applies to each student. The evaluation is just too long. Is there any difference between question 4 and question 8? Can't you combine those down? That would make it so much easier, and faster." (P2) "The evaluations are incredibly time consuming. I want them to be useful, so I think about it and I reflect back on my interactions with the students. If I don't do it right away, it takes longer for me to come up with specific things to put in the evaluation that I think would be helpful to them; it just takes time." (P8) "If you do these things right, you just can't fire them off. It takes a lot of time to do them well, and I get behind, especially when I have a lot of evaluations to complete. But I just sacrifice that time, because the students will get more net benefit out of the evaluations if I do." (P1) "It's that initial, perceived amount of time to make sure you get all of your evaluations completed. The perceived time demand is significant." (P3) "The evaluations are long, don't get me wrong, but I also perceive them to be very time consuming." (P8)

Themes in 'Evaluative Practicability'

The fourth and final sub-research question inquired into participants' views of the 'practicability' or feasibility of completing CCEs. From participants' transcribed interviews, two themes emerged. Presented below, these themes demonstrate that participants' perceptions of the practicability of completing CCEs were influenced by the timing of evaluation delivery and evaluators' prioritization of administrative responsibilities. A comprehensive summary of the themes, their descriptions, and representative quotes are presented in Table 4-12.

For the participants, "The Timing of Evaluative Delivery" heavily influenced the practicability of CCEs. Nearing the end of a clerkship rotation, evaluators receive an electronic request to complete performance evaluations for specific students. The timing of these evaluation requests varies by department, with most requests arriving after the clerkship rotation has ended. Some departments, however, ask evaluators to complete CCEs prior to the official conclusion of the clerkship rotation period. This premature request for evaluations distressed a few participants, who felt pressured to complete evaluations before they had time to finish working with a student. Moreover, this haphazard dissemination of evaluation requests was exacerbated by participants' other responsibilities. In addition to their evaluations of third-year clerkship students, nearly all participants were responsible for evaluating fourth year medical students, graduate students (e.g., nurse practitioner students), residents, or fellows. Without a common 'rotation' calendar among departments, the timing of evaluation requests for all evaluative populations overwhelmed participants and decreased the practicability of completing evaluations. To one participant, the creation and implementation of a campus-wide rotation calendar is believed to be a way to increase evaluative practicability, as such a calendar would provide structured observation and assessment opportunities for evaluators across all departments and programs.

In addition to the perceived influence of evaluative delivery, one's "Prioritization of Administrative Responsibilities" also influenced one's perceptions of evaluative practicability. Though participants' repeatedly acknowledged the importance of providing students with feedback, the completion of CCEs was not a pressing priority for most participants. Lack of protected time to complete evaluations, competing evaluative responsibilities (e.g., other performance evaluations or societal/professional questionnaires), and the preferential completion of patient notes over student clinical assessments decreased the perceived practicability of completing CCEs.

Table 4-12. Themes, descriptions, and representative quotes for 'Evaluative Practicability'

Theme	Theme Description	Representative Quotes
The Timing of Evaluative Delivery	The timing of electronic requests to evaluate clerks and the lack of a coordinated, campus-wide rotation calendar decreased the feasibility of completing CCEs.	"I would like for the surveys to come out immediately after the students leave rotations, not before, because I put off doing them until they leave anyways and sometimes I forget that I need to do them. All the surveys do is stress me out when they come before the rotation ends." (P5) "Like everybody else, they have medical students, they have residents, they have nurse practitioner students; they may have different health professions on campus that they're evaluating. It's just a lot of evaluations. And that's only the educational ones. That doesn't count the surveys for all the hospital organizations and all the other stuff that's happening. It's just a lot to do." (P4) "We have six perfectly siloed calendars that suit individual department needs but do nothing for a coordinated teaching effort across the school. We need to create a coordinated, campus-wide rotation calendar. We could have dedicated time for orientation, for early rotation/direct observation of students in the clinical setting, and protected time for mid-rotation and end-of-rotation feedback. It would make evaluations easier and less burdensome." (P3)
Prioritization of Administrative Responsibilities	Lack of protected time to complete CCEs and competing evaluative responsibilities decreased the practicability of completing CCEs.	"I am sad that I do not have protected time to sit and think or reflect on students and be able to give them helpful and timely feedback. Sometimes time doesn't present itself until two months after the rotation has ended, and then I struggle to give accurate, and most importantly, helpful feedback to my medical students. If there was a way to build in protected time to evaluate students, that would be great." (P5) "And it's not so much that I don't have time to do them, it's that I have a certain amount of time to do my administrative responsibilities, and if I've got 60-70 patient notes to write, they generally take priority over my student evaluations; so my student evaluations linger behind all of the other stuff I have to do on a daily basis." (P2)

CHAPTER SUMMARY

The purpose of this two-phase, sequential explanatory mixed-methods study was twofold. The quantitative phase investigated the extent to which the quality of CCEs was directly affected by physicians' evaluative load and strain and indirectly mediated by time delays in clerkship evaluation submissions. A path analysis was used to estimate the magnitude of the hypothesized causal relationships among the five evaluative variables (i.e., Load, Strain, Delay, Diff, and QOF) and two demographic variables (i.e., Department and Gender). As the findings pertain to the research questions, the path analysis revealed:

- One's evaluative load did not directly influence one's perceptions of evaluative quantity or one's conceptualization of the mental demands required to complete the assigned evaluative tasks (Research Question 1).
- Evaluative load positively, significantly, and directly affected the degree of score differentiation, but did not affect the quality of written feedback. Evaluative strain influenced neither the degree of score differentiation nor quality of feedback directly (Research Question 2).
- Evaluative load did not significantly influence the length of time delay in clerkship evaluation submission, but evaluative strain did, such that physicians with higher evaluative strain took longer to submit CCEs (Research Question 3).
- The length of time delay in clerkship evaluation submission directly and significantly influenced the degree of score differentiation on CCEs, such that physicians with lengthy time delays in CCE submission varied their scores on the closed-ended items within and across learners more than physicians with shorter time delays. Time delay had no influence on the quality of written feedback (Research Question 4).
- Evaluative load had no indirect effect on either the degree of score differentiation or the quality of written feedback when mediated by time delays. Evaluative strain, however, did

significantly influence the degree of score differentiation when mediated by time delays. In other words, physicians with high evaluative strain were more likely to have more variation among the closed-ended CCE items they produced when they also had lengthy time delays in evaluation submission. Evaluative strain had no such indirect effect on the quality of written feedback (Research Question 5).

The qualitative phase expanded upon the results of the quantitative phase by exploring how physicians perceive the utility, quality, cost, and practicability of CCEs they complete for third-year medical students. A total of eight themes emerged from participants' transcribed interviews that promoted an understanding of participants' perceptions of CCEs and may be used to help explain the quantitative findings. Participants shared their perceptions of the 'utility' of the evaluative instrument, finding it to be either (1) a useful model for evaluating medical students' clinical conduct, or (2) an ineffective measure for assessing relevant aspects of students' clinical performances. Participants also commented on the utility of feedback provided to students, noting that (3) verbal feedback is useful, while written feedback is less valuable. In regards to the 'quality' of the evaluations, participants believed the (4) breadth of the student-faculty interaction and (5) the evaluative culture of the department to be influential determinants of CCE quality. Participants' perceptions of the 'cost' of completing CCEs were unanimous and believed to be (6) entirely temporal. Finally, the 'practicability' of completing CCEs was perceived to be influenced by (7) the timing of evaluative delivery and (8) one's prioritization of administrative responsibilities.

CHAPTER FIVE

Discussion and Conclusions

Introduction

Integration and Interpretation of the Findings

Limitations

Suggestions for Future Research

Conclusions

INTRODUCTION

Within undergraduate and graduate medical education, performance evaluations of medical trainees remain the foremost means of assessing one's clinical prowess and behavioral conduct.[4,5] Recently, efforts to improve the rigor of these subjective evaluations have included an increase in the number of evaluations required for each trainee,[6] though the consequences of implementing higher assessment demands on evaluating physicians remains under-investigated. Accordingly, this study sought (1) to investigate the extent to which the quality of clinical clerkship evaluations (CCEs) was directly influenced by physicians' evaluative responsibilities and indirectly mediated by delays in evaluation submissions; and (2) to understand how physicians perceive the utility, quality, cost, and practicability of the CCEs they complete for third-year medical students.

A two-phase, sequential explanatory mixed-methods approach was utilized to achieve the study's central aims. The first, quantitative phase permitted the calculation of the five evaluative variables (e.g., evaluative load, evaluative strain, time delays, differentiation, and quality of feedback) and three demographic variables (i.e., gender, medical department affiliation, and academic rank), and culminated in a path analysis used to estimate the magnitude of the hypothesized, causal relationships among the first seven variables. The analysis resulted in several key findings: First, physician evaluative load did not influence physician evaluative strain; second, the degree of score differentiation on the closed-ended CCE items (i.e., the quality of the closed-ended items) was directly influenced by both physician evaluative load and the length of time delay in evaluation submissions, and was indirectly influenced by physician evaluative strain when mediated by time delay; third, evaluative strain had a direct effect on the length of time delay; and finally, neither evaluative load, nor evaluative strain, nor the duration of time delay in evaluation submissions had any effect on the quality of the written feedback included in the CCEs.

The second, qualitative phase permitted continued exploration of these findings and promoted an in-depth understanding of physicians' perceptions of CCEs. Semi-structured, one-on-one interviews were conducted with a subset of the physicians sampled and resulted in the emergence of eight themes related to participants' perceptions of the utility, quality, cost, and practicability of CCEs. Participants' perceptions of the 'utility' of the evaluative instrument varied, with some finding the instrument to be a useful model for evaluating students' clinical skills, while others believed it did not permit assessment of the more relevant aspects of students' performances. Despite bipolar views on the utility of the instrument, participants generally agreed on the usefulness of providing students with feedback, but found informal *verbal* feedback to be more useful than formal *written* feedback in promoting students' professional development. Additionally, participants reported the breadth of the student-faculty interaction and the evaluative culture of their medical department to be influential determinants of CCE 'quality.' Moreover, all participants believed the temporal demands required by CCEs were the most 'costly' component of the evaluation process. Finally, participants perceived the 'practicability' of the evaluations to be influenced by both the timing of requests for evaluation submissions and one's prioritization of administrative/clinical responsibilities.

This chapter will present an integrated discussion of the study's most salient quantitative and qualitative findings and the implications of these findings, beginning with the observed lack of influence of physician evaluative load on evaluative strain. Next, the influence of evaluative load on the quality of the closed-ended CCE items will be presented, followed by a combined explanation of the influences of evaluative strain and time delay on the quality of the CCE items. The body of this chapter will conclude with a rationalization of the apparent lack of influence of either physician evaluative responsibilities or time delay

on the quality of written feedback. Lastly, the study's limitations and ideas for future research will conclude the chapter.

INTEGRATION AND INTERPRETATION OF THE FINDINGS

The Influence of Evaluative Load on Evaluative Strain

As a budding area of medical education research, investigations into evaluator ('rater') perceptions have begun to elucidate factors believed to influence rater perceptions and rating quality. Though the principal focus of this work was to investigate how the quality of CCEs was influenced by physicians' evaluative responsibilities, this study additionally sought to add to this growing body of research by discerning the nature of the relationship between one's imposed and perceived evaluative responsibilities. Accordingly, this study explored the influence of evaluators' measurable evaluative quantity (i.e., evaluative load) on their evaluative perceptions (i.e., evaluative strain). Interestingly, the path analysis reported no significant association between one's evaluative load and strain. As the first study to formally investigate the causal association between one's imposed and perceived evaluative responsibilities, it is difficult to place these findings within the context of the extant medical education literature.

At the broadest level, these findings suggest physicians' evaluative perceptions are influenced by factors other than the number of evaluations one completes. Although task perceptions are commonly guided or modeled by a task's more objective characteristics,[34] perceptions ultimately remain subjective, with their reasoning often known only to the task performer.[91,92] Research that has investigated the influence of 'rater demands' on rater perceptions suggests that the number of items one is asked to rate/evaluate,[26,35] the time required to complete such ratings/evaluations,[93] and the influences of one's environment[5] may contribute more heavily to one's perceptions of the task than one's assigned task quantity (i.e., number of evaluations one is asked to complete). Indeed, participants'

qualitative admissions revealed the length and usefulness of the evaluation instrument, the time requirements imposed by the evaluation process, and concerns regarding the evaluative environment (e.g., allocation of evaluative responsibilities; the duration and relevance of student-evaluator interactions; departmental hierarchy; and departmental salience of evaluation proficiency) all strongly influenced their evaluative perceptions (see Tables 4-9, 4-10, 4-11, and 4-12 for a summary of these themes).

Although the qualitative findings help interpret the lack of association between physician evaluative load and strain, they do not provide a complete answer. Additional studies that explore the influence of physicians' imposed evaluative responsibilities on their perceptions of the evaluative task are needed. Because this study used CCEs as a surrogate measure for physician evaluative load, it may be pertinent to examine how the full extent of physicians' measurable survey, questionnaire, and evaluative burden influences their perceptions of this load.

