PREDICTING SUCCESS IN THE FIRST YEAR OF RESIDENCY TRAINING

CHRIS ASHLEY FOX

JULY 2015
The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

PREDICTING SUCCESS IN THE FIRST YEAR OF RESIDENCY

presented by Chris Ashley Fox,

a candidate for the degree of doctor of philosophy in Educational Leadership and Policy Analysis

and hereby certify that, in their opinion, it is worthy of acceptance.

__________________________
Professor Joe Donaldson

__________________________
Associate Professor Bradley Curs

__________________________
Associate Professor Casandra Harper

__________________________
Associate Professor Jeni Hart

__________________________
Associate Research Professor Kimberly Hoffman
ACKNOWLEDGEMENTS

I would like to thank Dr. Joe Donaldson for his thoughtfulness and insight over the years; and most of all for his dedication at ensuring this project was completed. You have given so much to the field of Education and are an inspiration to me. Many thanks go out to my committee members Dr. Curs, Dr. Hart and Dr. Harper for their valuable insights along the journey. Special thanks go out to Dr. Kimberly Hoffman who I consider ‘my favorite co-worker ever’ and greatest examples of how you can both be a great person, great boss and do great work.
# Table of Contents

**Acknowledgments** ........................................................................................................... ii  
**List of Illustrations** ........................................................................................................... iii  
**List of Tables** .................................................................................................................... iv  
**Abstract** ............................................................................................................................ vi  

Chapter 1 Introduction

- **Introduction** ................................................................................................................... 1  
- **Purpose** .......................................................................................................................... 6  
- **Research Questions** ....................................................................................................... 6  
- **Methodology** .................................................................................................................. 6  
- **Limitations, Assumptions and Design Controls** ......................................................... 7  
- **Definitions** ..................................................................................................................... 8  
- **Significance of the Study** ............................................................................................... 10  

Chapter 2 Review of Literature

- **Cognitive Predictors of Success** ..................................................................................... 13  
- **Cognitive Variables with Predictive Significance** ......................................................... 14  
  - **MCAT**  
  - **STEP Examinations**  
  - **In-Training and Knowledge Examinations**  
- **Predictive Validity of Exam Performance and Board Examinations** ............................ 16  
- **Non-Cognitive Predictors of Success** .......................................................................... 17  
- **Climate in Medical School** ........................................................................................... 18
DESCRIPTIVE VARIABLES .................................................................20

  Minority Status

  Self-Efficacy

  Sex

  Approach to Learning

  Personality traits and Interpersonal Skills

PREDICTING SUCCESS AFTER ENROLLMENT .........................27

PROFESSIONAL COMPETENCE ..............................................28

CONCLUSION .............................................................................29

Chapter 3 Methods

RESEARCH DESIGN .................................................................33

POPULATION ..............................................................................34

CONSTRUCTS AND DATA .........................................................35

DEPENDENT VARIABLES .........................................................36

  AKT

  In-Training Examination

  Remediation/Probations

INDEPENDENT VARIABLES .......................................................39

  New Innovations Resident Evaluation Database

  Gestalt Survey

  Pre-admittance Residency interview scores

  Sex

  Minority Status
CONTROL VARIABLES .........................................................43

   Step 1 and Step 2 scores

TIMING OF DATA COLLECTION ..............................................43

DATA COLLECTION AND THE INSTITUTIONAL REVIEW BOARD .................................................................45

ANALYSIS ..................................................................................45

   Research Question 1

   Research Question 2

LIMITATIONS AND CONCLUSIONS ............................................50

Chapter 4 Results and Findings

POPULATION DESCRIPTION ..................................................52

DEPENDENT VARIABLES .........................................................53

   AKT 1 month Examination

   AKT 6 month Examination

   In-Training Examination

   Remediation Status

KEY INDEPENDENT VARIABLES ..............................................57

   Step 1 scores

   Step 2 scores

   Interview Scores

   Gestalt Survey Scores

   Professionalism Scores
Crisis Management Scores

Overall Clinical Competency Scores

RESEARCH QUESTIONS .........................................................65

Question 1

RESULTS OF ANALYSIS WITH INTERVIEW VARIABLE ..........66

AKT 1 month Examination

AKT 6 month Examination

In-Training Examination

RESULTS OF ANALYSIS WITH GESTALT SCORE VARIABLE …68

AKT 1 month Examination

AKT 6 month Examination

In-Training Examination

RESULTS OF ANALYSIS WITH GESTALT SCORE/INTERVIEW VARIABLES .................................................................70

AKT 1 month Examination

AKT 6 month Examination

In-Training Examination

RESULTS OF ANALYSIS WITH TOP 5 INDEPENDENT VARIABLES ...............................................................73

AKT 1 month Examination

AKT 6 month Examination

In-Training Examination

SUMMARY OF RESULTS FOR QUESTION 1 .........................76

Step 1 Scores
Step 2 Scores
Gestalt Scores
Sex
Minority
Interview Scores

SUMMARY OF RESULTS FOR QUESTION 2 ……………………….83

Sex
Minority Status
Step 1 Scores
Step 2 Scores
Gestalt Scores
Interview Scores

SUMMARY ………………………………………………………………………….85

Chapter 5 Summary and Discussion

SUMMARY OF ANALYSIS ………………………………………………..86

LIMITATIONS OF STUDY ………………………………………………..87

Internal Validity

Limitations Summary

DISCUSSION OF FINDINGS FOR QUESTION 1 …………………….90

Descriptive Variables
Minority Status
Sex
Test Performance
Non-Cognitive Predictors of Success

Gestalt Scores

Interview Scores

DISCUSSION OF FINDINGS FOR QUESTION 2 .........................94

Summary

Variables of Significance

Sex

Interview Scores

Gestalt Scores

CONCLUSION .................................................................100

IMPLICATIONS OF FINDINGS ........................................101

FUTURE DIRECTIONS FOR RESEARCH ...............................103

APPENDIX A .................................................................107

BIBLIOGRAPHY .............................................................109

VITA ........................................................................123
LIST OF ILLUSTRATIONS

Figure
1. Histogram of AKT 1 month Percentile rank scores ..................55
2. Histogram of AKT 6 month percentile rank scores ..................55
3. Histogram of In-Training scores for study cohort .....................56
4. Histogram of Step 1 raw scores for study cohort .....................58
5. Histogram of Step 2 raw scores for study cohort .....................59
6. Histogram of Interview scores for study cohort .....................60
7. Histogram of Gestalt scores for study cohort .......................62
LIST OF TABLES

Table

1. Collection Timetable for Dependent and Independent Variables used in Research Question 1 ..............................................44

2. Collection Timetable for Dependent and Independent Variables Used In Research Question 2 .............................................49

3. Descriptive Statistics of Study Population ........................................53

4. Descriptive Analysis of Key Dependent variables ............................54

5. Descriptive analysis of Key Independent Variables .........................57

6. Regressions Run with the Dependent Variables AKT 1, AKT 6, and In-Training Exam with independent variables of Interview Score .................................................................67

7. Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variable of Interview, Missing, Sex, and Minority ..................................................67

8. Regressions Run with Dependent Variables AKT 1, AKT 6, and In-Training Exam with independent variable of Interview, Missing, Sex, Minority, and Step 1 & 2 .........................................................68

9. Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variable of Gestalt ..........69

10. Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt Scores, sex, minority .................................................................70

11. Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt Scores, sex, minority, Step 1 & 2 .........................................................70

12. Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt, Interview, and Missing .................................................................72

13. Regressions Run with Dependent Variables AKT 1, AKT 6, and In-Training Exam with independent variables of Gestalt, Interview, Missing, Sex, and Minority ..................................................72
14. Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam and independent variables of Gestalt, Interview, Missing, Sex, Minority, Step 1, and Step 2 ……………..73

15. Regressions run with Dependent variable AKT6 and 5 independent variables with lowest Variance Inflation Factor (VIF) …………………………………………………………...75

16. Regression run with Dependent variable Remediation status and independent variables Interview/Missing ……………….79

17. Regression run with Dependent variable Remediation status and independent variables Interview/Missing, Sex, and Minority ………………………………………………………….....79

18. Regression run with Dependent variable Remediation status and independent variables Interview/Missing, Sex, Minority, Step 1 and Step 2 ……………………………………………………………..80

19. Regression run with Dependent variable Remediation status and independent variable Gestalt …………………………………….81

20. Regression run with Dependent variable Remediation status and independent variable Gestalt, Sex, Minority status …………....81

21. Regression run with Dependent variable Remediation status and independent variable Gestalt, Sex, Minority status, Step 1 and Step 2 …………………………………………………………..81

22. Regression run with Dependent variable Remediation status and independent variable Gestalt, Interview, and Missing …………...82

23. Regression run with Dependent variable Remediation status and independent variable Gestalt, Interview, Missing, Sex, and Minority …………………………………………………………..83

24. Regression run with Dependent variable Remediation status and independent variable Gestalt, Interview, Missing, Sex, Minority, Step 1 and Step 2 …………………………………………………………..83
ABSTRACT

This dissertation is motivated by two research questions 1) What are the independent variables that are associated with performance for first year residents in an Anesthesiology Residency? and 2) What linear combination of factors can predict success (not entering remediation/probation) for the first year of residency? A post-hoc correlational analysis of 3 years of first year resident data was conducted using linear regression to determine any significant predictive relationships. Previous literature has indicated that past test performance is likely to predict future performance and those models of prediction that incorporate both cognitive and non-cognitive attributes tend to be the most successful. Results indicated a strong association between Step 1 scores and minority status for predicting performance on first year examinations with weaker associations for the variables of gestalt and interview scores. Analysis also indicated a strong association between gestalt scores and sex and prediction of whether individuals would end up on remediation or probation within their first year of training.
Chapter 1

Introduction

Performance on cognitive exams throughout one’s academic career has proven to be highly predictive of future performance on cognitive examinations. However the same cause and effect relationship cannot be said of other variables that may hinder successful completion of medical residencies, such as lack of professionalism or poor communication skills. While there are robust screening strategies for academic performance prior to acceptance into any residency program, there is always a subset of residents who enter a specialized training program and get into immediate “academic” difficulty and are consequently placed on remediation or probation (ACGME, 2014). This difficulty may be due to perceived deficiency in a core competency (social and clinical) area or a lack of demonstrable knowledge, or a combination of both. The creation of an accurate model of prediction of success in residency that evaluates both cognitive and non-cognitive data would afford both institutions and individuals a better opportunity to choose the correct specialty and minimize expense in years of training in a field to which the individual might not be suited.