The Influence of Evaluative Load on the Quality of Closed-ended CCE Items

Though several works have provided evidence supporting theorized relationships among physicians' measurable evaluative responsibilities and the quality of physician-performed evaluations of medical trainees,[26,35-37] this study was the first to formally investigate this association. An examination of the direct and indirect effects of evaluative load on the quality of the closed-ended CCE items revealed a direct, positive association between the constructs. Though this finding suggests that physicians who regularly complete a high quantity of CCEs produce evaluations with a high degree of variation within and across clerks, participants orally described a contrasting association between the constructs. Specifically, participants with high evaluative load reported experiencing marked difficulty in noticing subtle aspects of students' clinical performances; consequently, participants felt unable to accurately report their judgments using the closed-

ended CCE rating scales and believed the CCEs they produced to be of poor quality and little utility to evaluees. Moreover, participants' desire to 'not harm' evaluees with their CCE judgments further exacerbated their concerns regarding the utility of the evaluations they produced. At their broadest level, these divergent accounts of the influence of evaluative load on evaluative quality highlight how little is known regarding the determinants of evaluative quality in performance evaluations. Within the context of this study, however, this ambiguity suggests that participants' initial ratings of students' clinical performances may not be reflective of students' true abilities; and in turn, this misrepresentation of students' scores likely resulted in the production of a 'false positive' association between evaluative load and the quality of the closed-ended CCE items. Evidenced by anecdotal reports and little empirical data, this claim remains largely speculative. Nevertheless, a working explanation for this claim follows.

Though this study was the first to investigate the influence of physicians' evaluative load on the quality of CCEs, others have sought to investigate the influence of rater demands on evaluative quality. In two separate studies, Tavares and colleagues[26,35] demonstrated that increasing rater demands, such as the number of behaviors to be identified during a performance evaluation, may overextend raters' mental resources and prevent them from attending to nuanced aspects of a trainee's performance. As a consequence, raters who experience difficulty in recalling the specifics of observed performances are often 'cognitively overextended' and more likely to engage in 'load-avoidance strategies' or 'satisficing tendencies' to prevent further mental exhaustion.[26] The employment of such satisficing behaviors typically results in the selection of a rating or response deemed 'good enough' to satisfy the task without exhausting one's mental resources further and may result in the production of overinflated ratings based on raters' seemingly random selection of responses. Among many, examples of satisficing behaviors include 'endorsing the *status*

quo,' or choosing a response simply because it is easier for the evaluator to "keep things as they are," than to provide a more individualized or representative response; and producing ratings that vary only slightly from a 'reasonable' response given to the first item included on the rating instrument.[78] For example, a rater utilizing the latter satisficing behavior who gave an evaluee a '4' (out of 5 possible points) for the first item on the rating instrument would rate all subsequent items as a '3,' '4,' or '5,' regardless of the evaluee's actual performance. Moreover, such satisficing behaviors have the propensity to manifest as 'range restriction' and 'grade compression/inflation' in summative assessments, as producing responses that vary only slightly tends to cluster ratings to one area of the rating instrument and results in a narrowing of the grading scale.[13,78,93]

As the study's conceptual model did not include a measure of participants' cognitive exertion (evaluative strain measured only *perceived* cognitive demands needed to complete the evaluative task), it is not possible to empirically conclude that the participants in this study were 'cognitively overloaded' by their attempts to recall the details of students' performances. However, participants' oral testimonies revealed that many felt "overwhelmed" and "fatigued" by their evaluative responsibilities and that participants generally experienced difficulty in recalling students' observed behaviors during the completion of the evaluations. Given these admissions, it is reasonable to suggest that the participants in the current study experienced some degree of cognitive exertion while trying to recall the nuances of students' performances and likely engaged in satisficing behaviors to ease their cognitive burden. Post hoc descriptive analyzes were performed to ascertain the validity of this sentiment and to examine the score distribution for the closed-ended CCE items. Had participants utilized the full range of the rating instrument (and not engaged in satisficing behaviors and/or grade compression behaviors), one would theoretically expect a mean evaluation score of 2.5/5 and a normal distribution throughout

the mid-point of the data. Examination of the data, however, revealed an average evaluation score of 3.62±0.65/5 and a score distribution skewed toward higher grades, with the majority of scores falling between '3' and '4.25' (see Figure F-1 in Appendix F for a histogram of these findings).

With evidence of range restriction at the high-end of the evaluative instrument, these data suggest that participants may have experienced difficulty in recalling nuanced aspects of students' performances and engaged in satisficing behaviors that produced clustered responses around the high-end of the evaluative instrument. Moreover, participants' strong resolve to 'not harm' students with their recorded judgments likely contributed to this skewed distribution of scores. During the qualitative interviews, several participants noted a hesitancy to include 'negative judgments' of students' observed behavior, as it was believed that any negative ratings included on a clerk's CCE would influence the clerk's residency matching potential. As a consequence, participants' admitted to censoring their ratings of negative judgments so that evaluees were portrayed positively and not heavily affected by evaluators' ratings. Given participants' oral testimonies, it seems likely that this self-censorship further contributed to the skewed distribution of scores at the high-end of the evaluative instrument. Furthermore, this oral recognition of the prevalence of grade inflation among the clerkship rotations further supports the use of satisficing behaviors among participants and strengthens the claim that participants' recorded ratings were not representative of students' behaviors.*

*Though this working explanation primarily suggests participants likely engaged in satisficing behaviors that produced a skewed distribution of scores at the high-end of the evaluative instrument, it is important to note that additional 'rating errors,' such as the 'halo effect' (i.e., the tendency of a rater to judge a trainee's clinical competence based on a single observed trait)[13] may have occurred, as well.

If this anecdotal and empirical evidence is true, the misalignment between observed behaviors and recorded scores across the closed-ended CCE items might explain why the path analysis found a positive association between physicians' evaluative load and the degree of score differentiation. An alternative yet competing interpretation of these findings is that physicians with high evaluative load may have produced higher quality closed-ended CCE items as a simple consequence of observing more students. Colloquially summarized by the old adage, 'practice makes perfect,' deliberate, repeated task exposure and experience is well-believed to promote 'expert' performance in that task.[94] As others have shown that evaluators tend to display increased sensitivity to performance nuances with increasing assessment practice,[94,95] it is possible that physicians with high evaluative load may exhibit 'expert' rater tendencies as a consequence of evaluating more students. Accordingly, physicians with high evaluative load may have been more able to discern behavioral and contextual cues more readily than those with less evaluator experience and were able to differenatiate among those behaviors on the CCEs they completed. This ability to aptly differentiate among observed behaviors would have yielded a higher degree of score differentiation within and across learners and manifested as 'good' quality among the closed-ended CCE items.

Though the anecdotal and empirical evidence provided above offer reasonable explanations for the production of a positive association between physicians' evaluative load and the quality of the closed-ended CCE items, it remains unclear which of these interpretations (if any) is the most accurate. Additional investigations into the associations between physicians' imposed evaluative demands and the resulting evaluation quality are warranted to develop a more complete understanding of these relationships. Future works into the role of satisficing behaviors among physician evaluators and the influence of

evaluators' concern for their evaluees are also needed to understand the full implications of this work.

The Influence of Evaluative Strain and Time Delay on the Quality of Closed-ended CCE Items

As an extension of the work previously performed by Williams et al.,[37] this study additionally sought to determine how the length of delayed evaluation submissions directly influenced the quality of CCEs produced and indirectly mediated the influences of evaluative load and strain. The delayed completion of performance evaluations has been occasionally cited in the medical education literature as a negative determinant of evaluative quality,[37,96,97] yet the results of this study revealed a direct, positive association between the length of time delay in CCE submissions and the quality of the closed-ended CCE items (see Figure 4-4). Furthermore, this study revealed a positive, indirect association between evaluative strain and the quality of the scaled-data through the mediating effects of time delay (see Figure 4-6). These findings directly contradict those observed by Williams and colleagues, whose retrospective study on the quality of resident operative performance ratings (OPRs) demonstrated that lengthy delays in evaluation submissions negatively impacted the amount of "item-to-item variation."[37] Expanding on the hypothesis purported to rationalize the positive findings between evaluative load and score differentiation, this discrepancy in findings may be reflective of satisficing behaviors and/or participants' concerns for evaluees' residency matching. Engaging in such satisficing behaviors and showing hesitation to report negative feedback in students' CCEs may have skewed the data toward the positive end of the evaluative instrument and caused the path analysis to produce an artificial, positive association among the variables. Though speculative, it is further possible that participants' concerns for trainees and their perceptions of the evaluative process were compounded by delays in evaluation submission and resulted in

the production of inflated ratings that varied across the high-end of the instrument (see Figure F-1, in Appendix F). Such actions would have satisfied the participants' desire to provide students with high scores that could not harm their residency chances and would have manifested as a high degree of score differentiation both within and across evaluations as theorized above.

The results of this study seemingly defy the logical contention that ratings completed immediately following an observation provide a rater with an ideal opportunity to recall and document specific nuances of a trainee's clinical performance. An examination of the factors perceived to influence the delayed submission of evaluations, however, helps to rationalize these findings. According to participants' oral testimonies, the culture of some medical departments results in requests for evaluators to complete CCEs prior to the official end of the clerkship rotation period. Many participants were distressed by this request, as they felt pressured to complete evaluations before they had had enough time to formulate a judgment of students' performances. This pressure to complete evaluations based on insufficient contact time was exacerbated by a low prioritization of CCEs and lack of protected time to complete the evaluations, which resulted in high evaluative strain and lengthy delays in evaluation submissions. When presented with time to complete the evaluations, participants reported difficulty in recalling the specifics of students' clinical performances, despite good intentions and a strong desire to provide students with constructive feedback. Classically, these difficulties are considered impediments to evaluative quality, though the path analysis reported alternative findings.

Clearly, additional work is needed to understand the true implications of these findings. Though the exact relationship between physicians' perceived evaluative responsibilities, time delays in evaluation submission, and the quality of closed-ended CCE items remains unclear, these findings still hold implications for improving the timeliness of

evaluation submissions. Principally, reminding participants of the need to submit timely evaluations through emails may help to promote cognizance of timely submissions. Although some IUSM medical departments currently use deadline notifications built-in to the evaluation platform to remind evaluators of deadlines, these 'reminders' may be taken more seriously if delivered by leaders in undergraduate medical education, such as Clerkship Directors, Coordinators, or even Vice Chairs. As an extreme example, one participant admitted being months behind in submitting evaluations, but felt no need to complete them because the Clerkship Director had not directly asked for their submissions. Though alarming, this example highlights the need for increased monitoring of the timeliness of evaluation submissions. Finally, the implementation of an incentive program may help decrease delays in evaluation submissions. Such incentives may be reward-based, or could factor into physicians' relative value units (RVUs). Either way, increasing the perceived salience of these evaluations is likely to improve the timeliness of evaluation submissions as a whole.

The Lack of Influence on the Quality of Written Feedback

Evaluative Responsibilities and the Quality of Written Feedback

Despite the observed influences of physicians' evaluative responsibilities on the quality of the closed-ended data, the path analysis revealed no association between physician evaluative load or strain on the quality of the written feedback provided to students. Broadly, these findings suggest physicians' written comments are influenced by other factors. Given participants' qualitative testimonies, it is clear that one such factor influencing narrative quality is the perceived utility and salience of written feedback. Though all participants desired to provide students with meaningful feedback that could be used to improve their clinical performance, they feared including any written feedback that could negatively impact students' chances for residency matching. As a consequence,

participants reported providing students with timely, actionable, and realistic *verbal* feedback that could be used to address weaknesses in their clinical performance, while providing self-censored *written* comments devoid of negative feedback on their CCEs that could be used in students' residency applications. In this way, participants perceived the purpose of the written feedback not as a way to offer advice aimed at improving students' clinical performances, but as a way to satisfy the demands of the evaluative task without creating negative ramifications for students. As a result, the written feedback included on students' CCEs was perceived by evaluators to be of little utility to students. These results are consistent with findings produced by Govarets et al.,[98] whose study on the quality of feedback given after performance assessments demonstrated that narrative comments tend to be "less useful for learning and professional development than verbal feedback." As indicated by the findings, the written feedback was generally global, non-specific, and limited in practical application. Overall, only 18% (17/93) of participants received full credit on their written feedback score for including of a 'specific comment' in each of their evaluations and less than one percent (1/93) received credit for including a 'practical comment.' Similar findings have been reported in the literature, with Brutus[99] and Cook[100] noting that unmotivated raters tend to produce more global, non-specific comments.

With both a fear of written feedback having a detrimental effect on trainees' residency matching potential[93,101,102] and a general unwillingness to document negative clinical behaviors[96,103] commonly cited as causes of grade inflation within clerkship rotations, it may be prudent to formally designate an area of the evaluation as formative feedback. Though some participants believed that the narrative portion of the evaluation was reserved just for students, others perceived the evaluation as a whole to represent a summative measure of student performance intended for the eyes of administrators. With literature indicating that evaluators prefer to 'speak' differently to the evaluee than the

administrative recipient of the evaluation,[99,104] providing a 'safe haven' box for formative comments, or making the usage of this section explicitly clear to evaluators, may encourage the delivery of realistic, constructive feedback for students. Moreover, the formative purpose of such a 'safe haven' box should be clearly articulated to clerks, as well. With the 'threat of student nuisance' cited as the most frequent cause of grade inflation in clerkship rotations,[93,102,105] clerks may feel less inclined to argue for altered evaluations if they understand that the formative feedback provided on their evaluations will not impact their overall clerkship grade. This, in turn, may decrease the number of evaluator-student confrontations and may help evaluators feel more comfortable providing clerks with constructive formative feedback. Finally, asking evaluators to provide a rationale for their written feedback has been shown to increase the accuracy and reliability of comments,[95] and may prove a relatively simple mechanism for improving the quality of the open-ended feedback provided on performance evaluations.

Alternatively, recent work into the cognitive processes underlying the evaluative process suggests that the structuring or wording of narrative question prompts may additionally influence the quality of written feedback provided on performance evaluations. Though questions used to elicit performance ratings using rating scales narrow raters' focus by calling attention to a single aspect of the performance experience (e.g., questions regarding the caring disposition of a clerk tend to facilitate rater recall of any behaviors demonstrating empathy or apathy towards patients), questions used to elicit narrative comments provide much less guidance.[106,107] Unable to simply select the most appropriate descriptor of the student's performance as is done using rating scales, narrative comment sections require the evaluator to articulate their judgments and consequently result in increased difficulty in recalling specific and relevant behaviors.[99,107] With numerous participants citing difficulties in recalling aspects of students' performances and an average

quality of feedback score of 2.48±0.32/4 across all participants, it's plausible that the physicians in this study experienced general difficulty in writing narrative comments. Accordingly, asking evaluators to briefly annotate/dictate their daily interactions with students may be one additional and simple way to increase memory recall and improve the caliber of written comments. This behavior was described by one participant, who believed that access to previous notes taken during the observation greatly improved his ability to recall performance specifics and lessened the difficulty of crafting written comments.

The Timeliness of Evaluation Submissions and the Quality of Written Feedback

Despite the positive association between time delay and quality of the closed-ended CCE items, time delay had no influence on the quality of written feedback. These findings directly contrast the work of Williams and colleagues[37] whose study on the quality of resident operative performance ratings (OPRs) demonstrated that lengthy delays in evaluation submissions negatively impacted the specificity of written feedback. The discrepancy between these findings may be reflective of the type of evaluation under investigation. In their study, Williams et al.[37] investigated the influence of time delay in completing OPRs for surgery residents. As a consequence of evaluative protocols mandated by the American Board of Surgery, each general surgery resident's performance within the operating room must be directly observed and judged by a surgery faculty member using an OPR on six distinct occasions.[6] Consequently, OPRs represent an evaluator's judgment of a *single* observed performance.[108] With an evaluator required to note the nuances of each resident's performance, it is likely that written feedback provided in OPRs is generally highly focused and detailed. It follows, therefore, that the written feedback included in OPRs may be highly susceptible to time delays in evaluation submissions as evaluators are likely to forget most of the observed subtleties of residents' performances in the days following the observation. Indeed, the results of the study performed by Williams et al.[37] indicated

that evaluations completed even 72 hours post-observation maintain only a reasonable amount of clarity and detail. In contrast, CCEs are not required for each direct observation or encounter between a supervising physician and a clerk. Rather, CCEs typically represent an evaluator's cumulative judgment of a trainee's overall clinical performance and can be based on a single observation or multiple encounters with varying numbers of patients. With evaluators' judgments based on a wide range of student observations, it is reasonable to suggest that evaluators may be unable to recall clerks' performances with the same level of detail as an evaluator completing an OPR immediately following an operative experience. Consequently, the delayed submission of CCEs may not be as detrimental to the quality of written feedback as the delayed submission of OPRs, as feedback included on CCEs is likely to be of lower specificity and clarity even if completed immediately following an observation. Stated otherwise, it is possible that delays in CCE submissions did not influence the quality of written feedback because delays only served to lessen the specificity of already predominately global comments.

Clearly, evaluations completed immediately following the direct observation of a student's clinical performance should still be encouraged, as the immediacy of such a rating provides an ideal opportunity for the rater to recall and document nuances of a trainee's clinical performance. Minimally, the amount of time that elapses between an observation and a rating should be monitored, with Clerkship Directors perhaps weighting the feedback of these evaluations less than evaluations submitted in a more timely fashion. Ideally, the implementation of a smartphone-based system with dictation features similar to that proposed by Williams and colleagues[37] or Ferenchick et al.[109] may be beneficial in promoting the timely completion of CCEs. Such platforms permit easy completion of performance-based evaluations, reduce delays in evaluation submissions, provide a record of elapsed time between the last observation and the rating, and can be used at the program

director's discretion.[37] Finally, the creation and implementation of a campus-wide rotation calendar, as proposed by one participant, would also improve the practicability of these performance-based evaluations. Such a calendar would provide structured observation and assessment opportunities for evaluators across all departments and programs. This structure, in turn, has the potential to reduce the temporal demands of completing CCEs by giving physicians a little 'protected time' to complete their evaluations and may help to prioritize evaluation completion.

LIMITATIONS

Though efforts were made to minimize potential sources of bias, this work is not without limitations. Concerns regarding the generalizability of the findings, the study's sample size, response rate and demographic characteristics, missing data, construct conceptualization, measurement error, and researcher bias all warrant acknowledgement.

Generalizability

Though considered a form of "causal modeling," path analysis provides only estimates of the direction and magnitude of hypothesized causal relationships among a grouping of variables. Because path analysis does not utilize an experimental design, it does not permit the deduction of causal conclusions. Additionally, this study occurred at a single institution. As a consequence, these findings may not generalize to physicians involved in the evaluation of clerks at other medical schools. Differences in departmental evaluative culture, the length and wording of the evaluative instrument, and the frequency with which such evaluations are performed are likely to produce different results.

Sample Size

One notable limitation is the study's modest sample size (n = 93). A power analysis was conducted to determine the likelihood of accepting the fit of the conceptual model based on sample size and degrees of freedom. Despite the strength of the model fit, the

power analysis estimated a power of 0.08, which did not reach the accepted standard of 0.8. According to the values provided by MacCallum et al.,[110] a total sample size of 1069 physicians would be needed to ensure the model maintained a "close-fit." Despite inadequate study power, it is important to note that there are no definitive standards for determining adequate sample size in path analysis. In the absence of such regulations, several 'rules of thumb' are commonly followed. In this study, the 10:1 ratio of observations to variables or parameters was respected. Though a 20:1 ratio is suggested as the 'ideal' situation, path analysis can still be conducted in situations that maintain only a 5:1 ratio.[111] This ratio was confirmed by the statistical software package used to conduct the analysis (*LISREL*, Scientific Software International, 2016, Version 9.2) and indicated that the model possessed an adequate number of observations per variable. Such confirmations help to mitigate the potential limitation of the study's small sample size.

Response Rate and Participant Demographic Characteristics

In a similar vein, the response rate to the electronic study was only 11% (160/1518). The timeliness of survey administration may have contributed to the low response rate, as surveys were distributed during the first two months of the third year clerkship rotation schedule. With most third-year clerkship rotations lasting four or eight weeks, it is likely that the survey administration coincided with administrative requests for CCEs and was perceived as a lower priority task. Administering the survey between mid-May to mid-June may have resulted in a higher response rate, as no third year clerkship rotations were scheduled during that time. Moreover, survey administration relied on the willingness of Clerkship Directors and Vice Chairs of Medical Education to distribute the survey link to physicians within their respective departments. With many Clerkship Directors maintaining private email addresses used to alert their physicians to matters requiring their attention, it was believed that the delivery of the survey link through

familiar channels would result in higher participation than the research team was able to recruit alone. Nevertheless, it is likely that a higher response rate may have been achieved had the research team additionally distributed the survey link to the physician population.

With a low response rate, the demographic characteristics of the participants also became a limitation. In spite of the low response, the participants for the quantitative phase remained fairly well distributed. Gender proportions were nearly equitable, with males comprising 55% of the sample. Despite the absence of participation from fellows or instructor/lecturers, the remainder of the sample remained appropriately balanced among residents (34%) and assistant (26%), associate (27%), and full professors (13%). Moreover, at least one participant from each of the targeted medical departments participated, though the sample was over-representative of internists and pediatricians, with each representing nearly 37% of the sample. Though efforts were made to ensure a broad representation of participants for the qualitative phase, it should be noted that females (37.5%) and associate professors (12.5%) were underrepresented, while internists were again overly present.

Missing Data

Though 11% (160/1518) of the total physician population completed the survey, 67 participants were excluded from the study due to missing data. Several of these participants did not complete all necessary components of the EVAL-TLX needed to calculate their numerical evaluative strain, while others did not have complete E*Value records that made the computation of one or more of their evaluative variables (e.g., evaluative load or time delay) incalculable. While some missing EVAL-TLX data were collected through follow-up emails or inquiries made during the semi-structured interviews, participants with incomplete datasets were ultimately excluded.

Construct Conceptualization and Measurement

As this work was the first to formally characterize the evaluative load and strain constructs and quantitatively investigate their relationship, it is possible that the constructs may not have been fully measured. Within this study, 'evaluative load' was defined as the measurable quantity of evaluations, ratings, or surveys completed by a physician during a specific period-of-time and was measured by counting the number of CCEs a participant completed during the study timeframe. The decision to use CCEs as a surrogate measure for one's evaluative load was based on the assumption that physicians view their evaluative tasks separately (e.g., the number of resident performance evaluations one needs to do would be conceptualized differently than the number of CCEs one has to do). However, exploration of participants' qualitative admissions suggests that physicians perceive their evaluative load as a singular, summative construct; stated otherwise, the physicians in this study did not appear to differentiate among their evaluative tasks; each survey, evaluation, or questionnaire to be performed added to their cumulative evaluative load. Consequently, it is possible that the decision to use CCEs as a surrogate measure for physician evaluative load may have resulted in an incomplete measure of the variable. Similarly, strain may have been limited in its conceptual design. By limiting measure of one's evaluative strain to perceptions of only six cognitive dimensions experienced during the evaluative process (i.e., mental, physical and temporal demands; task complexity; situational stress; and distractions), it is likely that the more nuanced facets of the construct may not have been elucidated. Indeed, the qualitative interviews demonstrated that participants have unique perceptions regarding the quality, utility, cost, and practicability of completing CCEs, yet the EVAL-TLX did not include direct measures of these perceptions.

Measurement of Time Delay

Participants' time delay was computed by calculating the time difference between the 'date submitted' and the last date of observance listed on their E*Value records. This calculation was predicated on the assumption that participants' CCEs were completed the same day they were submitted. Had participants been able fill-out the CCE form, save their work, and submit completed evaluations on a subsequent day, participants' calculated time delays would not have been representative of the true lapse in time between their observation and judgment. In turn, this could have altered the results of the path analysis. This is an important study limitation, as the researcher had no way to ensure that this assumption was merited.

Researcher Bias

Finally, it should be noted that this research took place at the primary researcher's home institution. Though it is possible that the researcher's relationship with the institution may have influenced participants' survey data or qualitative admissions, the maintenance of a rigorous research protocol, assurance of anonymity, and continual reflection on the researcher's bias may have helped to mitigate such influences.

SUGGESTIONS FOR FUTURE RESEARCH

To the researcher's knowledge, this study was the first to (1) quantitatively investigate the extent to which physicians' evaluative responsibilities and the length of delay in evaluation submissions influenced the quality of CCEs; and (2) qualitatively explore physicians' perceptions of CCEs and the evaluation process. Though this study has provided new insights into the role of evaluative responsibilities, perceptions, and the timeliness of evaluation submissions on evaluative quality, further investigation is needed to understand the true implications of this work. The following suggestions for future research are

believed to be pertinent extensions of this work and may help to elucidate the true nature of the causal associations unearthed in this study.

- Because the scope and nature of the relationship between physicians' evaluative load evaluative strain remains largely unclear, expanding research efforts to investigate both of these evaluative constructs is warranted. As this study used CCEs as a surrogate measure for physician evaluative load, it may be pertinent to examine how the full extent of physicians' evaluative responsibilities (beyond CCEs alone) influence their perceptions of the evaluations they complete. Moreover, the instrument used to measure participants' evaluative strain examined only six aspects of human perception believed to be prominently experienced during task completion (i.e., mental, physical, and temporal demands; task complexity; situational stress; and distractions). Given participants' testimonies regarding their mixed perceptions on the utility, quality, cost, and practicability of CCEs, including a measure of these areas may promote a more detailed understanding of how physicians' conceptualize their evaluative responsibilities. Likewise, it is feasible that procrastination tendencies and levels of motivation contribute to physician evaluative strain and the timeliness of evaluation submissions. Further investigation into the role of these constructs on physicians' perceptions of their evaluative tasks and time delays in evaluation submissions are likely to demonstrate influential effects.
- Though this work presented anecdotal and empirical evidence suggesting that the physician population may have succumbed to satisficing behaviors that influenced grade inflation/grade compression, future investigations into the role of satisficing behaviors among physician evaluators are needed to understand the full implications of this work. Moreover, a thorough examination of the ubiquity of grade inflation within the institution (and the contributing factors of such inflation) may illuminate deficiencies in either the

evaluative process or the evaluative instrument that could ultimately increase evaluative utility and quality.