With an aging populous and increased staffing required in the medical fields the demand to increase class size and to graduate competent physicians is becoming more and more urgent (IHS, 2015). Graduating competent independent physicians is however a long and expensive process for both the individual and the institution involved. On top of the pressure to graduate well rounded competent professionals, medical schools and residencies are faced with the institutional and programmatic desire to maintain student diversity while maximizing the cost/time to train ratios. By the time individuals reach
residency they have already invested significant time and monies into their training and it can be a very costly mistake for both the program and the individual (Jolly, 2005) if the wrong specialty is chosen or a good match (whether academic or personal) is not made. If a program could identify characteristics of residents who are likely to graduate as competent physicians with minimal program intervention during either the pre-review or interview process or during their first 6 months of residencies, a more effective use of resources could be obtained. Early prediction of residents who are likely to be in distress, can help a program twofold, first, to make better decisions of who to rank in their program and second to help direct the use of educational resources towards individuals most in need/at risk. Both of these will result in a higher overall quality of education for all trainees and programs.

Being certified as an ‘independent consultant’ (physicians who are qualified to work independently in their specialty) is a costly and time consuming process for both the individual receiving the education and for the providing institution(s). Typically to become a ‘consultant physician’ in a chosen medical field, an individual would need to complete a 4-year undergraduate degree, a 4-year medical degree, and a 4-plus year residency. Compounding the pressure to select ‘the right’ residents, or ones who will not have academic difficulty, is the fact that Medicare reimbursement, which funds all residency spots at the researched institution, remains stagnant with constant threats of decreased funding from the government in every budget proposal (IHS, 2015). With fewer funds to pay residency slots and stressed overall resources in healthcare, programs are less willing to take risks with residents who on paper, and in person, do not “look” exceptional.
In order to minimize educational cost, maximize training efficiency, and time and effort expended trying to graduate "problem residents," researchers have long been searching for the answer to the question, “How do we select individuals who will be successful?” To answer this question many studies have been conducted to analyze individual performance data with/against admittance data for significant patterns. To this day, however, their questions have not fully been answered nor has a ‘best-fit’ prediction of success model been created for residency programs (Hendren, 1988; Julian, 2005; Shen & Comrey, 1997). The problem with answering the question of selecting ‘successful residents’ and creating an accurate model of prediction may lie in the variations in educational institutions at which the individuals were trained, institutional expectations for residency, and the applicants themselves and/or simply in the quality and type of data available for analysis.

I believe that any analysis conducted designed to predict resident success will always come up with the same responses: that cognitive (academic) variables will strongly predict future performance on academic examinations. However a thorough review of the literature (Bridgham, 1990; Gore, 2006; Shen & Comrey, 1997), discussed in Chapter 2, reveals that cognitive variables alone cannot predict success in medical school or residency (success is defined by the program as graduating a competent individual consultant). While a strong correlation between cognitive variables and academic performance has been demonstrated in the literature when examined in the context of the first 2 years of medical school (traditionally the didactic years), the same cannot be said for the clinical years (Huff, Koenig, Treptau & Sireci, 1999; Julien, 2005). In the clinical years and during residency, it is the non-cognitive variables that tend to
have the greatest impact on evaluation and performance scores (Daly, Levine & Adams, 2006; Hayden, Hayden & Gamst, 2005, Kron et al., 1985). Because residency is primarily the evaluation of an individual’s daily performance in a clinical setting (with several benchmark cognitive exams along the way), it is essential that we focus our research on all performance indicators or independent variables (including the non-cognitive ones) that make residents more likely to be placed on remediation. By predicting which residents are likely to be less successful, meaning that they fall into remediation and/or probationary status, institutions can either revise their selection process or implement interventions and remediation plans early on in the residency to ensure that most residents will graduate on time, with minimal intervention, as a competent physician.

As previously stated, significant research has been conducted on the ability to create accurate models of prediction but have either failed to come up with correlations of significance or ‘tell-tale’ indicators of successful performance in a residency program. However, no research in this area has incorporated performance on structured competency questions (designed to evaluate the interviewee’s knowledge base and aptitude in areas such as professionalism, communication, systems based practice, and practice based learning and improvement) during the interview process and a “gestalt” survey (a survey of resident performance by faculty that has them evaluate residents on three competencies) at one month into the program to further examine predictors of success. The importance of investigating these variables together to predict probability of remediation rather than looking at the simple prediction of academic success on cognitive exams lies in the understanding that a competent physician must be able to
demonstrate competency on both the ‘soft’ variables (ones that are often described as personality or communication skills) along with medical knowledge (cognitive exams and clinical skills). For example, many prediction models examine how cognitive variables such as test scores relate to academic success at a later date in one’s academic career, but few researchers have been able to predict overall success in a residency program (or during clinical years) because of independent variables that are much more difficult to measure such as communication skills, levels of professionalism, dedication to lifelong learning (Callahan, Hojat, & Veloski, 2010; Ferguson & Medeley, 2002; Silver & Hogson, 1997). As demonstrated in the review of the literature that follows this chapter, no one single predictive model has yet been able to account for a highly significant amount of the variance in resident performance, but a combination of several types of variables (cognitive and non-cognitive) has been demonstrated to be somewhat more effective at determining which individual is more likely to graduate or successfully complete the program. By incorporating both cognitive variables and non-cognitive competency variables in an analysis model the institution/program may be able to make more accurate predictions of residents who will end up in remediation and those who will continue on remediation into probationary status.

By examining these questions, programs can either revise their selection process and/or their resident promotion committee policies (to start the dismissal process earlier) or determine interventions that can ensure residents in distress complete the program successfully. To that end, this study will research the question of which components or individual characteristics best predict successful future residency performance (as determined by performance on cognitive exams) and whether or not any variables or
combination of variables can predict which trainee will be placed on remediation and/or probationary status within their first year of residency.

**Purpose**

The purpose of this study was to conduct an exploratory investigation of the accuracy of several independent variables for predicting overall success in an anesthesiology residency program in an academic medical center in the upper Midwestern United States. The research focused on one program in one institution in order to hold constant any factors that may vary among different programs at different institutions. I investigated the research questions outlined below by using ordinary least squares regression analysis (OLS). Ordinary least squares was chosen as the statistical technique, even with its tendency to contain more standard errors, for its high power of analysis.

**Research Questions**

To address the purpose I examined the following questions:

1) What are the independent variables that are associated with future Anesthesiology Knowledge Tests (AKT) and In-Service Examination test performance of first year residents?

2) What linear combination of factors, controlling for Step 1 and 2 examination scores, predicts success (not entering remediation and/or probationary status) during the first year of Residency?

**Methodology**

This research project is a post-hoc correlational analysis of 3 years of residents who interviewed and were accepted in an anesthesiology residency program in the
Midwest (to be called Midwest Medical School). The cohorts analyzed were comprised of current residents who entered the residency after July 2008 and completed the 3 year residency. The dependent variables in this research are individual performance in residency. Performance was measured by demonstration of medical knowledge on standardized examinations during the first year of residency (Anesthesia Knowledge Test and In-Training examination) and, for research question two, successful performance in residency was defined as a resident who remained in good standing (i.e., never enters remediation or probationary status (defined below)). For the independent variables, I examined Step 1 and Step 2 scores (standardized tests that all Medical Student must complete for admission to Residency); Anesthesiology Knowledge Test scores (a standardized exam taken by all US anesthesiology residents) at 1, 6, and 12 months when appropriate; In-service exam scores; Gestalt survey scores; structured competency question performance from residency interview (aggregated as a total score); and mean scores from the New Innovations (online evaluation and curriculum management tool) faculty evaluation of residents in the areas of crises management, overall competency and professionalism. The data were analyzed using ordinary least squares analysis.

**Limitations, Assumptions and Design Controls**

There are several assumptions that are made in regards to this study. The first is that all individuals are admitted on the basis of their individual merit and have been assessed using equivalent criteria. Second, there is no way to measure the impact of opinion on the decision making process; while we assume that every resident who meets the criteria for probation/remediation will be placed onto remediation/probation, instances occur when a judgment call by the committee is made to over-rule that
PREDICTING SUCCESS IN RESIDENCY

decision. Repeated measure bias is also a limitation of this research, as independent variables are often made up of evaluations by faculty, in that there are instances where the same faculty evaluate the same trainee on different independent variables. A further limitation of the study is the inability to make generalizable recommendations that could be applied to all anesthesiology residencies or residencies in general throughout the United States. Because all programs have individualized admissions processes, curriculum, and assessment tools, it is not plausible to suggest that a one-size fits all answer can be achieved. Still, the findings of this study will inform directors of residency programs to the extent that the programs they administer are comparable to the one studied in the present research.

Definitions

Success: A resident who completes the first year of residency training without entering remediation or probationary status (defined below) will be considered successful. Residents first enter into remediation status if they meet the criteria listed below; a resident who does not improve during the defined remediation cycle of 3 months will either be placed into a secondary remediation period (3 additional months) or into probationary status. A resident cannot be on remediation for more than 2 cycles (or 6 consecutive months) and a resident cannot be on probation for more than one cycle.

Criteria for Remediation: The Housestaff Education and Performance Review Committee of the Department of Anesthesiology at the Midwest Medical School place residents on remediation if they meet one or more of the following criteria:

- Their 1- or 6-month AKT exam score is <25% the national average
- Their overall clinical competence score is >1 standard deviation below their class mean at any quarterly evaluation
• Their evaluations have a score of >2 “goals and objectives of the rotation not met
• A serious event documented by a faculty member or a series of events that indicates the resident is lacking the appropriate knowledge base or clinical skill for their status
• Professional misconduct

Criteria for Probation: The Housestaff Education and Performance Review Committee of the Department of Anesthesiology at a large Midwest medical school places residents on probation if they meet two or more the following criteria, or if the resident fails to progress (continues to meet the criteria for remediation) while on remediation:

• A 1-month AKT exam score <25% the national average
• A 6-month AKT exam score <25% the national average
• A clinical competence score is >2 standard deviations below their class mean at any quarterly evaluation.
• Evaluations have >2 “goals and objectives not met”
• A faculty documented, serious event that indicates the resident is lacking the appropriate knowledge base for their status
• Professional misconduct

Gestalt Survey: A three question survey given to the faculty after the new class (CA-1) has been working in the operating rooms (OR) for one month. The survey asks whether or not the faculty perceives the residents to be below their peers, on par with peers, or above their peers in the areas of medical knowledge and clinical skills and also asks them to predict whether or not they feel the residents will be an average, below average, or a superior trainee.
Significance of the Study

The results of this study will be very informative to the Anesthesiology Department in the Midwest Medical School as well as for other anesthesia residency programs. If the Midwest Medical School anesthesiology residency program is able to proactively identify ‘at risk’ individuals then there would be a twofold benefit. First, the department could be more selective in the recruitment process and second, once a resident is admitted and begins to display indicators of probable performance difficulty we could be proactive in creating effective interventions that decrease attrition and/or program failure rate.