- During interviews, it was revealed that the sharing of evaluative duties among several physicians within a medical department is perceived to positively influence evaluative quality, while assigning a single physician to all medical student evaluations is believed to be a substantial detriment to evaluative quality. Additional studies are needed to confirm these findings. If true, promoting an equal distribution of evaluative responsibilities across evaluators through the use of sophisticated evaluation monitoring systems may prove a solution for augmenting the quality of CCEs.
- As previously indicated, this study failed to identify an association between physician evaluative responsibilities, the timeliness of evaluation submission, and the quality of written feedback provided on CCEs. Further work is needed to understand which factors influence written feedback. Continued investigations into the validity, reliability, and usability of CCEs may illuminate some of these factors.
- Finally, future studies should investigate a causal association between physicians' evaluative responsibilities, the timeliness of evaluation submission, evaluative quality, and burnout. Defined as a psychological syndrome involving a chronic response to interpersonal job stressors,[112] job burnout is a well-established phenomenon within the physician population. Numerous aspects of one's job may influence the prevalence of burnout syndrome (e.g., a general misalignment of values, lack of workplace-based flexibility,[113,114] etc.), yet workload-related factors are believed to most prominently contribute to burnout among physicians.[115] Although work hours[116] and high job demands[113] have been demonstrated to contribute to burnout, no studies to date have investigated the influence of physicians' evaluative workload (imposed or perceived) on burnout syndrome. Accordingly, the influence of evaluative load and strain on burnout

should be investigated. The implications of such work are likely to be substantial, and may permit a more complete understanding of factors perceived to influence evaluative quality, as well.

CONCLUSIONS

Despite longstanding criticisms of the validity and reliability of performance-based evaluations of medical trainees' clinical performances, few research studies have investigated the causal influences of certain factors on evaluative quality. In an attempt to illuminate such influences, this work explored the extent to which physicians' evaluative responsibilities (both imposed and perceived) and temporal delays in evaluation submissions influenced the quality of both the Likert scaled-data and written feedback included in clinical clerkship evaluations (CCEs) completed for third-year medical students. To qualitatively explore these relationships in more-depth, semi-structured interviews gathered perceptions on the utility, quality, cost, and practicability of completing these evaluations. In spite of the theoretical evidence in support of the relationships among these variables, the results of this study have produced more questions than answers.

A path analysis was used to estimate the influence imparted by physicians' evaluative responsibilities and delays in evaluative submission on evaluative quality. Though the results produced several significant associations, nearly all associations contradicted the tangential evidence initially provided in support of these relationships. Physicians' perceptions of their evaluative responsibilities were not influenced by their assigned evaluative quantity, but were casually associated with the length of time delay in evaluation submission and (indirectly) the quality of the closed-ended CCE items. The quality of the closed-ended items was additionally influenced by physicians' evaluative load, but load had no influence on either time delay or the quality of written feedback. Interestingly, no variable included in the model was causally associated with the quality of

written feedback. Though participants' perceptions of CCEs helped rationalize these findings and may have highlighted the role of grade inflation in evaluative quality, the true implications of these findings remain cloudy.

In its entirety, this dissertation illuminates the need for continued investigation into the influence of physicians' evaluative responsibilities, perceptions, and submission timeliness on evaluation quality. Above all else, this work provides a foundation for further research and accentuates the complexities and nuances of performance evaluations. It is my hope that this work will ultimately benefit the physicians and students at the heart of these evaluations, and that these findings will help to ease the evaluative burden of physicians so they might reclaim the joy inherent in teaching medical students.

APPENDICES

Appendix A: Evaluation of Student Clerkship Performance ('Medium' Form)

Appendix B: Evaluation of Student Clerkship Performance ('Long' Form)

Appendix C: The Electronic Study Survey

Appendix D: Quality of Clinical Clerkship Evaluation (CCE) Rubric

Appendix E: Interview Protocol

Appendix F: Histogram Illustrating Participants' Scores Across the Closed-ended CCE Items

APPENDIX A: EVALUATION OF STUDENT CLERKSHIP PERFORMANCE ('MEDIUM' FORM)

Subject:
Evaluator:
Site:
Period:
Dates of Activity:
Activity: General Surgery Evansville
Form: Educator of Medical Student

Please evaluate this student's clinical performance compared to other students you have supervised (during a similar time in the academic year). Using the Likert scale with the representative descriptive anchors below, please check the most appropriate box, or N/A (not applicable/not observed) next to each category listed. Comments are required for all evaluations.

PROFESSIONAL ATTRIBUTES

(Question 1 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Unreliable, irresponsible. Unexcused attendance and/or tardiness.	Occasionally unprepared or unwilling to take on responsibility, lackadaisical. Unreliable in being accountable for own errors and shortcomings.	Reliable, punctual, fulfills responsibilities. Ensures proper transfer of patient care responsibilities. Accountable for own errors and shortcomings.	Very dependable. Seeks additional responsibilities. Devises and implements plans to correct own errors and shortcomings.	Exceptionally conscientious. Remarkable dedication to patient care; checks on patients after hours.	
Duty: Responsibility/Accountability	1.0	2.0	3.0	4.0	5.0	0

(Question 2 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Demeaning or condescending towards patients. Refuses to deal with patient distress. Avoids talking with patients and their families. Lacks effective communication skills.	Some deficiency of empathy and compassion for patients. Often does not elicit or identify patient's barriers to care. Communication skills need improvement.	Generally courteous towards patients. Appropriate communication skills. Generally elicits and identifies at least one barrier to care.	Strong empathic skills. Makes effort to seek out and talk with patients and their families. Relatively advanced communication skills. Routinely elicits and identifies most of patient's barriers to care.	Deals with sickness, death and dying in a highly professional and effective manner with patients and their families. Highly advanced communication skills; easily adjusts communication style to each individual patient or situation.	
Caring, Compassion, & Communication	1.0	2.0	3.0	4.0	5.0	0

(Question 3 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Poorly integrated into the team. Antagonistic or disruptive. Disrespectful of others. Unprofessional attire.	Occasional difficulty in relating to health care team members. Occasional lapses in use of professional language or in professional appearance.	Cooperative, productive team member. Uses professional language with patients and colleagues. Appropriate professional attire.	Relates well to health care team members and functions well on team.	Outstanding in respecting the feelings, needs, rights, and diversity of team members and other co-workers. Highly integrated team member.	
Respect for Others: Working Relationships	1.0	2.0	3.0	4.0	5.0	0

DATA ORGANIZATION & REPORTING

(Question 4 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Inaccurate data or major omissions. Reflects poor understanding of disease process and patient's psychosocial situation.	Lacks supporting detail, unfocused, needs organization. Incomplete understanding of disease process and patient's psychosocial situation.	Generally accurate and complete. Reflects basic understanding of disease process and patient's psychosocial situation.	Complete, appropriately focused, and organized.	Complete, but concise. Reflects thorough understanding of disease process and patient's psychosocial situation.	
Written Histories & Physicals	1.0	2.0	3.0	4.0	5.0	0

KNOWLEDGE BASE AND EDUCATIONAL INITIATIVE

(Question 5 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Lacks knowledge to understand patient problems.	Inconsistent understanding of patient problems. Limited differential diagnoses. Needs prompting to read and learn about own patients.	Knows basic differential diagnoses of patients' active problems. Reads appropriately.	Expanded differential diagnoses. Demonstrates extra reading about patients' problems. Searches literature for best evidence.	Self-directed learner; routinely poses insightful questions and effectively searches literature for best evidence. Differential diagnoses are expanded and well-prioritized.	0
Knowledge Base and Educational Initiative	1.0	2.0	3.0	4.0	5.0	0

DATA INTERPRETATION & INTEGRATION

(Question 6 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Cannot integrate and interpret basic data.	Often reports data without analysis; often does not consider pertinent psychosocial factors. Problem list incomplete or unprioritized. Often unable to defend differential diagnoses.	Integrates and interprets data, including psychosocial factors, at a basic level. Generally appropriate problem list & differential diagnoses.	Thoughtful integration and interpretation of data, including pertinent psychosocial factors. Consistently able to defend differential diagnoses.	Highly thoughtful integration of data to identify, prioritize, and solve patient problems. Understands complex issues, interrelates patient problems.	0
Data Interpretation & Integration	1.0	2.0	3.0	4.0	5.0	0

CLINICAL JUDGMENT/MANAGEMENT

(Question 7 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Fails to recognize, identify, or prioritize urgent problems. Makes inappropriate diagnostic and therapeutic decisions. Lack of knowledge of routine screening. Unable to identify patient risk factors.	Inconsistent prioritization of clinical issues and decision-making ability. Often does not consider patient's psychosocial factors in decision making.	Appropriate basic diagnostic and therapeutic decision-making, taking into consideration patient's psychosocial factors. Performs appropriate health screening and usually identifies/addresses health risk factors. Generally aware of own limitations and seeks appropriate help.	Consistently makes sound diagnostic and therapeutic decisions, taking into consideration patient's pertinent psychosocial factors.	Highly thoughtful approach to management plans; consistently integrates new data, adjusts plans accordingly, and carefully weighs all alternatives, including patient's preferences and psychosocial factors. Comprehensive in application of routine health screening; addresses health risks.	0
Clinical Judgment/Management	1.0	2.0	3.0	4.0	5.0	0

DEVELOPMENTAL ASSESSMENT *(Question 8 of 13 - Mandatory)*

Selection	Option
	INCONSISTENT REPORTER: Performing consistently at the "Reporter" level is the minimum requirement for clinical clerkships.
	REPORTER: Consistently acquires and communicates clinical information accurately; this includes properly identifying patient problems and constructing appropriate problem lists. Also consistently demonstrates satisfactory professional behavior.
	REPORTER/INTERPRETER: Starting to integrate and interpret the collected data to develop reasonable differential diagnoses, but not yet on a consistent basis.
	INTERPRETER: Competent "Reporter" skills, and now consistently integrates and interprets data in reasonable fashion. Good fund of knowledge. Appropriately prioritizes patient problems. Thoughtful development and defense of differential diagnoses based on data.
	INTERPRETER/MANAGER: Starting to offer reasonable diagnostic and therapeutic plans based on interpretation of data, but not yet on a consistent basis.
	MANAGER: Competent "Reporter" and "Interpreter" skills and now consistently offers reasonable and thoughtful management plans and appropriately adjusts plans in response to new incoming data. Excellent fund of knowledge with good applicability to patient care.
	MANAGER/EDUCATOR: Starting to pose insightful questions and search all available sources for answers, but not yet on a consistent basis
	EDUCATOR: Accomplished "Reporter", "Interpreter", and "Manager" skills. Outstanding fund of knowledge and clinical skills with exceptional self-directed learning traits. Consistently poses insightful questions and highly motivated to expand knowledge and share knowledge with others.
	N/A

PROCEDURAL KNOWLEDGE AND SKILLS

(Question 9 of 13 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Lacks knowledge of risks/benefits/limitations of medical procedures. No improvement even with coaching.	Limited understanding of relevant characteristics of medical procedures. Awkward, reluctant to try even basic procedures; insensitive.	Generally has basic understanding of relevant characteristics of medical procedures. Reasonable skills.	Good understanding of relevant characteristics of medical procedures. Proficient, compassionate.	Thorough understanding of relevant characteristics of medical procedures. Highly proficient and skillful.	0
Procedural Knowledge and Skills	1.0	2.0	3.0	4.0	5.0	0

OVERALL SCORE

(Question 10 of 13 - Mandatory)

	1	2	3	4	5	6	7	8	9
	Unacceptable		Below Expectations		Meet Expectations		Above Expectations		Exceptional
Overall Clinical Performance Score	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0

COMMENTS (required for the student's final grade sheet) Please comment on this student's performance based on all of the above categories. Indicate any significant areas of strength and areas needing improvement demonstrated by this student.

General Comments: *(Question 11 of 13 - Mandatory)*

Areas for Improvement: *(Question 12 of 13)*

Summation of Contact Time with Student: *(Question 13 of 13)*

Selection	Option
	Please Select One
	<1 Day
	1-7 Days
	8-14 Days
	>14 Days

Figure A-1. Evaluation of student clerkship performance ('medium' form)

APPENDIX B: EVALUATION OF STUDENT CLERKSHIP PERFORMANCE ('LONG' FORM)

Subject:
Evaluator:
Site:
Period:
Dates of Activity:
Activity: Internal Medicine Clerkship Evansville
Form: Educator of Medical Student

Please evaluate this student's clinical performance compared to other students you have supervised (during a similar time in the academic year). Using the Likert scale with the representative descriptive anchors below, please check the most appropriate box, or N/A (not applicable/not observed) next to each category listed. Comments are required for all evaluations.

PROFESSIONAL ATTRIBUTES

(Question 1 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Insensitive or inattentive to patients' needs or feelings. More concerned with own interests rather than patients' or team's. Routinely fails to recognize and act to resolve conflicts of interest.	Occasionally does not subordinate own self-interests. Occasionally fails to recognize and act to resolve conflicts of interest.	Puts patients' needs before his/her own. Promotes the common good of the team above self. Capable of recognizing most conflicts of interest but does not always resolve them.	Strong patient advocate. Goes beyond requirements of expected service to patients and team members. Recognizes conflicts of interest and is often able to resolve them satisfactorily.	Outstanding patient advocate; endures inconveniences to meet patients' and/or team's needs. Shares credit with other team members for work achieved. Deals effectively with conflicts of interest.	
Altruism	1.0	2.0	3.0	4.0	5.0	0

(Question 2 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Unreliable, irresponsible. Unexcused attendance and/or tardiness.	Occasionally unprepared or unwilling to take on responsibility, lackadaisical. Unreliable in being accountable for own errors and shortcomings.	Reliable, punctual, fulfills responsibilities. Ensures proper transfer of patient care responsibilities. Accountable for own errors and shortcomings.	Very dependable. Seeks additional responsibilities. Devises and implements plans to correct own errors and shortcomings.	Exceptionally conscientious. Remarkable dedication to patient care; checks on patients after hours.	
Duty: Responsibility/Accountability	1.0	2.0	3.0	4.0	5.0	0

(Question 3 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Demeaning or condescending towards patients. Refuses to deal with patient distress. Avoids talking with patients and their families. Lacks effective communication skills.	Some deficiency of empathy and compassion for patients. Often does not elicit or identify patient's barriers to care. Communication skills need improvement.	Generally courteous towards patients. Appropriate communication skills. Generally elicits and identifies at least one barrier to care.	Strong empathic skills. Makes effort to seek out and talk with patients and their families. Relatively advanced communication skills. Routinely elicits and identifies most of patient's barriers to care.	Deals with sickness, death and dying in a highly professional and effective manner with patients and their families. Highly advanced communication skills; easily adjusts communication style to each individual patient or situation.	
Caring, Compassion, & Communication	1.0	2.0	3.0	4.0	5.0	0

(Question 4 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Unaware of own performance deficiencies. Not receptive to and lack of improvement with feedback. Lacks self-motivation to expand knowledge and skills. Maladaptive, inappropriate coping.	Defensive or resistant to feedback; inconsistent improvement. Occasionally loses composure.	Accepts feedback when offered and generally improves. Generally makes appropriate adjustments in response to stress.	Seeks and is receptive to feedback, and consistently improves with feedback. Adapts well in response to stress.	Consistently self-reflects. Outstanding in soliciting, receiving, and incorporating feedback to improve. Excellent self-motivation and initiative to expand knowledge and skills. Highly adaptive in stressful situations.	
Excellence through Self-awareness	1.0	2.0	3.0	4.0	5.0	0

(Question 5 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Poorly integrated into the team. Antagonistic or disruptive. Disrespectful of others. Unprofessional attire.	Occasional difficulty in relating to health care team members. Occasional lapses in use of professional language or in professional appearance.	Cooperative, productive team member. Uses professional language with patients and colleagues. Appropriate professional attire.	Relates well to health care team members and functions well on team.	Outstanding in respecting the feelings, needs, rights, and diversity of team members and other co-workers. Highly integrated team member.	
Respect for Others: Working Relationships	1.0	2.0	3.0	4.0	5.0	0

(Question 6 of 27 - Mandatory)

	No	Yes	N/A
Honesty & Integrity: This student always demonstrated honesty and integrity in all his/her interactions with patients and their families, faculty, colleagues, and others.	1.0	2.0	3.0

If a LOW SCORE was given on any of the previous questions in this section, please explain.

Comments on Professional Attributes
(Question 7 of 27)

DATA ACQUISITION

(Question 8 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Inaccurate, major omissions. Unresourceful. Inefficient.	Incomplete or unfocused; often omits pertinent psychosocial factors. Chronology out of order.	Obtains basic history, including standard psychosocial factors; accurate.	Thorough, yet focused; can appreciate certain subtleties, including pertinent psychosocial factors. Clarifies sequence of events.	Resourceful, searches other sources for information, comprehensive, appreciates subtleties, efficient, insightful.	
History taking skill	1.0	2.0	3.0	4.0	5.0	0

(Question 9 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Unreliable, major omissions.	Incomplete; occasionally misses major findings.	Identifies major findings.	Systematic approach, organized, focused. Can elicit some subtle findings.	Thorough, but appropriately focused. Routinely elicits subtle findings.	
Physical examination skill	1.0	2.0	3.0	4.0	5.0	0

If a LOW SCORE was given on any of the previous questions in this section, please explain.

Comments on Data Acquisistion
(Question 10 of 27)

DATA ORGANIZATION & REPORTING

(Question 11 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Inaccurate data or major omissions. Reflects poor understanding of disease process and patient's psychosocial situation.	Lacks supporting detail, unfocused, needs organization. Incomplete understanding of disease process and patient's psychosocial situation.	Generally accurate and complete. Reflects basic understanding of disease process and patient's psychosocial situation.	Complete, appropriately focused, and organized.	Complete, but concise. Reflects thorough understanding of disease process and patient's psychosocial situation.	
Written Histories & Physicals	1.0	2.0	3.0	4.0	5.0	0

(Question 12 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Inaccurate data or major omissions. Clinical assessments have major gaps.	Needs organization, relevant data and/or patient problems occasionally missing. Clinical assessments incomplete.	Generally accurate and complete. Clinic notes appropriately focused.	Complete, but concise, organized. Problem list consistently and appropriately updated.	Thorough. Identifies and appropriately prioritizes new and even subtle patient problems. Thoughtful clinical assessments.	0
Progress Notes/Clinic Notes	1.0	2.0	3.0	4.0	5.0	0

(Question 13 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Consistently ill-prepared, disorganized, omits key data. Easily distracted or flustered.	Occasionally omits key data or includes irrelevant facts, rambles, needs organization.	Generally maintains format, includes all basic information, usually focused.	Fluent, focused, complete.	Poised, able to adjust length according to situation (type of rounds, time limitations, etc.) without compromising content.	0
Oral Presentations	1.0	2.0	3.0	4.0	5.0	0

If a LOW SCORE was given on any of the previous questions in this section, please explain.

Comments on Data Organization & Reporting
(Question 14 of 27)

KNOWLEDGE BASE AND EDUCATIONAL INITIATIVE

(Question 15 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Lacks knowledge to understand patient problems.	Inconsistent understanding of patient problems. Limited differential diagnoses. Needs prompting to read and learn about own patients.	Knows basic differential diagnoses of patients' active problems. Reads appropriately.	Expanded differential diagnoses. Demonstrates extra reading about patients' problems. Searches literature for best evidence.	Self-directed learner; routinely poses insightful questions and effectively searches literature for best evidence. Differential diagnoses are expanded and well-prioritized.	0
Knowledge Base and Educational Initiative	1.0	2.0	3.0	4.0	5.0	0

If a LOW SCORE was given, please explain.

Comments on Knowledge Base and Educational Initiative
(Question 16 of 27)

DATA INTERPRETATION & INTEGRATION

(Question 17 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Cannot integrate and interpret basic data.	Often reports data without analysis; often does not consider pertinent psychosocial factors. Problem list incomplete or unprioritized. Often unable to defend differential diagnoses.	Integrates and interprets data, including psychosocial factors, at a basic level. Generally appropriate problem list & differential diagnoses.	Thoughtful integration and interpretation of data, including pertinent psychosocial factors. Consistently able to defend differential diagnoses.	Highly thoughtful integration of data to identify, prioritize, and solve patient problems. Understands complex issues, interrelates patient problems.	0
Data Interpretation & Integration	1.0	2.0	3.0	4.0	5.0	0

If a LOW SCORE was given, please explain.

Comments on Data Interpretation & Integration
(Question 18 of 27)

CLINICAL JUDGMENT/MANAGEMENT

(Question 19 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Fails to recognize, identify, or prioritize urgent problems. Makes inappropriate diagnostic and therapeutic decisions. Lack of knowledge of routine screening. Unable to identify patient risk factors.	Inconsistent prioritization of clinical issues and decision-making ability. Often does not consider patient's psychosocial factors in decision making.	Appropriate basic diagnostic and therapeutic decision-making, taking into consideration patient's psychosocial factors. Performs appropriate health screening and usually identifies/addresses health risk factors. Generally aware of own limitations and seeks appropriate help.	Consistently makes sound diagnostic and therapeutic decisions, taking into consideration patient's pertinent psychosocial factors.	Highly thoughtful approach to management plans; consistently integrates new data, adjusts plans accordingly, and carefully weighs all alternatives, including patient's preferences and psychosocial factors. Comprehensive in application of routine health screening; addresses health risks.	0
Clinical Judgment/Management	1.0	2.0	3.0	4.0	5.0	0

If a LOW SCORE was given, please explain.

Comments on Clinical Judgment/Management
(Question 20 of 27)

DEVELOPMENTAL ASSESSMENT *(Question 21 of 27 - Mandatory)*

Selection	Option
	INCONSISTENT REPORTER: Performing consistently at the "Reporter" level is the minimum requirement for clinical clerkships.
	REPORTER: Consistently acquires and communicates clinical information accurately; this includes properly identifying patient problems and constructing appropriate problem lists. Also consistently demonstrates satisfactory professional behavior.
	REPORTER/INTERPRETER: Starting to integrate and interpret the collected data to develop reasonable differential diagnoses, but not yet on a consistent basis.
	INTERPRETER: Competent "Reporter" skills, and now consistently integrates and interprets data in reasonable fashion. Good fund of knowledge. Appropriately prioritizes patient problems. Thoughtful development and defense of differential diagnoses based on data.
	INTERPRETER/MANAGER: Starting to offer reasonable diagnostic and therapeutic plans based on interpretation of data, but not yet on a consistent basis.
	MANAGER: Competent "Reporter" and "Interpreter" skills and now consistently offers reasonable and thoughtful management plans and appropriately adjusts plans in response to new incoming data. Excellent fund of knowledge with good applicability to patient care.
	MANAGER/EDUCATOR: Starting to pose insightful questions and search all available sources for answers, but not yet on a consistent basis
	EDUCATOR: Accomplished "Reporter", "Interpreter", and "Manager" skills. Outstanding fund of knowledge and clinical skills with exceptional self-directed learning traits. Consistently poses insightful questions and highly motivated to expand knowledge and share knowledge with others.
	N/A

PROCEDURAL KNOWLEDGE AND SKILLS

(Question 22 of 27 - Mandatory)

	Unacceptable	Below expectations	Meet expectations	Above expectations	Exceptional	N/A
	Lacks knowledge of risks/benefits/limitations of medical procedures. No improvement even with coaching.	Limited understanding of relevant characteristics of medical procedures. Awkward, reluctant to try even basic procedures; insensitive.	Generally has basic understanding of relevant characteristics of medical procedures. Reasonable skills.	Good understanding of relevant characteristics of medical procedures. Proficient, compassionate.	Thorough understanding of relevant characteristics of medical procedures. Highly proficient and skillful.	0
Procedural Knowledge and Skills	1.0	2.0	3.0	4.0	5.0	0

If a LOW SCORE was given, please explain.

Comments on Procedural Knowledge and Skills *(Question 23 of 27)*

OVERALL SCORE

(Question 24 of 27 - Mandatory)

	1	2	3	4	5	6	7	8	9
	Unacceptable		Below Expectations		Meet Expectations		Above Expectations		Exceptional
Overall Clinical Performance Score	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0

COMMENTS (required for the student's final grade sheet) Please comment on this student's performance based on all of the above categories. Indicate any significant areas of strength and areas needing improvement demonstrated by this student.

General Comments: *(Question 25 of 27 - Mandatory)*

Areas for Improvement: *(Question 26 of 27)*

Summation of Contact Time with Student: *(Question 27 of 27)*

Selection	Option
	<1 Day
	1-7 Days
	8-14 Days
	>14 Days

Figure B-1. Evaluation of student clerkship performance ('long' form)

APPENDIX C: THE ELECTRONIC STUDY SURVEY

Resize font:

Measuring Your Perceptions of Clinical Clerkship Evaluations

The purpose of this survey is to gather your perceptions on the experience of evaluating a medical student's clinical performance via a clinical clerkship evaluation form. This survey is part of a larger project aimed at determining how physician evaluative load and evaluation strain may be associated with clerkship evaluation data.

By completing this survey, you agree to allow the research team to access records of clinical clerkship evaluations you have completed during the 2015-2016 academic year. No one outside of the research team will have access to these files, nor will your data be used for any other purpose.

Submission of this survey will qualify you for a raffle to win a $100 Amazon.com gift card!

Thank you for your participation!

Research Team: Courtney Traser, B.S., James J. Brokaw, Ph.D., Adam B. Wilson, Ph.D., Ronald L. Shew, Ph.D., Gary R. Pike, Ph.D., Melissa R. Alexander, Ph.D.