Even though the sample size is small, and factoring in the variations in the curriculum and promotion process across anesthesiology residency programs, there remains the potential to be some generalizability in results. There is an inherent applicability of results built in because of the overall limitation of the applicant pool (accepted residents trend the same on admission score averages and educational backgrounds) and because the curriculum and evaluation foundation of most programs are comparable due to required Accreditation Council for Graduate Medical Education (ACGME) accreditation standards. One primary benefit that may be afforded by this research is that the results may give evidence to significant identifiable independent variables about residents who are likely to be placed on remediation/probation during their residencies. With that information, institutions would have the ability be more selective in their match process or be more proactive in designing and implementing effective interventions and/or remediation plans. A second significant impact of the study could be a refinement in the way practitioners look at how performance on the core
competencies impact future academic success. Instead of trying to predict what makes someone successful in his/her residency based on their previous academic success; which from a thorough review of the literature has proven to be difficult over the years, the integration of several key non-cognitive measure into the prediction model could be helpful in developing a more robust screening process and curriculum for medical residencies.
Review of Literature

Creating accurate models to predict academic success of individuals enrolled in residency programs is a daunting process that involves the analysis of not only academic variables, but psychosocial and/or demographic ones as well. If a highly accurate prediction model, designed to predict success during the residency training years and/or on board examinations at the end of training, were created and applied to selection methods of residency applicants, it could lead to a greater efficiency in the overall selection process as well as assist training institutions in restructuring their curriculum and remediation processes to better meet the needs of incoming and/or current residents.

In order to minimize educational cost, maximize training efficiency, and time and effort expended trying to graduate "problem residents," researchers have long been searching for the answer to the question, “How do we select residents who will be successful?” Success, of course, is sometimes in the eye of the beholder, but in the context of residency it is the ability to graduate a competent and independent consultant physician. To date the question of resident selection has neither fully been answered nor has a ‘best-fit’ prediction of success model been created. The problem with both answering the above question and creating an accurate model of prediction may lie in the variations in educational institutions and their respective curricula and evaluation methods as well as the applicants themselves.

In this chapter, I will report on selected research that has been conducted to date on the creation of models used to predict academic success both in residency and in medical schools and investigate the variables that add the most variance (or carry the most significance) in predicting academic success. Results of the literature review will be
categorized into several areas that include cognitive factors, non-cognitive factors, individual demographics, and institutional types as they relate to predicting academic success.

This chapter will examine studies conducted to predict success in undergraduate medical education degrees (M.D.) and during residency training (graduate medical education). The studies will be included as a composite for two reasons: (1) they track the same individuals over time, and in progression in their medical career training, and (2) entrance into both programs is highly determinant on standardized test scores that capture the cumulative progression of knowledge.

**Cognitive Predictors of Success**

Cognitive intelligence (thinking, learning, judgment skills) is the most common variable(s) used as a predictor of success. Cognitive intelligence is most often measured by standardized scores such as grade point averages (GPAs) received in high school or undergraduate courses, or an individual’s scores on the Standardized Aptitude Tests (SATs) or the American College Test (ACTs) (Ridgell & Lounsbury, 2004).

Specific to the field of medicine is the Medical School Admissions Test (MCAT) for admission into medical school and the Step 1 and Step 2 examinations that all medical students have to complete before being considered competitive for entry into residency. The United States Medical Licensing Examination (USMLE) exams consist of three parts, Step 1, Step 2, and Step 3 (Clinical Skills). The USMLE exams are intended to “assess a physician's ability to apply knowledge, concepts, and principles, and to determine fundamental patient-centered skills that are important in health and disease and that constitute the basis of safe and effective patient care” (USMLE, 2009).
To date, most of the predictive success models created by researchers have measured the role that individual and combined cognitive variables play in predicting success. Essentially all research studies have found that while each individual standardized exam adds a significant percentage to the variance, the best models include a combination of multiple examination scores. Simply stated, models with many cognitive variable measurements are almost always more accurate predictors of success (but still not as accurate as models that contain both cognitive and non-cognitive variables) (Fleming, 2002; Huff et al., 1999; Julian, 2005; Julian & Lockwood, 2000; Tross, Harper, Osher, & Kneidinger, 2000; Wiley & Koenig, 1996).

**Cognitive Variables with Predictive Significance**

**MCAT exams.** Numerous studies have demonstrated the predictive validity of the Medical College Admissions Test (MCAT) exam on future academic success (Jones & Thomae-Forgues, 1984), including the score’s relationship to performance on STEP examinations and even board exams. Researchers have consistently found that a combination of MCAT with other academic variable(s) almost always increases prediction accuracy dependent upon expected outcome (Downey, Collins & Browning, 2002; Julian, 2005; Solomon, Vancouver, Reinhart, & Haf, 1989). For example, Julian (2005) examined residents from 14 different medical schools through to residency to see if she could predict grades and or academic difficulty/distinction by examining undergraduate GPAs, MCAT scores and institution selectivity. She found that the MCAT was very efficient in predicting licensure exam scores (board exams) but that performance during medical school was best predicted by a combination of GPAs and MCAT scores.
STEP exams. Traditionally a significant predictor of success for residency is performance on the Step exams. The Step 1 and 2 examinations are tests that all medical students must take for entry into residency. There is also a clinical skills portion of the Step exam, often referred to as Step 3 CS, but that portion is not universally required to be completed before residency entry. The USMLE administers the Step exams with the intended purpose to assess the knowledge base and whether medical school students or graduates understand and can apply important concepts of the sciences basic to the practice of medicine (Step 1) and to assess critical thinking skills or whether medical school students or graduates understand and can apply important concepts of the sciences basic to the practice of medicine (Step 2). Many researchers have determined that there is a strong degree of correlation between success on the Step 1 and Step 2 exams and future success on in-training exams or board exams (Elam, Mitzi, & Haist, 1994; Myles, & Galvez-Myles, 2003).

Elam et al. (1994) examined a historical cohort to determine whether or not preadmission and medical school performance data could predict success in an internal medicine residency. They determined that residents (interns in this case) who had performed well in their first 2 years of medical school and on their USMLE exams (Steps 1 and 2) were closely correlated to those who had outstanding performance on medical knowledge tests (cognitive) in the ambulatory care settings, and those interns who had scored well in their clinical years (typically more hands on and interactive) of medical school accordingly scored higher in their clinical settings during residency.
In-Training and Knowledge Exams. Much of the research conducted on predicting academic success uses measures that occur before or after an individual has been admitted (or graduated from) a program with very little focus on examinations completed while training is in progress. Godellas and Huang (2000) attempted to identify the combination of variables that could predict which residents would be successful on the in-training examination (taken upon completion of the first year of residency). Their research found that a combination of conference attendance, previous academic performance, probationary status, and amount of sleep and study were the greatest contributors to scores on the in-training examinations (such as AKT, In-service standardized multiple choice exams given to all residents in certain specialties). To that end, it seems as though a combination of several non-cognitive variables were high predictors of success on in-training exams, much more so than just previous academic performance itself.

Predictive Validity of Exam Performance and Board Examinations

Beyond being certified an independent consultant in anesthesiology, one must become licensed for practice, which requires passing the board examination. The board examination occurs in two parts (a written and a verbal) with most residents choosing to take the written examination first. Passing the board exam within the first year of graduation is often considered a measure of programmatic success for residency programs (and it is often a condition of a contract for faculty appointment or clinical appointment for the individual).

Much research has been conducted on the relationship between board examinations and in-training exams across specialties. Several researchers have found that...
PREDICTING SUCCESS IN RESIDENCY

strong correlation between performance on In-training exams with board performance (Bailey, Yackle, Yuen, May, & Voorhees, 2004; Klein, Austin, Randolph; McClintock & Gravlee, 2010; Sharkey & Hilibrand, 2004). While all these researchers found strong correlations they also identified that incorporating other variables such as personal motivation, country of medical school, gender, and program accreditation length contributed to the strength of the predictive model (Klein et al., 2004).

While cognitive variables can account for a significant proportion of the variance in predictive models they are not able to account for all. To that end, researchers have also investigated the role that soft-competencies or non-cognitive variables play in predicting academic success.

**Non-cognitive Predictors of Success**

Many residencies and medical schools select residents purely on the basis of their exam results (whether it is course grades or results on standardized performance exams). It is much rarer to see non-cognitive variables included in predictive models. Non-cognitive variables are those such as behavioral competencies, individual personality traits, and personal motivation. Non-cognitive variables are traditionally competencies that can only be measured face-to-face in an interview setting or from the completion of a self-report assessment tool.

Traditionally many medical schools and residencies disregarded non-cognitive variables in the admission process, (other than in instances where they meet a truly egregious interviewee) because they are difficult to evaluate reliably and consistently. A second reason for the exclusion of non-cognitive variables is that barring the personal interview, non-cognitive factors such as student motivation, parents’ income (socio-
economic status) and education levels rely mainly on self-reported data (which may or may not be an accurate reflection of the student’s ability) (Chen, Campbell, & Suleiman, 2001). Furthermore, institutional factors such as faculty effort and college environments may positively or negatively impact an individual’s success, but once again are difficult variables to consistently measure and evaluate.

This research project will evaluate the importance of non-cognitive variables in predicting success in residency. Selected literature demonstrates that variables such as self-efficacy, motivation, self-appraisal skills, organizational skills, and tolerance of ambiguity have some promise for contributing to predictions of success, but in combination with other measures. As stand-alone variables they have been found to be not very significant.