Please provide the following demographic information.

First and Last Name:

* must provide value

Which IUSM medical department are you affiliated with?

- ○ Family Medicine
- ○ General Surgery
- ○ Internal Medicine
- ○ OB/GYN
- ○ Pediatrics
- ○ Psychiatry

reset

What is your academic rank?

- ○ Instructor/Lecturer
- ○ Assistant Professor
- ○ Associate Professor
- ○ Full Professor
- ○ Resident
- ○ Fellow

reset

What is your gender?

- ○ Male
- ○ Female
- ○ Prefer not to answer

reset

To what extent do you feel overwhelmed by the number of clinical clerkship evaluations you are asked to complete on a regular basis?

- ○ I never feel overwhelmed.
- ○ I sometimes feel overwhelmed.
- ○ I often feel overwhelmed.
- ○ I very often feel overwhelmed.

reset

Below are six rating scales designed to gauge your experience in completing a clinical clerkship evaluation. Please assess the evaluative process by moving the 'slide bar' on each of the six scales to the point which best represents your experience. The scales range from "very low" on the left to "very high" on the right (or something similar). Please read the descriptions carefully.

Mental Demand: How mentally demanding is it to complete clerkship evaluations?

Very Low Moderate Very High

Click bar above and then drag to set response

reset

Physical Demand: How physically fatiguing is it to complete clerkship evaluations?

Very Low Moderate Very High

Click bar above and then drag to set response

reset

Temporal Demand: How hurried or rushed do you feel while completing clerkship evaluations?

Very Low Moderate Very High

Click bar above and then drag to set response

reset

Task Complexity: How complex is the task of completing clerkship evaluations?

Not Very Complex Somewhat Complex Very Complex

Click bar above and then drag to set response

reset

Situational Stress: How anxious do you feel when completing clerkship evaluations?

Not Very Anxious Somewhat Anxious Very Anxious

Click bar above and then drag to set response

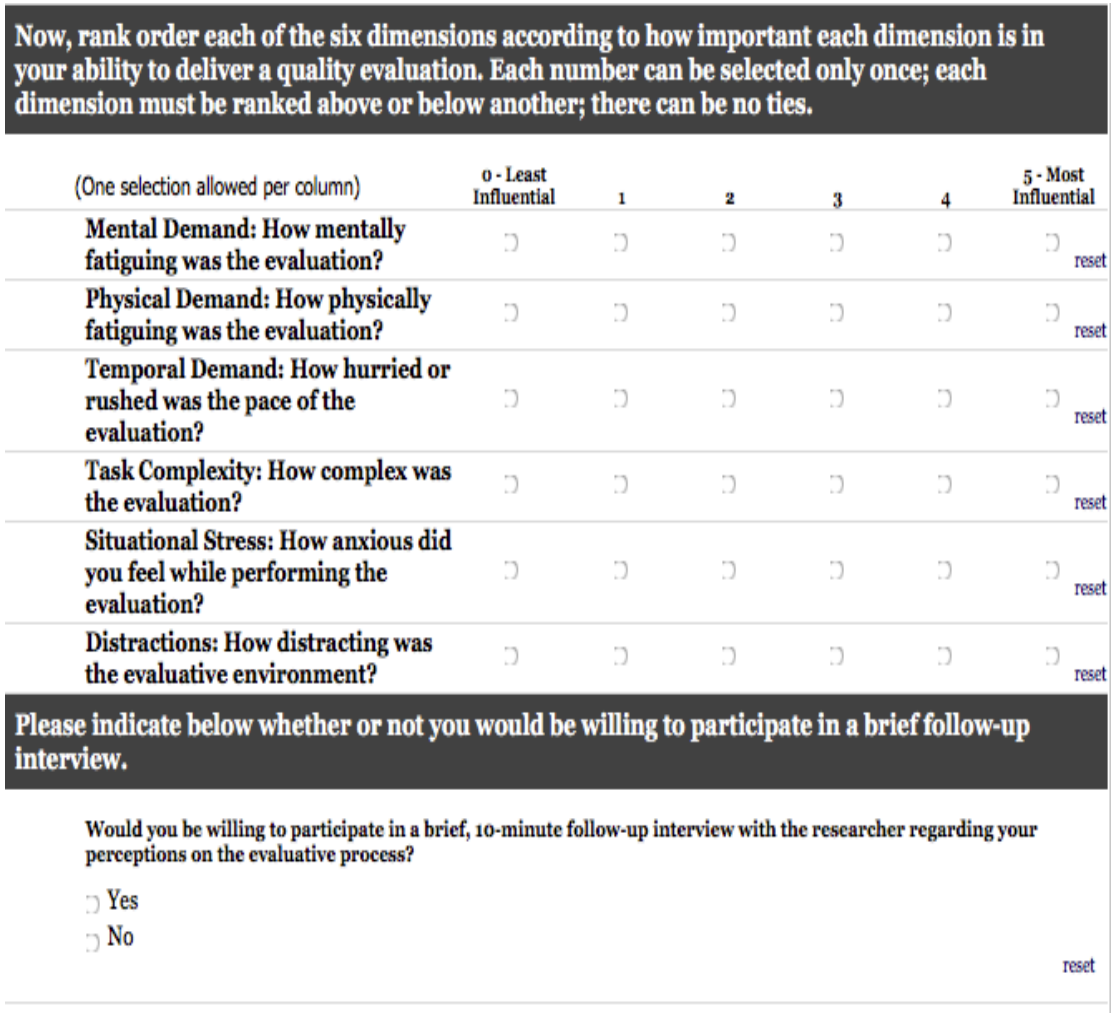
reset

Distractions: How distracting is the evaluative environment when you complete clerkship evaluations?

Not Very Distracting Somewhat Distracting Very Distracting

Click bar above and then drag to set response

reset



Now, rank order each of the six dimensions according to how important each dimension is in your ability to deliver a quality evaluation. Each number can be selected only once; each dimension must be ranked above or below another; there can be no ties.

(One selection allowed per column)	0 - Least Influential	1	2	3	4	5 - Most Influential
Mental Demand: How mentally fatiguing was the evaluation?	○	○	○	○	○	○ reset
Physical Demand: How physically fatiguing was the evaluation?	○	○	○	○	○	○ reset
Temporal Demand: How hurried or rushed was the pace of the evaluation?	○	○	○	○	○	○ reset
Task Complexity: How complex was the evaluation?	○	○	○	○	○	○ reset
Situational Stress: How anxious did you feel while performing the evaluation?	○	○	○	○	○	○ reset
Distractions: How distracting was the evaluative environment?	○	○	○	○	○	○ reset

Please indicate below whether or not you would be willing to participate in a brief follow-up interview.

Would you be willing to participate in a brief, 10-minute follow-up interview with the researcher regarding your perceptions on the evaluative process?

○ Yes
○ No

reset

Figure C-1. Electronic study survey

APPENDIX D: QUALITY OF CLINICAL CLERKSHIP EVALUATION (CCE) RUBRIC

Table D-1. Quality of CCE rubric

'Quality of Feedback' Quality Measures	**Eval 1**	**Eval 2**	**Eval 3**	**Eval 4**	**Eval 5**	**Sum**
Diagnostic comment: a comment based on observable behaviors and/or skills						
Formative comment: a two-part comment suggesting specific areas of strengths and/or weaknesses *with* a clear explanation of how to improve or what was done well						
Specific comment: a two-part comment that was both not globally descriptive, like "did well," and uniquely formulated for the evaluated student						
Practical comment: a comment that was thorough, useful, and clearly actionable						
					Total	**/4**

APPENDIX E: INTERVIEW PROTOCOL

1. What value do you ascribe to clinical clerkship evaluations?
    a. Are they important to you?
    b. Are they important to the students?
2. How confident are you that the information you are communicating to students on these evaluations is actionable?
3. Have you heard of any 'grade inflation' on the evaluations?
4. How would you describe the 'quality' of the evaluations you produce?
5. What would you change about the evaluative process, if anything?
6. How easy is it to complete the evaluations?
7. How do you conceptualize your evaluative load? Do you think that you have a lot of evaluations to complete, more so than your colleagues or those from other departments?
8. In the survey you completed, you were asked to describe how six dimensions affected your experience of completing these evaluations. These were mental, physical, and temporal demands; task complexity; situational stress; and distractions. Are there any other factors or influences that you believe contribute to or hinder your completion of these evaluations?
9. Of these six dimensions, you ranked _____ as the most influential dimension and _____ as the least influential. Can you describe why?
10. What is the most noteworthy aspect of the evaluation, in your opinion?

APPENDIX F: HISTOGRAM ILLUSTRATING PARTICIPANTS' SCORES ACROSS THE CLOSED-ENDED CCE ITEMS

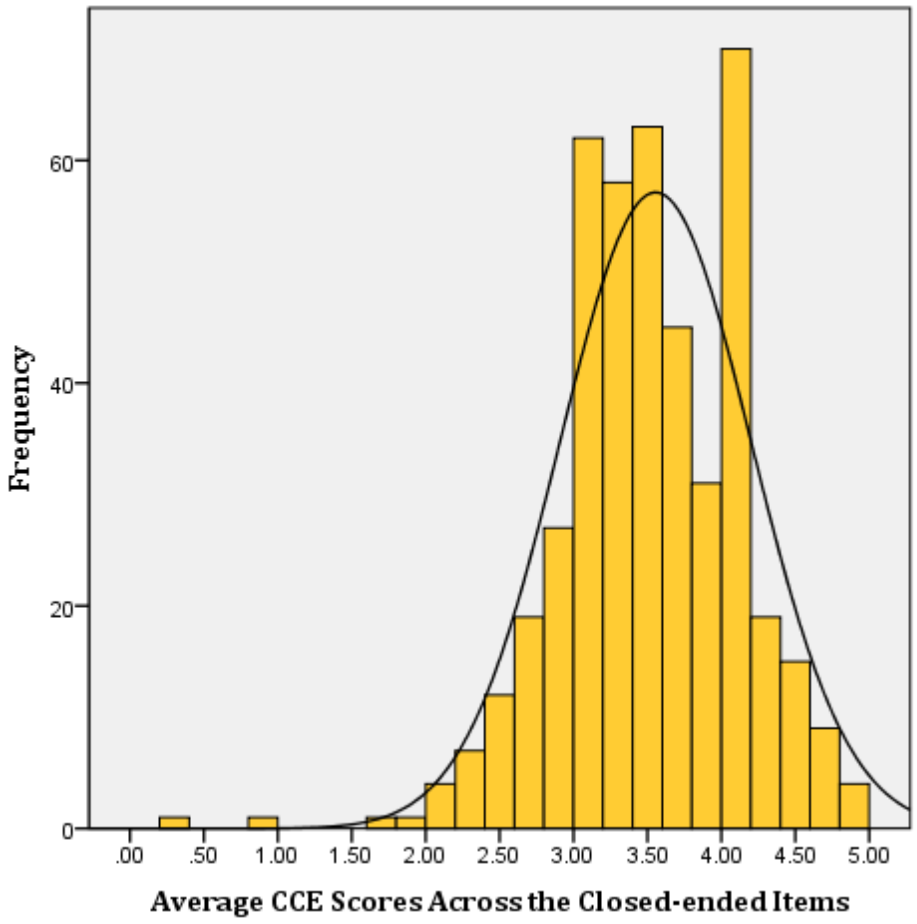


Figure F-1. Histogram illustrating participants' scores across the closed-ended items

REFERENCES

1. Nutter D, Whitcomb M. The AAMC project on the clinical education of medical students. Washington, DC: Association of American Medical Colleges. 2001.

2. Corbett EC, Whitcomb M. The AAMC project on the clinical education of medical students: Clinical skills education. Washington, DC: Association of American Medical Colleges. 2004.

3. Weinberger S, Whitcomb M. The clinical education of medical students: Report on millennium conferences I & II. Washington, DC: Association of American Medical Colleges. 2002.

4. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387-396.

5. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15(4):270-292.

6. Training and certification: Resident performance assessments. American Board of Surgery. Accessed January 3, 2016. Available: http://www.absurgery.org/default.jsp?aboutcontact

7. Bates J, Konkin J, Suddards C, et al. Student perceptions of assessment and feedback in longitudinal integrated clerkships. Med Educ. 2013;47(4):362-374.

8. Sinead OD, Deidre M, Walter Cullen G. How do longitudinal clerkships in general practice/primary care impact on student experience and career intention? A cross-sectional study of student experience. Educ Prim Care. 2015;26(3):166-175.

9. Dubé TV, Schinke RJ, Strasser R, et al. Transition processes through a longitudinal integrated clerkship: A qualitative study of medical students' experiences. Med Educ. 2015;49(10):1028-1037.

10. R. Kogan J, Lapin J, Aagaard E, et al. The Effect of resident duty-hours restrictions on internal medicine clerkship experiences: Surveys of medical students and clerkship directors. Teach Learn Med. 2015;27(1):37-50.