**Climate in Medical School and Residency**

When looking at descriptive variables it is also important to pay attention to the overall climate within the training institution(s) and programs enrolling trainees. The breakdown of sex, minority status, and even socio-economic status has the potential to be an impactful factor on predicting future success. An institution lacking racial and ethnic diversity may have trouble recruiting a more diverse population due to the under-represented minority not wanting to feel like the ‘token’ or to be isolated from like groups (Hung et al, 2007). Lacking a diverse, non-inclusive environment, the URM may end up feeling like they have to work twice as hard for the same recognition of their majority counterpart (Bright, Duefield & Stone, 1998). On the other hand, an institution with high levels of exposure to a diverse population (both trainee, trainer, and patient) is likely to produce individuals with positive attitudes on the value of diversity (Vela et al.,
2010; Guiton, Chang & Wilkerson, 2007). Without exposure to minority populations during training majority students are more likely to graduate lacking the cultural competency or comfort levels to work with diverse populations (Saha et al., 2008, Lum & Korenman, 1994) leading to underserved and medically under-represented areas.

The percentage of females in medical school has been on the decline since the all-time high of 50.8% reached in 2003 and as of 2012 sat around 47% of medical school matriculants (AAMC, 2012). In the field of Anesthesiology the gender breakdown of resident trainees is 63.4% male and 36.6% female (AAMC, 2014), so while the male/female ratio is fairly even in medical school, men are a much greater proportion of the trainees in the field of Anesthesiology. Women in medicine are also under-represented as faculty, comprising only 4% of full professors in academic medicine (Pololi et al, 2012). This underrepresentation has been attributed to many things including working in a perceived sexist climate (Fitzgerald et al., 1999; Hesson-McInnis & Fitzgerald, 1997) and gender discrimination (Carr et al, 2003; Higgins & Thomas, 2001) both which leads to higher job dissatisfaction and attrition.

Under-represented minorities are truly underrepresented at all levels of medical training. In 2011 blacks comprised 6.1%, Native Americans comprised 0.2%, and Hispanics comprised 8.5% of all medical school matriculants where as whites comprised approximately 60 percent of medical school faculty, with Hispanics and Blacks comprising only 4 and 2.9 percent respectively (AAMC, 2012).

With women and minorities being significantly under-represented in medicine (decreasing as the level of training or promotion increases) and studies suggesting that diversity in medical school (of both the trainees and trainers) can reduce health disparities
PREDICTING SUCCESS IN RESIDENCY

(Smedley, Stith & Nelson, 2003; Institute of Medicine, 2003) and increase access to healthcare in underserved areas (Saha & Shipman, 2008; Ko et al, 2007), institutions have become increasingly interested in investigating the climate in their institutions. However, studies have found that Caucasian men tend to perceive institutions as more fair and inclusive than women or minorities (Hung et al., 2007; Mor Borak, Cherin & Berkman, 1998) and medical students report an environment where minority groups (ethnic minorities, people with strong religious beliefs, non-English speakers, women, people with low socio economic status) are still treated with offensive behavior or disparaging remarks (Dhaliwal et al, 2013).

Descriptive Variables

**Sex.** Significance performance differences have been recorded between the sexes on examinations such as part 1 of the National Board of Medical Examiners (NBME) and Step 1 with males consistently outperforming females (Case, Becker & Swanson, 1993; Cuddy, Swanson, Clauser, 2008; Dawson et al., 1994; Weinberg & Rooney, 1978). This gap, however, traditionally narrows over the course of medical school training (Case, Becker & Swanson, 1993; Weinberg & Rooney, 1978), as the focus of medical education shifts from basic sciences to clinical applications of knowledge. The performance gap almost completely disappears by the time individuals take Step 2, with males and females performing comparably on examination (Case, Swanson, & Ripkey, 1996).

Females typically fare better than their male counterparts on Objective Structured Clinical Examinations (OSCE), which integrate medical knowledge into a clinical setting while also evaluating non-cognitive skills such as communication style and professionalism (Rutala, Witzke, Leko, & Fulginiti, 1991). However, one study by
Calkins, Arnold and Willoughby (1987) discovered that non-cognitive variables were the best predictors of how both males and females would perform on clinical clerkships.

**Minority Status.** Over the past 2 decades we have steadily seen the progressive elimination of affirmative action admittance policies in states such as Michigan, Texas, California, and Washington (Nobel, 2003). Further, we may begin to see more minority student/affirmative action enrollment programs disappear as increasingly American’s believe that the playing field has been equalized for persons of all race, ethnicity and gender. The move away from affirmative action policies has a significant impact on prediction models for two reasons: a) the skill set and life experience of the non-majority and majority population may be fundamentally different, which can impact how they approach academics and learning, and b) with the assumption that minority student population often are products of lower socio economic status neighborhoods and underfunded schools (with less access to standardized test preparatory education) it may be unreasonable to expect the same level of achievement on cognitive variables (Nobel, 2003; Nobel & Sawyer, 2002).

Achievement gaps, between white and black Americans, have been found to exist at all levels of training from K-12 to college level. African American students, already underrepresented in the math and science fields (Culotta & Gibbons, 1992) are far less likely to complete college compared to their Caucasian counterparts (42% to 70%) (American Council of Education, 1990) and when they do graduate they do so with grades that are significantly lower. (Nettles, 1988). There has been much debate on whether or not it is fair to judge minority students based on their performance on standardized tests because of a purported test bias. Furthermore, some decry that it is
unfair to judge minorities based on high school GPA because of the variance in the public
school systems and related socio economic status. However research has shown that
achievement gaps between whites and non-whites exist on standardized test scores
independent of the socio-economic class of the test takers (Miller, 1995, 1996)

While some research confirms that minority students score worse on standardized
exams and achievement measures used for admissions (Nobel, 2002), others note that no
bias or discrimination exists. Irrespective of whether or not test bias exists there is a large
body of literature that establishes that standardized tests, known for their ability to predict
subsequent school achievement for white and black students, demonstrate poorer future
achievement for black students when the test scores are equivalent at the present time
(Two students who score the same on the SAT; the black student tend to have poorer
future scores (Jensen, 1980).) This over-prediction of future performance for URM
students integrates well with the theory of stereotype threat. The concept of stereotype
threat, where one is worried about being judged based on stereotypes, should not be
overlooked when examining the performance of minority trainees in advanced
educational training situations. Stereotype threat exists “for the members of any group
about who there exists some generally known negative stereotype.” (Steele, 1997) This
threat can depress the performance on an individual who is in the ‘threatened’ class. In
the case of URM students the fear of being stereotyped and performing poorly leads to
actual poor performance (Steele and Aronson, 1995, Steele, 1997).

For medical students the initial barrier of admission to their training is the MCAT
examination, which has been extensively studied to see if there is inherent bias in the test.
Koenig, Sireci and Wiley (1998) conducted a study to analyze the claim that the MCAT
was biased against minorities and women but found no basis to the claim. Additionally, Vancouver et al. (1990) attempted to analyze whether tests such as the MCAT had test bias toward minorities by applying a moderated multiple regression and the Cleary model of regression analysis. They found that ethnicity did not play any significant factor in predicting an individual’s success and that it was fair to apply the same predictive model to students of all ethnicities.

Barring the notion that the MCAT has built-in racial bias, the MCAT has been proven to be a strong indicator of success not only in the first 2 years (the academic curriculum) of medical school but also a good indicator of success in the third (clinical) year with both minority and majority students. In fact, the MCAT may be better at predicting non-majority third year performance than the performance of majority students. Huff et al. (1999) conducted a predictive validity study on 3,249 students enrolled in 14 institutions throughout the US and determined that the MCAT and other preadmission variables were able to predict African American and Hispanic third year (clinical) performance fairly accurately, but were less accurate at predicting the performance of others.

Noble (2003) conducted a study that examined ACT composite scores and high school GPAs of two groups of students: an African American/Caucasian sample (262, 553 students from 43 institutions) and a Hispanic/Caucasian group (174,890 students from 25 institutions). She found that ACT and high school GPAs were slightly more accurate at predicting success in African American than Caucasian students but that was not true for Hispanics.
Moving away from achievement discrepancies in cognitive variables between majority and non-majority students, many researchers have extended their analysis to investigate whether or not race, gender and/or religion are significant independent variables to predict success. Chen, Campbell and Suleiman (2001) tested both logistic and linear regression models for a best fit predictor for students in a minority professional school. They found that medical school first year students’ GPA and number of second year courses failed were the greatest predictors of Step 1 performance in both positive and negative directions and that these demographic variables contributed little to the prediction model.

Hoffman and Lowitzki (2005) conducted a study to investigate whether or not factors such as minority status or religious affiliation had a significant impact on student success. They posited that high school grades would be a greater predictor of achievement for minority students due to the high level of culture shock experienced when entering (e.g. culture, diversity,) university. They also hypothesized that the best predictor of success for majority students would be standardized tests such as the ACT and/or SAT. To test this, the researchers created a model fit (priori path analysis) by examining 60.5% of a first year class at a private Lutheran university by comparing their high school GPAs or SAT scores with answers they self-reported on a student opinion survey. Ultimately the researchers found support (albeit weak) for all their hypotheses.

**Self-Efficacy.** Academic self-efficacy as defined by Schunk (1991) is an individual’s confidence in his/her own abilities to perform at pre-designated levels. Gore (2006) conducted two incremental validity studies examining the role that self-efficacy plays in the prediction of college success and in both studies found that self-efficacy is a
fairly low predictor of college success at the beginning of college but a better indicator of success as the student progressed through training. Therefore, one might assume that prior academic success in a program is predictive of current or future success.

**Self-appraisal and organizational Skills.** Bridgham (1990) conducted a study to determine whether non-academic variables were good indicators of undergraduate medical education performance. He found that ‘soft’ variables, such as self-appraisal skills, organizational skills, and tolerance of ambiguity, provided very little gain in predicting success over the traditional MCAT, and GPA measures. However, Bridgham did find that residents who had a number of withdrawals and incompletes in their undergraduate curriculum could be predicted fairly accurately to have a strong possibility of delays in completing their pre-clinical curriculum. To that end Bridgham noted that prior demonstrations of academic delays were highly predictive of current or future delays in progress towards residency completion.