11. Katowa-Mukwato P, Andrews B, Maimbolwa M, et al. Medical students' clerkship experiences and self-perceived competence in clinical skills. Afr J Health Prof Educ. 2014;6(2):155-160.

12. McLaughlin K, Vitale G, Coderre S, et al. Clerkship evaluation: What are we measuring? Med Teach. 2009;31(2):e36-39.

13. Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. Eval Health Prof. 2007;30(3):266-283.

14. Farrell TM, Kohn GP, Owen SM, et al. Low correlation between subjective and objective measures of knowledge on surgery clerkships. J Am Coll Surg. 2010;210(5):680-685.

15. Goldstein SD, Lindeman B, Colbert-Getz J, et al. Faculty and resident evaluations of medical students on a surgery clerkship correlate poorly with standardized exam scores. Am J Surg. 2014;207(2):231-235.

16. Dudas RA, Colbert JM, Goldstein S, et al. Validity of faculty and resident global assessment of medical students' clinical knowledge during their pediatrics clerkship. Acad Pediatr. 2012;12(2):138-141.

17. Awad SS, Liscum KR, Aoki N, et al. Does the subjective evaluation of medical student surgical knowledge correlate with written and oral exam performance? J Surg Res. 2002;104(1):36.

18. Oaks WW, Scheinok PA, Husted FL. Objective evaluation of a method of assessing student performance in a clinical clerkship. J Med Educ. 1969;44(3):207-213.

19. Hull AL. Medical student performance: A comparison of house officer and attending staff as evaluators. Eval Health Prof. 1982;5(1):87-94.

20. Kreiter CD, Ferguson K, Lee W-C, et al. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73.

21. Saguil A, Balog EK, Goldenberg MN, et al. The association between specialty match and third-year clerkship performance. Mil Med. 2012:47-52.

22. Hemmer PA, Hawkins R, Jackson JL, et al. Assessing how well three evaluation methods detect deficiencies in medical students' professionalism in two settings of an internal medicine clerkship. Acad Med. 2000;75(2):167-173.

23. Hemmer PA, Pangaro L. The effectiveness of formal evaluation sessions during clinical clerkships in better identifying students with marginal funds of knowledge. Acad Med. 1997;72(7):641-643.

24. Plymale MA, Donnelly MB, Lawton J, et al. Faculty evaluation of surgery clerkship students: Important components of written comments. Acad Med. 2002;77(10 Suppl):S45-47.

25. Pulito AR, Donnelly MB, Plymale M. Factors in faculty evaluation of medical students' performance. Med Educ. 2007;41(7):667-675.

26. Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: Rater performance and behavior when considering multiple competencies. Teach Learn Med. 2016;28(1):41-51.

27. Govaerts MJ, Van de Wiel MW, Schuwirth LW, et al. Workplace-based assessment: Raters' performance theories and constructs. Adv Health Sci Educ Theory Pract. 2013;18(3):375-396.

28. Kogan JR, Conforti L, Bernabeo E, et al. Opening the black box of clinical skills assessment via observation: A conceptual model. Med Educ. 2011;45(10):1048-1060.

29. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: Rethinking the etiology of rater errors. Acad Med. 2011;86(10 Suppl):S1-7.

30. Norcini JJ. Current perspectives in assessment: The assessment of performance at work. Med Educ. 2005;39:880-889.

31. Undergraduate medical education. Accessed Febuary 24, 2016. Available: http://www.evaluehealthcare.com/who-we-serve/undergraduate-medical-education

32. Gauthier G, St-Onge C, Tavares W. Rater cognition: Review and integration of research findings. Med Educ. 2016;50:511-522.

33. Hart SG. NASA-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet. 2006:904-908.

34. Hart SG, Staveland LE. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In PA Hancock and N. Meshkati (Eds). Human mental workload: Advances in psychology. Oxford, England: North-Holland. 1988:139-183.

35. Tavares W, Eva KW. Impact of rating demands on rater-based assessments of clinical competence. Education Prim Care. 2014;25(6):308-318.

36. Melchers KG, Kleinmann M, Prinz MA. Do assessors have too much on their plates? The effects of simultaneously rating multiple assessment center candidates on rating quality. Int J Select Assess. 2010;18(3).

37. Williams RG, Chen XP, Sanfey HA, et al. The measured impact of delay in completing operative performance ratings on clarity and detail of ratings assigned. J Surg Educ. 2014;71(6):e132-8.

38. Stahnisch FW, Verhoef M. The flexner report of 1910 and its impact on complementary and alternative medicine and psychiatry in North America in the 20th century. Evid Based Complement Alternat Med. 2012.

39. Hubble D. William Osler and medical education. J R Coll Physicians Lond. 1975;9(3):269-278.

40. Physicians for the twenty-first century: Report of the project panel on the general professional education of the physician and college preparation for medicine. J Med Educ. 1984;59(11):1-208.

41. Gastel, B, Rogers, DE. Adapting clinical medical education to the needs of today and tomorrow: Proceedings of the Josiah Macy, Jr. Foundation. Bermuda Islands: New York Academy of Medicine. 1988:118.

42. Report I: Learning objectives for medical students education: Guidelines for medical schools. Washington, DC: Association of American Medical Colleges. 1998.

43. Ouyang W, Cuddy M, Swanson D. U.S. medical student performance on the NBME subject examination in internal medicine: Do clerkship sequence and clerkship length matter? J Gen Intern Med. 2015;30(9):1307-1312.

44. Rotations and electives. Accessed February 23, 2016. Available: http://www.aafp.org/medical-school-residency/medical-school/rotations.html

45. Percentage of medical schools with separate required clerkships by discipline. Accessed February 23, 2016. Available: https://www.aamc.org/initiatives/cir/406450/05a.html

46. Sharp LM, Frankel J. Respondent burden: A test of some common assumptions. Public Opin Q. 1983;47(1):36-53.

47. Porter SR, Whitcomb ME, Weizter WH. Multiple surveys of students and survey fatigue. New Dir Inst Res. 2004:120.

48. Apodaca R, Lea S, Edwards B. The effect of longitudinal burden on survey participation. Proc Amer Assoc Pub Opin Res Annu Meet. 1998.

49. Sosdian CP, Sharp LM. Nonresponse in mail surveys: Access failure or respondent resistance. Public Opin Q. 1980;44(3):396-402.

50. Asiu BW, Antons CM, Fultz ML. Undergraduate perceptions of survey participation: Improving response rates and validity. Proc Assoc Inst Res Annu Meet. 1998.

51. Revilla M, Ochoa C. What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? Soc Sci Comput Rev. 2015;33(1):97-114.

52. Schaeffer NC, Presser S. The science of asking questions. Annu Rev Sociol. 2003;29:65-88.

53. Mavis BE, Cole BL, Hoppe RB. A survey of student assessment in U.S. medical schools: The balance of breadth versus fidelity. Teach Learn Med. 2001;13(2):74-79.

54. Pulito AR, Donnelly MB, Plymale M, et al. What do faculty observe of medical students' clinical performance? Teach Learn Med. 2006;18(2):99-104.

55. Howley LD, Wilson WG. Direct observation of students during clerkship rotations: A multiyear descriptive study. Acad Med. 2004;79(3):276-280.

56. Chisholm CD, Whenmouth LF, Daly EA, et al. An evaluation of emergency medicine resident interaction time with faculty in different teaching venues. Acad Emerg Med. 2004;11(2):149-155.

57. Cooper D, Beswick W, Whelan G. Intensive bedside teaching of physical examination to medical undergraduates: Evaluation including the effect of group size. Med Educ. 1983;17(5):311-315.

58. Reichsman F, Browning FE, Hinshaw JR. Observations of undergraduate clinical teaching in action. J Med Educ. 1964;39:147-163.

59. An analysis of U.S. student field trial and international medical graduate certification testing results for the proposed USMLE clinical skills examination. Accessed February 25, 2016.

60. The role of faculty observation in assessing students' clinical skills. Washington, DC: Association of American Medical Colleges. 1997.

61. Burdick WP, Schoffstall J. Observation of emergency medicine residents at the bedside: How often does it happen? Acad Emerg Med. 1995;2(10):909-913.

62. Ivankova NV, Creswell JW, Stick SL. Using mixed-methods sequential explanatory design: From theory to practice. Field Methods. 2006;18(3):3-20.

63. Creswell JW. Research design: Qualitative, quantitative, and mixed methods approaches. Los Angeles, CA: SAGE Publications, Inc. 2014.

64. Tashakkori A, Teddlie C. Introduction to mixed method and mixed model studies in the social and behavioral sciences. Mixed methodology: Combining qualitative and quantitative approaches. Thousand Oaks, CA: SAGE Publications, Inc. 1998:3-19.

65. Teddlie C, Tashakkori A. Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A Tashakkori, C Teddlie (Eds). Handbook on mixed methods in the behavioral and social sciences. Thousand Oaks, CA: SAGE Publications, Inc. 2003:3-50.

66. Creswell JW, Plano Clark VL, Gutmann ML, et al. Advanced mixed methdos research designs. In A Tashakkori, C Teddlie (Eds). Handbook of mixed methods in social and behavioral research. Thousand Oaks, CA: SAGE Publications, Inc. 2003:209-240.

67. Morgan D. Practical strategies for combining qualitative and quantitative methods: Applications to health research. Qual Health Res. 1998;8:362-376.

68. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377-381.

69. Wilson MR, Poolton JM, Malhotra N, et al. Development and validation of a surgical workload measure: The surgery task load index (SURG-TLX). World J Surg. 2011;35(9):1961-1969.

70. Pauzie A. A method to assess the driver mental workload: The driving activty load index (DALI). IET Intell Transp Syst. 2008;2:315-322.

71. Wetzel CM, Kneebone RL, Woloshynowych M, et al. The effects of stress on surgical performance. Am J Surg. 2006;191(1):5-10.

72. Moore LJ, Wilson MR, McGrath JS, et al. Surgeons display reduced mental effort and workload while performing robotically assisted surgical tasks, when compared to conventional laparoscopy. Surg Endosc. 2015;29(9):2553-2560.

73. Weigl M, Antoniadis S, Chiapponi C, et al. The impact of intra-operative interruptions on surgeons' perceived workload: An observational study in elective general and orthopedic surgery. Surg Endosc. 2015;29(1):145-153.

74. Berg RJ, Inaba K, Sullivan M, et al. The impact of heat stress on operative performance and cognitive function during simulated laparoscopic operative tasks. Surgery. 2015;157(1):87-95.

75. Malhotra N, Poolton JM, Wilson MR, et al. Conscious motor processing and movement self-consciousness: Two dimensions of personality that influence laparoscopic training. J Surg Educ. 2014;71(6):798-804.

76. Singh R, Carranza D, Morrow MM, et al. 3: Effect of different chairs on work-related musculoskeletal discomfort during vaginal surgery. Am J Obstet Gynecol. 2016;214(4):S456-S457.

77. Cole JS, McCormick AC, Gonyea RM. Respondent use of straight-lining as a response strategy in education survey research: Prevalence and implications. Proc Amer Educ Res Assoc Annu Meet. 2012.

78. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol. 1991;5:213-236.

79. Ovando MN. Constructive feedback: A key to successful teaching and learning. Int J Educ Manage.1992:12.

80. Keith TZ. Multiple regression and beyond: An introduction to muliple regression and structural equation modeling. New York, NY: Routledge. 2015.

81. Krosnick JA, Alwin DF. A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. Pub Opin Q. 1988;52:526-538.

82. Williams RG, Verhulst S, Colliver JA, et al. A template for reliable assessment of resident operative performance: Assessment intervals, numbers of cases and raters. Surgery. 2012;152(4):517-524.

83. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34.

84. Graham M, Milanowski A, Miller J. Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Washington, DC: Center for Educator Compensation Reform. 2012.

85. Mertler CA, Vannatta RA. Advanced and multivariate statistical methods: Practical application and interpretation. Los Angeles, CA: Pyrczak Publishing. 2002.

86. Olobatuyi ME. A user's guide to path analysis. Lanham, MD: University Press of America, Inc. 2006.

87. Cohen J. Statistical Power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. 1988.

88. Merriam SB. Qualitative research: A guide to design and implementation. San Francisco, CA: Jossey-Bass. 2009.

89. Kennedy T, Lingard L. Making sense of grounded theory in medical education. Med Educ. 2006;40(2):101-108.

90. Tabachnick BG, Fidell LS. Using multivariate statistics. New York, NY: Harper Collins. 1996.

91. Campbell DJ. Task complexity: A review and analysis. Acad Manage Rev. 1988;13(1):40-52.

92. Braarud PØ. Subjective task complexity and subjective workload: Criterion validity for complex team tasks. Int J Cogn Ergon. 2001;5(3):261-273.

93. Bowen RE, Grant WJ, Schenarts KD. The sum is greater than its parts: Clinical evaluations and grade inflation in the surgery clerkship. Am J Surg. 2015;209(4):760-764.

94. Berendonk C, Stalmeijer R, Schuwirth LT. Expertise in performance assessment: Assessors' perspectives. Adv Health Sci Educ Theory Pract. 2013;18:559-571.