**Approach to Learning.** Researchers have also investigated whether or not an individual’s approach to learning impacted his/her future success. Approach to learning as defined by Lindblom-Yanne, Lonka and Leskinen (1996) is either “broad” or “narrow.” Lindblom-Yanne et al. measured approach to learning both by the evaluation of text comprehension and essay writings skills (which involves the integration of multiple learning skills) as well as performance on traditional multiple-choice tests (which principally involves the memorization of facts). It was their hypothesis that medical students need a “broad approach to learning” being able to be successful on both these test types (multiple choice knowledge tests as well as text comprehension and problem solving exams) rather than have a “narrow” focus and only scoring well on one
type. Through the use of confirmatory factor analysis the researchers substantiated several of their hypotheses: first, they found that their test was able to measure deep text comprehension rather than basic school learning and second, science test scores were the best indicators of success in the basic science curriculum.

**Personality Traits and Interpersonal Skills.** There is something to be said for an individual’s personality and the impact on academic performance. A 1997 study of 97 entering medical residents at UCLA by Shen and Comrey analyzed the relationship between residents’ cognitive abilities and personality traits (the Comrey personality scale measures trust versus defensiveness, orderliness versus lack of compulsion, social conformity versus rebelliousness, activity versus lack of energy, emotional stability versus neuroticism, extraversion versus introversion, mental toughness versus sensitivity, and empathy versus egocentrism). They found that MCAT scores and residents’ performance in medical school were highly correlated, but when the M score (results of Comrey test) was incorporated the accuracy of prediction increased by 20% (to an overall accuracy of 51%).

Furthermore, Schwartz, Donnelly, Sloan, Johnson and Strodel (1995) conducted a longitudinal analysis of the relationship between clinical evaluations, objective structured clinical exam scores, and the American Board of Surgery In-training exams. They found that measures on interpersonal skills were strong measures of overall resident performance and that when added to other non-cognitive measures increased the reliability of the evaluation.

Results from all of the above mentioned studies indicate that non-cognitive variables are not great independent predictors of success but that they all significantly
add to the power of a predictive model. Currently global assessments, or holistic assessments, are one way in which researchers have tried to capture non-cognitive data and input the variable into prediction models. For this research project I use two types of holistic assessment tools in the analysis; the interview score variable and the Gestalt score variable. Both these variables attempt to capture both the cognitive ability and non-cognitive competencies of the individual being evaluated. Traditionally the predictive validity of global assessments is varied; previous research has determined that global assessment tools are significantly correlated with clinical performance (Brothers & Wethernolt, 2007; Donnon, Paolucci, & Violato, 2009; Gough, Hall Harris, 1964; Meredith, Dunlap & Baker, 1982; Reiter, Eva, Rosenfeld & Norman, 2007) but tend to lose predictive power when medical knowledge was a significant component of the clinical competency metric (Elam & Johnson, 1992; Reiter et al. 2007; Smith, 1991, Meredith et al., 1982).

**Predicting Success after Enrollment**

After examining whether or not it is possible to predict an individual’s success in medical school based on their admission/performance data, one last area to consider is if the same model could also be applied to predict residents in distress while in the process of completing the program.

Hendren (1998) analyzed the records of 41 medical students who were at risk for expulsion to see if he could identify their key characteristics. The variables he analyzed were sex, age, race, undergraduate GPAs, MCAT scores, nature of problem that put them at risk for dismissal/failure, course type failed, and personality characteristics. Hendren divided his participants into 3 subsets, the first included residents who were having
difficulty because of academics, the second included residents who were having problems
due to interpersonal events, and the last group included residents who were in trouble due
to intrapersonal issues. Across all groups, the variables that he found to be predictive of
success or failure were GPA and MCAT. He also discovered that the group that was at
risk of failure due to intrapersonal problems was the most likely to bounce back and
successfully complete training.

**Professional Competence**

To graduate as a competent consultant physician the physician, in training needs
to not only demonstrate a mastery of clinical knowledge and skills but master the five
core competencies (professionalism, systems based practice, practice based learning and
improvement, communication, and medical knowledge) as well. Frequently it is the soft-
competencies (or the non-cognitive) variables that get a resident into trouble but at the
same time are the most difficult to measure. While cognitive measures such as exam
scores can weed out poor performers and sometimes lead to dismissal or non-graduation,
poor performers due to soft competencies are difficult to determine and evaluate.

Eisenach, in a 2009 *Anesthesiology* journal article claimed that professionalism (one of
the principal core competencies) was more of an “I know it when I see it” kind of thing
rather than a hard measure, something that is measured by what not to do rather than
what to do. A 2004 study by St. George, Kaigas, and McAvoy determined that poorly
performing physicians are rarely identified and removed from practice and therefore
some measure of clinical competence was necessary. They determined that regular
assessment of individual faculty was both impractical and not always able to identify the
truly poor performers. They suggested that a screening tool for identification of
PREDICTING SUCCESS IN RESIDENCY

performance be implemented over the lifetime of a physician’s career, which could be significantly improved by the early identification, and perhaps remediation, of these deficiencies during residency.

Conclusion

The literature has strongly indicated that the best predictors of success in residency and/or medical school are cognitive variables such as previous performance on knowledge based examinations. Secondary to these variables are non-cognitive variables such as personality type, motivation, or approach to learning which while not accounting for significant variance as independent variables have tended to added significance in regressions when added into the analysis with other cognitive variables.

One limitation to the data and research studies reviewed in this chapter is the subjects being evaluated compared to the actual number of residents applying to residency. As Julian (2005) noted, most of the predictive models discussed above lack data from the full set of applicants because analysis is normally conducted only on admitted students/residents. She further goes on to note that any observed relationships (or predictive significance) are probably an understatement because applicants who were not selected to enter the program likely had lower scores on cognitive examinations.

A further limitation of this type of research is the collinearity or multicollinearity of the independent variables being used that can lead to high levels of standard errors. It is to be assumed that several of the independent variables are already highly correlated (such as Step 1 scores correlate to Step 2 scores or to AKT exam scores). Having highly correlated independent variables can mar the predictive significance of any one independent variable. To address the issue of multicollinearity I will use a variance
inflation factor (VIF) in the data analysis. VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient (the square of the estimate's standard deviation) is increased because of collinearity.

Lindblom-Yanne, Lonka, and Leskinen (1996) also adeptly noted that the problem with trying to identify a best fit prediction model is that the medical field trains residents to enter a wide variety of residency programs and professional fields, all of which require different skill sets and aptitudes. They suggested that it would be presumptuous to assume that any one single model could be successfully used to predict graduate medical education in all fields of medicine based on undergraduate performance.

There is always the limitation of creating too directed and linear of a prediction model that leads little room for individuality. If programs develop a precise model and become experts at predicting student success which in turn leads to the creation of more stringent application procedures will programs run the risk of losing applicants or not filling in the match? According to Siegert (2008) the answer is no, that the decision to apply is not dependent upon the rigor of the admission process. However researchers and institutions should always be cognizant that they may be losing a perfectly fantastic applicant whose cognitive performance indicators and actual performance are inaccurate.

If programs could make more accurate predictions of academic success they would have to deal with few, if any, expulsions in the middle of training (after much financial and personal investment on behalf of the individual and institution has been expended). But perhaps dealing with a few bumps in the road is what spurs change and
programmatic improvement. After all, it is in times of great distress that we often see the greatest amounts of innovation.
Methods

In this chapter, I will explain the methods employed to answer the two research questions. This methods chapter will be divided into several sections. First the design will be described and a rationale provided for its choice. Second, the data set will be described. Third, a justification for the method of analysis will be discussed and finally the analysis method will be described.

In order to answer the first research question of what are the independent variables that are associated with future AKT and In-service examination test performance of first year residents, a regression analysis was employed. The independent variables assessed included Gestalt survey scores, minority status, sex, and New Innovations Scores (variables included overall clinical competency, crisis management, professionalism, communications, problem based learning and improvement, systems based practice). Step 1 and Step 2 scores were used as control variables. Multiple regression (a linear regression, ordinary least squares (OLS)) was utilized to analyze data after a variance inflation factor analysis (VIF) was conducted in order to decrease the number of independent variables in the regression to a maximum of six. The number of independent variables per regression needed to be decreased because of the sample size; with a small sample size, too many independent variables will make it difficult to identify significance (Studenmund, 2011).

To investigate the second research question, what linear combination of factors, controlling for Step 1 and 2 examination scores, predicts success (not entering remediation and/or probationary status) during the first year of residency?, a linear regression (OLS) was conducted. The independent variables analyzed for this question
included Gestalt survey scores, minority status, sex, Step 1 and Step 2 scores, and New Innovations Scores (variables included overall clinical competency, crisis management, professionalism, communications, problem based learning and improvement, systems based practice). Independent variables were significantly reduced by conducting a primary analysis of collinearity and VIF so as to have no more than six independent variables in the regression.

**Research Design**

The study’s research design is quantitative and post-positivistic in that it is a correlation study that explores the relation between independent and dependent variables (Creswell, 2013). The research is exploratory analysis of a limited sample and involved secondary analysis of data about three cohorts (entering in 2008, 2009, 2010) of anesthesia residents at a major academic medical center in the Midwest during their respective first year in the residency. This particular design is appropriate for the research questions because it is an investigation into the relationship between variables with a set outcome measure. Quantitative analysis is appropriate for use because of the discrete measurable nature of the independent variables (data) available for analysis that is common across all participants and because quantitative analysis can easily examine relationships between those variables. In this instance, I investigated the statistically significant associations between individual resident outcomes based on a large set of independent variables. The sample for this study is composed of 67 residents over the course of 3 years (approximately 22 residents per annum out of a total population of 67 residents); these residents entered the anesthesia residency in 2008, 2009, and 2010. The total study sample includes 44 males and 23 females. The sample comprises 100% of
the eligible first year (CA-1) residents enrolled in the anesthesiology residency during the 3 year span listed above (the residency is 3 years in length and has approximately 22 residents in each class per year, with an occasional resident extending their training due to an alternative pathway (research) or because of extended leave, for a total of 67 residents).

The research takes on a post-positivistic outlook in its design because of several factors that include the breadth of the research, the use of data to investigate the validity of personal theories, and what prior literature has said about predictive analysis. The research study is also post-positivistic in nature because of the assumption by the researcher that any data collected is based on an initial theory. In this case I, the researcher, assume that the data and the interpretation of the data are co-dependent (Willis, 2007).

Population

The population for this study is all first year residents who have ever participated at the study institution. The sample for this study is composed of all first year residents who participated in the Anesthesiology Residency at the study institution for the years 2008-2010. Typically, a residency in anesthesiology is 4 years long, but the first year, the intern year, is a general medicine primer and does not need to be completed at the same institution. Therefore, anesthesiology residency is classified as a stand-alone intern year (Clinical Anesthesiology 0 year, CA-0) and a 3-year residency (CA1-3 years). All residents must complete board minimums and be certified as a competent anesthesiologists by the program to graduate. Successful completion of the residency means completing a minimum of varying rotations and minimum case numbers and
demonstrating that the individual meets all core competencies. Core competencies are graduate medical education standards in areas such as professionalism and systems based practice. While individual institutions have great leeway on how their curriculum is developed and delivered all graduates in the USA must meet the same minimum standards.

The sample included 67 residents who entered the anesthesiology residency in 2008, 2009, and 2010. The sample includes all CA-1 residents (first year residents) for the aforementioned years and was extracted from the total population of residents in the program. Each class has approximately 22 residents although class size fluctuates based on time in training and a required cap on enrollment by Medicare. The sample consists of residents who completed their medical training (M.D.) at a variety of institutions (with different types of training (e.g. PBL, lecture based), differing methods of evaluations, and of different sizes) across the US with a high percentage of trainees from the Midwest. Under-represented minorities in each class range from 7-20%. The sample to be used for the research study comprises 100% of the eligible first year residents enrolled in the anesthesiology residency during the 3 year span listed above.

This sample set was selected because of the uniformity of data available. For the entering class of 2008-2010 the same metrics were employed by the institution. These metrics include a uniform interview form, rules, regulations, and assessment types for advancement and promotion.

**Constructs and Data**

Data about the three cohorts of residents came from multiple sources; these sources and the variables themselves are described below.
Dependent Variables

Anesthesia Knowledge Test (AKT). Performance on the AKT exam was examined as a dependent variable via the use of multiple regressions to answer the first research question, what are the independent variables that are associated with future AKT and In-Service Examination test performance of first year residents? Performance was classified as an individual’s raw score on the examination. The score on this test reflects the level of anesthesia knowledge that an individual learner has at a given time compared to his/her peers. The higher the score the better the performance. At the institution being analyzed a score that is less than 25% of the national norm (score varies by year) would put a resident in risk of being placed into a remediation status.

The AKT is a nationally normed multiple choice examination that all residents in anesthesia take at set times. Residents at the Midwest Medical School take the examination at 1, 6, 12 and 24 months and their scores are normed against all other residents at the same level of training taking the examination within the same year. For the purpose of this study (and to avoid too much data collinearity) only the one month and 6 month AKT was used.

The AKT is independently created (by a national committee) and machine scored with no input from the faculty at the institution being studied. The Anesthesia Knowledge Test is continuously tested and evaluated for reliability and validity by a group of MDs called the Inter-Hospital Study Group for Anesthesia Education (HSGAE). The tests themselves are developed and revised every 2 years. The Inter-Hospital Study group’s principal approach to validation is content validity, which assumes the test questions represent the domain of knowledge appropriate for a one month resident in the case of the
AKT-1 or a sixth month resident in the case of the AKT-6 or a resident at 24 months in the case of the AKT-24. To ensure construct validity, the IHSGAE meets annually to review and update the examinations to ensure that they contain timely questions, in a proper balance. The reliability of the AKT instruments is also analyzed on an annual basis. For the AKT-1, the reliability as measured by Cronbach Alpha is consistently about .90/.91. (http://metricsinc.org/mancontents.htm)

**In-training Examination.** A second dependent variable in the regression analysis (used to address the first research question listed previously) is resident performance on the In-training exam (ITE). Achievement on the ITE is measured as an individual’s raw performance score. In-training exam scores are reported as scores ranging from 1-40 with a corresponding percentile rank identification for the 25, 50, 75th and 90th percentiles. Data used for this analysis were the reported score (from 1-40). Much like the AKT, the In-training exam is independently created (by the American Board of Anesthesiology) and machine scored with no input from the faculty at the institution being studied.

The In-training exam is a nationally normed multiple choice standardized test that residents take in each year of anesthesiology training. The In-training exam closely mirrors the written board examination and is representative of the comprehensive knowledge base that a resident will be required to have to become board certified upon completion of residency. For the ITE, internal consistency is checked to determine reliability. Testing internal consistency involves analyzing the test to determine if halves (split halves), containing equal numbers of items, correlate at a sufficient level. The overall reliability represents an average of the reliability coefficients obtained from all split halves (coefficient α). For the ITE, there is a strong correlation coefficient of α >0.8.
PREDICTING SUCCESS IN RESIDENCY

(McClintock & Gravlee, 2010). Furthermore the in-training exam has been found to be highly reliable as an evaluation measure and learning tool (Hall & Cotsonis, 1990). Residents complete the ITE in March of every year (about 9 months into their CA-1 training) and are required by the study institution to have a score that is over the 25% level of the nation to remain in good standing.

**Remediation/Probation.** A final dependent variable for this study is success as indicated by whether or not a resident is placed on remediation or probation during their first year of residency. The dependent variable of remediation/probation (indicated as a simple yes they were put on remediation/probation, or no they were not) was used to address the second research question of what linear combination of factors, including independent variables and test performance, predicts success (not entering remediation and/or probationary status) during the first year of residency?

The possibility for a resident to be placed into a remediation/probationary status occurs during the Housestaff Evaluation of Performance Review Committee (HEPRC) meetings (held three times a year at the approximate 4 month mark). The faculty members on the HEPRC committee consist of about 15 members at large from the total teaching faculty of about 105 members. Faculty on the HEPRC committee are selected as representatives from varying institutions and or rotations where residents will be assigned throughout their training career. Faculty on the HEPRC overlap with the faculty who interview/evaluate in New Innovations/complete the Gestalt survey. However many of the faculty who sit on the HEPRC committee would have not worked in-person with the study participants as they represent more advanced blocks/locations. The in-depth description for what qualifies a resident to be placed on remediation/probation is found in
chapter 1, but is basically a resident who has been found by a committee to be deficient in clinical competencies, professional competencies or both.

**Independent Variables**

New Innovations data (and their subset scores), Gestalt survey, interview scores, sex, minority stats, and Step 1 and 2 scores were used as independent variables in the linear regression analysis.

**New Innovations Resident Evaluation Database.** Eight points of data were pulled from the New Innovations evaluation database in order to address current/past resident performance. The measure of these variables are the mean scores from all faculty evaluations of the residents in each of the areas of (a) crisis management, (b) professionalism, (c) communication, (d) medical knowledge, (e) systems based practice, (f) practice based learning and improvement (PBLI), (g) patient care, and (h) overall competency.

The independent variables pulled from the New Innovations database are the first of several metrics that are independent evaluations of resident by faculty within the study department. The faculty who evaluate residents at the study institution total about 105, with an approximate male to female distribution during the study years of 55% to 45%, with about 8% under-represented minority population. The age range distribution of the faculty evaluators are approximately the following: a) 25-35 – 16% b) 36-45 – 21% c) 46-55 – 33% d) 55-62 – 23% and e) over 62 – 5%. About 40% of the faculty are consistent evaluators (complete regular monthly evaluations of residents), with the variance in number of evaluations arising for a number of reasons ranging from amount of contact with a high volume of residents to apathy in participating in the formal
PREDICTING SUCCESS IN RESIDENCY

evaluation process. Faculty from all rotations and all locations/rotations of training use the New Innovations evaluation database to give feedback on resident performance. All faculty who have contact with trainees at the institution are expected to evaluate a minimum of 4 residents per month so this evaluation metric has the greatest range of faculty (locations, specialty) giving input on resident performance.

The range of scores on the New Innovations evaluations is between 1 and 5, with one being the lowest score. Variables were selected to mirror the variables that the Clinical Competency Committee (the committee that makes decisions about resident performance) use in their quarterly evaluations of the residents and because these items are the core competencies in which the accrediting board requires all residents to be proficient upon graduation. Data on these variables were pulled quarterly (July-Oct, Nov-Feb, and Mar-June). Typically, multiple faculty assessed one resident during an evaluation period; with a representative resident receiving eight independent faculty evaluations that make up a mean score for the time period. This average score of all the independent evaluations of a given competency was used as a single independent variable. The independent variable of overall competency is not an average score of previously identified questions but its own independent question on the evaluation designed to measure overall performance by the resident on a given rotation.

Gestalt Survey. Residents’ average score on the Gestalt Survey (a “perception based or gut feeling” survey assessment of newly admitted residents administered by on-site faculty one month in to the new training year) was included as an independent variable. The Gestalt survey is distributed to faculty who have had contact with first year residents during the first month of training. These faculty are located at multiple sites and
work on a variety of training specialty rotations (e.g. the veterans hospital, birth center and the central hospital system) but excludes faculty who work at specialty locations (children’s hospital, neuroanesthesia etc.). Faculty who rate residents on the Gestalt survey have the potential to overlap with those who will later rate them on the New Innovations evaluation tool and could have possibly interviewed them in the prior year during recruitment season. The Gestalt survey is one that is not shared with residents and the purpose is not intended to be formal evaluation; rather the Gestalt survey is intended to classify residents early in training into quintiles to identify those in the lowest performing quintiles as an early ‘risk assessment’ metric. The Gestalt survey has three questions, the first two ask faculty to evaluate if the resident is a) above their peers, b) on par with peers, or c) below their peers in the areas of clinical performance (question 1) and medical knowledge (question 2). For the last question, faculty are asked to predict whether or not they feel that this resident will a) be a superior trainee, with leadership potential; b) be an average resident; or c) be at risk for failing, and need significant intervention. The resident was assigned a mean score on the Gestalt survey that is a combination of the score achieved for all three questions (points are allocated 1-low or below average, 2-mid-point indication average resident, and 3-high score indication above average for all three questions). The mean overall score on the Gestalt survey was used as an independent variable.

**Pre-Admittance Residency Interview Scores.** Finally, included as an independent variables were the mean scores resulting from the use of standardized questions during the interview for residency. Typically an applicant is interviewed by 4 independent faculty, and those faculty will be selected from a variety of locations where
residents will rotate during training. The program tries to ensure that a faculty member from each central teaching hospital will be interviewing each interview day. Faculty who interview prospective applicants will evaluate them again during training; however the time between the initial interview and consecutive evaluation can range from 1-3 years depending upon the primary location and specialty of the faculty interviewer.