95. Govaerts MJB, Schuwirth LWT, Van der Vleuten CPM, et al. Workplace-based assessment: Effects of rater expertise. Adv in Health Sci Educ. 2011;16:151-165.

96. Colletti LM. Difficulty with negative feedback: Face-to-face evaluation of junior medical student clinical performance resutls in grade inflation. J Surg Res. 2000;90:82-87.

97. Albanese MA. Challenges in using rater judgements in medical education. J Eval Clin Pract. 2000;6(3):305-319.

98. Govaerts MJB, van de Wiel MWJ, van der vleuten CPM. Quality of feedback following performance assessments: Does assessor expertise matter? Eur Jour Train Dev. 2013;37(1):105-125.

99. Brutus S. Words versus numbers: A theoretical exploration of giving and receiving narrative comments in performance appraisal. Hum Res Manage Rev. 2010;20:144-157.

100. Cook DA, Kuper A, DPhil RH, et al. When assessment data are words: Validity evidence for qualitative educational assessments. Acad Med. 2016;91(10):1359-1369.

101. Schmahmann JD, Neal M, MacMore J. Evaluation of the assessment and grading of medical students on a neurology clerkship. Neurology. 2008;70:706-712.

102. Fazio SB, Papp KK, Torre DM, et al. Grade inflation in the internal medicine clerkship: A national survey. Teach Learn Med. 2013;25(1):71-76.

103. Cacamese SM, Elnicki M, Speer AJ. Grade inflation and the internal medicine subinternship: A national survey of clerkship directors. Teach Learn Med. 2007;19(4):343-346.

104. Jawahar IM, Williams CR. Where all the children are above average: The performance appraisal purpose effect. Pers Psychol. 1997;50:905-925.

105. Iris Franz W-J. Grade inflation under the threat of students' nuisance: Theory and evidence. Econ Educ Rev. 2010;29(3):411-422.

106. Borman WC. Job behaviour, performance, and effectiveness. In MD Dunnette, LM Hough (Eds). Handbook of Industrial and Organizational Psychology. Palo Alto, CA: Consulting Psychologist Press. 1991.

107. Wherry RJ, Bartlett CJ. The control of bias in ratings: A theory of rating. Pers Psychol. 1982;35:521-551.

108. Williams RG, Verhulst S, Mellinger JD, et al. Is a single-item operative performance rating sufficient? J Surg Educ. 2015;72(6):e212-217.

109. Ferenchick GS, Solomon D, Foreback J, et al. Mobile technology for the facilitation of direct observation and assessment of student performance. Teach Learn Med. 2013;25(4):292-299.

110. MacCallum RC, Browne MW, Sugawara HM. Power analysis and determination of sample size for covariance structure modeling Psychol Methods. 1996;1:130-149.

111. Kline RB. Principles and practice of structural equation modeling. New York, NY: The Guilford Press. 1998.

112. Maslach C, Leiter MP. New insights into burnout and health care: Strategies for improving civility and alleviating burnout. Med Teach. 2016;39(2):160-163.

113. Maslach C, Leiter MP. The truth about burnout: How organizations cause personal stress and what to do about it. San Francisco, CA: Jossey-Bass. 1997.

114. Seritan AL. How to recognize and avoid burnout. In LW Roberts (Ed). The academic medicine handbook: A guide to achievement and fulfillment for academic faculty. New York, NY, US: Springer Science + Business Media. 2013:447-453.

115. Tayfur O, Arslan M. The role of lack of reciprocity, supervisory support, workload and work–family conflict on exhaustion: Evidence from physicians. Psychol Health Med. 2013;18(5):564-575.

116. Shirom A, Nirel N, Vinokur AD. Work hours and caseload as predictors of physician burnout: The mediating effects by perceived workload and by autonomy. Appl Psychol. 2010;59(4):539-565.

CURRICULUM VITAE

Courtney Jo Traser

EDUCATION

Doctor of Philosophy: Anatomy and Cell Biology June 2017
Minor: Education
Indiana University, Indianapolis, Indiana
Dissertation Title: "Do I *really* have to complete another evaluation?" Exploring Relationships Among Physicians' Evaluative Load, Evaluative Strain, and the Quality of Clinical Clerkship Evaluations

Bachelor of Science: Biology and Spanish 2013
Summa Cum Laude
Clarke University, Dubuque Iowa

TEACHING EXPERIENCE

Indiana University School of Medicine
Medical Human Structure Lecturer 2016

- Created and delivered five lectures on the lower limb.
- Served as a laboratory instructor for labs covering lower limb dissections.
- Aided with laboratory examination set-up.

Functionally-Oriented Human Gross Anatomy Associate Instructor 2015-2016

- Gave three lectures on the lower limb.
- Designed examination question based on lecture material and assisted with the set-up, proctoring, and grading of laboratory examinations.
- Guided graduate students through complete cadaveric dissection and served as a private tutor for students.

Medical Neuroscience and Clinical Neurology Associate Instructor 2015

- Acted as a laboratory instructor for two wet lab sessions.
- Created detailed problem-based worksheets for three small group sessions on the spinal cord, brainstem, and forebrain, respectively.
- Served as an instructor and group leader for three small group sessions.

Medical Human Gross Anatomy Associate Instructor 2015

- Designed and delivered three lectures on the lower limb.
- Performed bi-weekly prosections on cadaveric specimens and assisted medical and physical therapy students with complete cadaveric dissections.
- Set-up, proctored, and graded all laboratory examinations.

Basic Histology Associate Instructor 2014

- Created and delivered one team-based learning (TBL) module on the Integumentary and Endocrine Systems.
- Assisted graduate students with the histological identification of structures using light microscopes.
- Formulated examination questions on the Integumentary and Endocrine Systems and assisted with laboratory examination set-up.

Indiana University – Purdue University Indianapolis (IUPUI)

Biology Tutor for the Athletic Department 2015-2017

- Aided undergraduate athletes with lecture and laboratory materials for various biology courses (e.g., human anatomy, human physiology, concepts of biology I and II, etc.).
- Provided students with study tips and personalized learning strategies.
- Worked collaboratively with athletic advisement staff to promote good study habits among athletes.

Undergraduate Cadaveric Human Anatomy Associate Instructor 2017

- Contributed to course design, including creation of the course syllabus.
- Responsible for laboratory instruction and preparation of prosected cadaveric specimen.
- Assisted with the design, implementation, and conduction of practical examinations.

Human Anatomy Adjunct Instructor in the Department of Biology 2014-2016

- Designed and presented two pre-laboratory lectures on relevant histological and gross anatomical structures for a hybrid-laboratory section.
- Guided undergraduate students through the study of the human body using light microscopes, models, photographs, and non-cadaveric dissections.
- Wrote, proctored, and graded laboratory examinations.

Marian University – College of Osteopathic Medicine

Essential Clinical Anatomy and Development Laboratory Instructor 2016

- Served as the sole laboratory instructor for 6-7 dissection tables consisting of 5-6 medical students, each.
- Helped students with complete cadaveric dissection and emphasized the clinical significance of relevant anatomical structures.
- Aided in practical examination set-up.

Essential Clinical Anatomy and Development Prosector 2013

- Performed thrice weekly prosections for the inaugural medical school class.
- Worked collaboratively with Marian faculty.
- Tutored students outside of course hours using the prosected specimen.

Clarke University
Human Gross Anatomy Teaching Assistant and Course Tutor 2012

- Assisted physical therapy students with complete cadaveric dissection and identification of relevant anatomical structures.
- Assisted the course director with administrative duties, including the set-up, proctoring, and grading of laboratory examinations.
- Held weekly open-lab review sessions for students and served as a tutor for both lecture and laboratory material.

Human Anatomy and Physiology I & II Teaching Assistant and Tutor 2010-2012

- Assisted undergraduate biology, psychology, and other pre-health professional students with the identification of bones, prosected cadaveric structures, and anatomical models.
- Offered bi-weekly review sessions for students outside of structured class time.
- Set-up, proctored, and graded laboratory practical examinations and other course assignments.

INVITED ACADEMIC PRESENTATIONS, LECTURES, OR TALKS

Using Web-based Resources, Discussion Boards, and Mobile Apps in Anatomy Teaching June 2016
Advanced Medical-Level Anatomy Workshop
Hosted by the Division of Biomedical Sciences, Marian University – College of Osteopathic Medicine
Indianapolis, Indiana

"Do I *really* have to complete another evaluation?" Exploring relationships among physicians' evaluative load, evaluative strain, and the quality of clinical clerkship evaluations June 2016
Anatomy Education Seminar
Hosted by the Department of Anatomy and Cell Biology, Indiana University Medical School
Indianapolis, Indiana

Enhancing Instruction with Mobile Apps and Web-based Resources June 2014
Anatomy Education Summer Camp
Hosted by the Department of Anatomy and Cell Biology, Indiana University Medical School
Indianapolis, Indiana

Investigating the Use of Quick Response (QR) Codes in the Gross Anatomy Laboratory April 2014
5th Annual Spring Colloquium Series
Hosted by the Center for Urban and Multicultural Education, Indiana University – Purdue University Indianapolis
Indianapolis, Indiana

PUBLISHED SCHOLARLY WORKS & PHOTOGRAPHS

Brokaw, J. J., Byram, J. N., **Traser, C. J.** and Arbor, T. C. (2016). How the distinctive cultures of osteopathic and allopathic medical schools affect the careers, perceptions, and institutional efforts of their anatomy faculties: A qualitative case study of two schools. Anat Sci Ed., 9: 255-264.

Traser C, Condon K, Brokaw J. (2015). Endocrine and Integumentary Systems: A Team-Based Learning Module for Histology. MedEdPORTAL Publications. Available from: https://www.mededportal.org/publication/10290

Traser, C. J., Hoffman, L. A., Seifert, M. F. and Wilson, A. B. (2015), Investigating the use of quick response codes in the gross anatomy laboratory. Anat Sci Ed, 8: 421–428. doi: 10.1002/ase.1499
Note: This article was highlighted in the October 2015 issue of *Anatomy Now*. (See: http://amasan.informz.net/admin31/content/template.asp?sid=40980&ptid=1247&brandid=3960&uid=825287387&mi=4722378&ps=40980)

(2015). Anatomical Sciences Education Vol. 8, Issue 5, 2015 Cover Image. Anat Sci Ed, 8: C1. doi: 10.1002/ase.1565

POSTERS/PRESENTATIONS AT PROFESSIONAL MEETINGS

"Do I *really* have to complete another evaluation?" Exploring relationships among physicians' evaluative load, evaluative strain, and the quality of clinical clerkship evaluations October 2016
Anatomy and Cell Biology Fall Research Forum
Hosted by the Department of Anatomy and Cell Biology, Indiana University Medical School
Indianapolis, Indiana

Emphasizing the Importance of Qualitative Research in Anatomy Education: A "How-to-Guide" on Case Study Design, Implementation, & Data Analysis April 2016
Annual Meeting of American Association of Anatomists
Experimental Biology Conference
San Diego, California

Exposing the Gaps: A Review of Anatomy Education March 2015
Annual Meeting of the American Association of Anatomists
Experimental Biology Conference
Boston, Massachusetts

Investigating the Use of Quick Response (QR) Codes in the Gross Anatomy Laboratory April 2014
Annual Meeting of the American Association of Anatomists
Experimental Biology Conference
San Diego, California

Investigating the Use of Quick Response (QR) Codes in the Gross Anatomy Laboratory April 2014
Edward C. Moore Symposium on Teaching Excellence
Indiana University – Purdue University Indianapolis
Indianapolis, Indiana

CONFERENCES ATTENDED AS A REGISTRANT

Anatomy and Cell Biology Research Forum October 2016
Indiana University School of Medicine
Indianapolis, Indiana

Experimental Biology April 2016
Annual Meeting of the American Association of Anatomists
San Diego Convention Center
San Diego, California

Experimental Biology March 2015
Annual Meeting of the American Association of Anatomists
Boston Convention and Exhibition Center
Boston, Massachusetts

Experimental Biology April 2014
Annual Meeting of the American Association of Anatomists
San Diego Convention Center
San Diego, California

Edward C. Moore Symposium on Teaching Excellence April 2014
Indiana University – Purdue University Indianapolis Student Center
Indianapolis, Indiana

OTHER EDUCATIONAL & SERVICE ACTIVITIES

Indiana University Center for Anatomical Sciences Education (IU-CASE)
Gross Anatomy Laboratory Tour Instructor/Educator 2015-2017
Department of Anatomy and Cell Biology, Indiana University School of Medicine
Indianapolis, Indiana

Celebrate Science Indiana Exhibitor 2015
Indiana State Fairgrounds
Indianapolis, Indiana

Professional Development Achievements and/or Certifications
Tier One Academic Teaching Scholar 2016
Office of Faculty and Professional Development, Indiana University School of Medicine
Indianapolis, Indiana

Journal/Educational Resource Reviewer

Journal of Biomedical Education	2016
Anatomical Sciences Education	2015

Professional Organizations

American Association of Clinical Anatomists	2015-Present
Human Anatomy and Physiology Society	2014-Present
American Association of Anatomists	2013-Present