Each interviewed applicant should have received a score from 1-5 in the areas of communication, professionalism, systems based practice, and practice based learning and improvement. The questions used for the standardized competency assessment questions were closely modeled after the Vanderbilt behavioral interview questions (Easdown, Castro, Shinkle, Small & Algren, 2005). Interview scores were analyzed as an average interview score, meaning that each individual interviewer’s average score was totaled and then averaged for an overall average interview score.

Sex. Whether or not a resident is male or female was examined to see if either of those variables accounted for any of the variance in the dependent variables. Sex was recorded as male or female and for the purpose of the analysis male was listed as 0 and female as 1.

Minority Status. The institution under investigation considers under-represented minorities to include individuals who self-report themselves as Black/African American, Mexican Americans/Hispanic, and/or Native-Americans. Individuals were identified as either 0 or 1 for minority status. If individuals received a score of 0 they were not an underrepresented minority (URM) and if they were labelled a 1 then they were a URM. Minority status is included as an independent variable in the analysis but the low total number of minorities in the cohort makes it difficult to infer significance in findings;
therefore the minority status will be reported as an independent variable with the caveat that no inferences will be made from the findings.

Control Variables

Step 1 and Step 2 scores. The United States Medical Licensing Examination (USMLE) exams consist of three parts, Step 1, Step 2, and Step 3 (Clinical Skills). The USMLE exams are intended to “assess a physician's ability to apply knowledge, concepts, and principles, and to determine fundamental patient-centered skills that are important in health and disease and that constitute the basis of safe and effective patient care” (USMLE, 2009). The United States Medical Licensing Examination (USMLE) board administers the Step Exams with the intended purpose to assess the knowledge base and whether medical school students or graduates understand and can apply important concepts of the sciences basic to the practice of medicine (STEP 1) and to assess critical thinking skills or whether medical school students or graduates understand and can apply important concepts of the sciences basic to the practice of medicine (STEP 2). Scores for the USMLE are reported as raw scores and as a Pass/Fail. The pass or fail rate is determined based on overall performance of the test takers in a given year. Minimum passing scores are periodically reevaluated and readjusted. The raw scores of each exam will be used as control variables.

Timing of Data Collection

The following Table (1) outlines all the independent and dependent variables and the time at which the data were collected and/or individual performance was assessed. Independent variables are labeled with an I and Dependent with a D and control variables are labeled with a C.
Table 1

*Collection Timetable for Dependent and Independent Variables used in Research Question 1*

<table>
<thead>
<tr>
<th>Entrance</th>
<th>Month 1</th>
<th>Month 4</th>
<th>Month 6</th>
<th>Month 8</th>
<th>Month 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1 (C)</td>
<td>Gestalt(I)</td>
<td>HEPRC (D)</td>
<td>AKT(D)</td>
<td>HEPRC(D)</td>
<td>In-Training (D)</td>
</tr>
<tr>
<td>Step 2 (C)</td>
<td>AKT(D)</td>
<td>NI Data(I)</td>
<td>NI Data(I)</td>
<td>HEPRC (D)</td>
<td></td>
</tr>
<tr>
<td>Interview Scores (I)</td>
<td></td>
<td></td>
<td></td>
<td>NI Data(I)</td>
<td></td>
</tr>
<tr>
<td>Sex (I)</td>
<td></td>
<td></td>
<td></td>
<td>AKT (D)</td>
<td></td>
</tr>
<tr>
<td>Minority Status (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HEPRC denotes the meetings of the performance evaluation committee where the New Innovations Data were evaluated and remediation/probationary status was determined. The New Innovations performance data (HEPRC notation) included faculty evaluations of resident performance on the areas of medical knowledge, patient care, systems based practice, life-long learning, crisis management, professionalism and overall competency. The overall competency variable in the New Innovations data set is not an aggregate of previous scores but the overall impression that faculty have of a resident’s performance on a given rotation. The interview scores are the scores from interviews conducted.
before acceptance in to the program and represented the average score given by the
interviewees (typically four faculty interviewers) on interview day. In-training exam
score is the resident performance on the standardized assessment (the In-training exam)
given to all residents in ACGME accredited programs.

Data Collection and Institutional Review Board

The investigator completed and sent the research proposal for review by the
Midwest Medical College IRB review board and the University of Missouri-Columbia
IRB review board. Approval to conduct the study was granted by both entities. The
project was deemed exempt with no risk of harm to the participants, as all data were
stripped of all identifying information.

The data itself were extracted from paper records (files) by departmental support
staff and entered in to the statistical software program, SPSS, where they were then
double checked by the researcher for accuracy and completeness. The data were
combined into one database, collated and “cleaned” for errors and/or completeness.
Where missing data were present, an attempt to locate the data was made. If data were
missing for an individual, an Expectation Maximization (EM algorithm) was run to
determine multiple imputation (MI) values for missing variables. EM algorithms are
especially suited for this type of analysis as MI performs very well with small sample sets
with large amounts of multiple regressions (Graham, 2009). Because of the availability of
data, no student records were deleted for non-completeness.

Analysis

Linear regression, as an ordinary least squares analysis was employed to address
the research questions.
RQ1. To address the first research question of “Which independent variables are associated with future test performance” linear regression analysis was used to determine significant variables for predicting resident performance on AKT and In-training exams. Regression was used to analyze the relation of the independent variables (NI scores, Gestalt survey, interview scores etc.) to the dependent variables (Anesthesia Knowledge Test (AKT) and in-training exam scores (ITE)) while analyzing the first research question. Multiple regression was used because it helps researchers understand how the value of the dependent variable can change when any one of the independent variables is varied and the other independent variables are held fixed. The regression analysis helps us to understand which among the independent variables are related to the dependent variable and to what extent that relationship is significant.

The independent variables that were evaluated included interview scores, Gestalt survey scores, New Innovations performance data, and In-Training exam scores, minority status, and sex. Step 1 and Step 2 scores were controlled for in the regression(s). As there are so many independent variables a variance inflation factor (VIF) was run to measure the severity of multicollinearity. Variables with a VIF greater than 5 were discarded before the regressions were conducted. Additionally a correlation table was run to see how closely correlated were the independent variables. Once correlations and significance of independent variables was conducted the multiple regressions were run with a maximum of six independent variables.

For research question 1, the following regressions were performed:

<table>
<thead>
<tr>
<th>Interview Score</th>
<th>1 MONTH AKT / 6 MONTH AKT / IN-TRAINING</th>
</tr>
</thead>
</table>
Running the regressions in this order allowed me to investigate each independent variable for its contribution. Progressing from the simple to the full models, with and without Step 1 and 2 scores as controls, allowed for the determination of significance the two principal variables gestalt and interview scores.

**Research Question Two**

To answer the second research question of what independent variables predict success (not entering remediation and/or probationary status) during the first year of residency a linear regression was employed. Because the sample size is small it can be difficult to conduct a robust analysis in instances where the outcome is binary (remediation/no remediation) therefore it was important to conduct a robust analysis. OLS was once again employed because it is a simple analysis method that can generate predictions or a model for the relationships between a dependent variable to a set of explanatory variables. OLS affords the researcher a higher power of the analysis but a greater chance for invalid standard errors and hypothesis tests (Studenmund, 2011).

The data that were analyzed for research question 2, as well as the dependent and independent variables are outlined in the following chart. Independent variables are labeled with an I and Dependent with a D and control variables are labeled with a C.
Table 2

Collection Timetable for Dependent and Independent Variables used in Research Question 2

<table>
<thead>
<tr>
<th>Entrance</th>
<th>Month 1</th>
<th>Month 4</th>
<th>Month 6</th>
<th>Month 8</th>
<th>Month 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>(C)</td>
<td>Gestalt(I)</td>
<td>Remediation (D)</td>
<td>AKT(I)</td>
<td>Remediation (D)</td>
</tr>
<tr>
<td>Step 2</td>
<td>(C)</td>
<td>AKT(I)</td>
<td>NI Data(I)</td>
<td>NI Data(I)</td>
<td>Remediation (D)</td>
</tr>
<tr>
<td>Interview Scores (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NI Data (I)</td>
</tr>
<tr>
<td>Sex (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minority Status (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the regression analysis the independent variables were analyzed against the dependent variables as follows:

\[
\begin{align*}
\text{Interview Score} & \quad \text{Remediation Status} \\
\text{Sex} & \quad \text{Remediation Status} \\
\text{Minority Status} & \quad \text{Remediation Status} \\
\text{Step 1} & \quad \text{Remediation Status} \\
\text{Step 2} & \quad \text{Remediation Status} \\
\text{Interview} & \quad \text{Remediation Status}
\end{align*}
\]
By running these regressions, I was able to determine whether any of the independent variables significantly contribute to the likelihood of being put on remediation/probation during the first year of residency.

**Limitations and Conclusions**

The generalizability of the findings will be somewhat limited due to the fact that the data are from one institution and a single anesthesiology program. However, due to the flexibility that programs are given to manage and direct their programs (with no
standardized assessment tools or metrics) analyzing multiple programs across the country would result in too many confounding factors that could not be controlled, thereby compromising the validity of results. There was also multicollinearity in the data set, as several independent variables are correlated to each other. For example first quarter professionalism scores were highly correlated with the second quarter crisis management score. A correlation matrix, Appendix A, will be discussed in detail in the next chapter. Repeated measure bias is also a limitation of the study because while there are several independent evaluation variables there is a limited faculty set doing the evaluations; therefore the potential for the same faculty to be rating the same trainee on different independent variables exists.

Several limitations exist for question two. For example, while all ACGME accredited programs must follow the same standards, much leeway is present in how particular programs test the proficiency of its residents or decide their good standing/remediation or probationary status. In addition a limitation to using the data set for a count analysis of successful performance is the very small number of residents who are actually dismissed. Dismissals are rare due to reasons that include the rules and policies of the ACGME, accrediting boards, and residency programs and also the litigious nature of many individuals. That being said, this study provides some insightful information on significant variable that most Anesthesiology programs can apply to their own data sets.

In the next chapter I will report on the results of the data analysis in accordance to the research questions.
Results and Findings

This study was designed to investigate the relationship between individual performance and non-performance metrics and success in residency. Specifically, this study was designed to answer the following research questions:

1) Which independent variables are associated with future test performance?
2) What independent variables predict success (not entering remediation and/or probationary status) during Residency?

Each research question was addressed using separate data analysis. Research question one and two employed linear regression for statistical analysis.

Each participant in the cohort had an almost complete data set, which was made complete by using missing data procedures prior to analysis. That is, each individual had a data set that indicated his/her sex, minority status. Step 1, Step 2, gestalt, interview, AKT, In-Training, and New Innovations scores of crisis management, professionalism and overall performance.

Population Descriptions

Data from 67 participants, representing the complete training class each year, were used in the research study; 24 for 2008-09 year, 19 for 2009-10, and 24 for the 2010-11 academic year. The inconsistency in the class sizes across years is because an incoming class (CA-1) size is dependent upon the class being replaced. So while typically 22 spots per residency are available each year the number can increase or decrease based on individuals who leave the program and create an unexpected vacancy, take longer to graduate and use a spot for longer than anticipated, or stay longer because of personal leaves of absence. The program has 66 funded spots over the 3 year residency so
PREDICTING SUCCESS IN RESIDENCY

administrators try and maintain that total number. This means that class size can vary from the allotment amount of 22. Table 3 outlines the descriptive variables of the Cohort.

Table 3

Descriptive Statistics of Study Population

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>URM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008-09</td>
<td>17</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>2009-10</td>
<td>13</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2010-11</td>
<td>14</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>23</td>
<td>6</td>
</tr>
</tbody>
</table>

The cohort consisted of 44 men and 23 women. A total of six underrepresented minorities (URM) (African American, Hispanic and American Indian) were in the three cohorts spread out over the years with one in the 08-09 academic year, one in the 09-10 academic year, and four in the 2010-11 academic year.

While the data available for the study were robust with almost complete data sets for each individual, in instances where data (primarily with the interview score variable) were missing, adjustments were made as described later for those individual variables.

Dependent Variables

Three multiple choice exams were used as the dependent variables in the investigation of research question one while remediation status was the dichotomous dependent variable for the second set of analysis. Table 4 outlines the respective minimum, maximum and average scores for each variable as well as their standard
deviation with the exception of remediation status. A detailed description of each variable follows.

Table 4

*Descriptive analysis of key dependent variables*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT 1 Month</td>
<td>80%</td>
<td>11%</td>
<td>99%</td>
<td>24.15</td>
</tr>
<tr>
<td>AKT 6 Month</td>
<td>63%</td>
<td>6%</td>
<td>99%</td>
<td>25.84</td>
</tr>
<tr>
<td>In-Training Exam</td>
<td>26</td>
<td>19</td>
<td>37</td>
<td>4.39</td>
</tr>
</tbody>
</table>

**AKT 1 Month Examination.** The one-month Anesthesiology Knowledge Test (a national standardized exam given to all anesthesiology residents at set time intervals) had a 100% participation rate over the three years with the minimum score being 11% (scores reported as a percentile rank against all test takers in the US) and the maximum score 99% with the average score being 80%. Percentile rank of a score is the percentage of scores in a frequency distribution that are the same or below it; therefore a score of 50% on the AKT 1 month examination would indicate that an individual had scored better than 49% of other test takers for that specific examination. See Figure 1 for the distribution of scores for the AKT 1 month examination

54
Figure 1: Histogram of AKT 1 month percentile rank scores for study cohort.

**AKT 6 Month Examination.** The 6 month Anesthesiology Knowledge Test reports scores as a percentile rank and residents must maintain a ranking of greater or equal to the 25\textsuperscript{th} percentile (score must be fall in the 25-99 percentile range) to be in good standing. The AKT had a 100% participation rate over the 3 years with the minimum score being 6\% and the maximum score 99\% with the average score being at the 63\textsuperscript{rd} percentile. Figure 2 illustrates the distribution of the AKT 6 month examination scores for the study population.

Figure 2: Histogram of AKT 6 month percentile rank scores for study cohort.
In-Training Examination. The In-Training Exam (ITE) is the national normed performance exam given to all Anesthesiology residents each spring during the three years of residency training. In-training exam scores are reported as scores ranging from 1-40 with a corresponding percentile rank identification for the 25, 50, 75th and 90th percentiles. The In-training exam scores for this cohort, illustrated in Figure 3, ranged from a low of 19 to a high of 37 with the mean being 26 for all 3 years. Historically, at the national level, a score of 24 will put a resident at the 25th percentile, 27 at the 50th percentile and 33 at the 90th percentile.

![In-Training Exam Score Distribution](image)

Figure 3: Histogram of In-Training scores for study cohort. Scores have a possible range of 1-40.

Remediation Status. The dependent variable of remediation status was a dichotomous variable of yes/no based on whether or not a resident was placed onto a remediation or probationary status within his/her first year of residency training. Remediation status was coded as a 1 for good standing and 0 for remediation. For the cohort, over the 3 years, a total of 13 (about 19%) individuals were placed into remediation/probation.
PREDICTING SUCCESS IN RESIDENCY

Key Independent Variables

Step 1, Step 2, interview, Gestalt, professionalism, crisis management and overall clinical competency were the key independent variables used in the investigation. Table 5 outlines their mean, minimum, maximum and standard deviations with an in-depth analysis to follow.

Table 5

*Descriptive analysis of key independent variables*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>217</td>
<td>183</td>
<td>251</td>
<td>17.9</td>
</tr>
<tr>
<td>Step 2</td>
<td>223.5</td>
<td>188</td>
<td>264</td>
<td>18.8</td>
</tr>
<tr>
<td>Interview</td>
<td>7.27</td>
<td>3.25</td>
<td>9.9</td>
<td>1.46</td>
</tr>
<tr>
<td>Gestalt</td>
<td>6.64</td>
<td>4</td>
<td>9</td>
<td>1.18</td>
</tr>
<tr>
<td>Professionalism</td>
<td>4</td>
<td>2.43</td>
<td>5</td>
<td>.47-.52</td>
</tr>
<tr>
<td>Crisis Management</td>
<td>3.65-4</td>
<td>2.25</td>
<td>5</td>
<td>.48-.55</td>
</tr>
<tr>
<td>Overall Competence</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>.46-.60</td>
</tr>
</tbody>
</table>

Step 1. Step 1 scores for each study participant were used in the analysis of both research questions. Because the Step 1 score is a cognitive exam score that is normed with a national mean and is similar in nature to the dependent variables for research question one (that are also normed cognitive exams) it was also used as a control variable to examine its potential effect on results. Step 1 scores, demonstrated in Figure 4, in the cohort ranged from a minimum of 183 to a maximum of 251, the mean and median
scores of the three cohorts were both 217 with a standard deviation of 17.9. Individuals with Step 1 scores of lower than 200 are generally not invited for an interview. However, if other variables are superior (e.g., strong letters of reference, an individual with strong ties to the Midwest), individuals may still gain a request for an interview because they meet screening minimums in a different way. Eight CA-1 residents in the cohort had Step 1 scores lower than the 200 baseline.

**Figure 4:** Histogram of Step 1 raw scores for study cohort.

**Step 2.** Step 2 scores, illustrated in Figure 5, in the cohort ranged from a minimum of 188 to a maximum of 264; the mean and median score of the three cohorts was 223 and 224 respectively with a standard deviation of 18.8. Step 2 scores do not weigh heavily in the decision to interview an applicant as they are often not completed prior to submitting applications to residency (but are required prior to the start of the residency). Typically, the program being investigated expects an upward trend between the Step 1 and Step 2 scores and will look more closely at instances of declining Step 2 scores during the interview process.
Interview Scores. Interview scores as an independent variable consisted of the average of the overall interview score given to each applicant on the day he/she conducted his/her in-person interview for residency. Figure 6 illustrates the distribution of interview scores for the study population. The rating scale for interviews was 1-10 with a 10 being the highest score possible. Typically, each individual would be evaluated by four separate interviewers with the independent scores being averaged. Each interviewer is advised to pay attention to several competencies that include communication skills, professionalism, enthusiasm for program and genuine interest in specialty, and ability to work in a team. The lowest interview score was 3.25 and the highest was 9.9, with the average score being 7.27.
Figure 6: Histogram of Interview score distribution for study cohort. Interview scores are averaged across 4 separate raters.

An interview score was not available for every study participant because about 20% of the residents who matched with the program did so in an ‘unconventional’ way. Most often this was an outside-of-the-match position, which did not have the same formal interview process as the residents who matched into the program through the traditional manner. Typically residents go through a process called a “match” to get into residency; this is a formal application process involving submitting materials electronically to online databases to multiple institutions and programs. When individuals enter a match, they are bound by the rules of the match which include honoring a commitment to whatever program with which they are matched. The match process is time consuming; residents are selected to interview by programs, personal interviews are conducted, and then both the individual and institutions rank their selection preferences. A match occurs when both the individual and institution ‘match’ by a computerized prioritization process. Occasionally people enter the program ‘outside-the-match;’ these individuals are sometimes non-traditional learners who are coming from other specialties, research
backgrounds, or are foreign medical graduates. Outside-of-the-match applicants are still interviewed but typically instead of meeting with four faculty members they will meet only with the program director and therefore lack a formal interview score.

To address the missing interview variable, for those who were outside of the match, a dummy variable called missing was created and in the instance in which individuals had no interview score they were assigned a missing score of 1; conversely individuals who had an interview score were given a missing value of 0. By creating this dummy variable I hoped to account for any artificially inflated significance of the interview score that might arise in the analysis.

**Gestalt Survey.** The Gestalt survey is administered one month into residency and is a short (3 questions) written assessment of the residents by faculty who have worked with them to date. The Gestalt survey consists of the following three questions and checkboxes: a) Based on your initial experiences with this resident would you rate their clinical performance: above, on par, below their peers. b) their medical knowledge: above, on par, below their peers; and c) prediction: do you believe this resident will: be a superior trainee with leadership potential, be an average resident, or be at risk of failing, need significant intervention? Each checkbox is designated a point value (3 being the best, 1 being the worst) and a resident’s score is totaled and then averaged depending on how many evaluations he/she receives. The average score is used in their overall rank for the Gestalt score. Figure 7 illustrates the distribution of Gestalt scores for the study population.
Figure 7: Histogram of Gestalt score distribution for study cohort. Gestalt scores are an average score from multiple raters who complete and pencil and paper survey.

Residents received between 9-14 independent evaluations from which an overall average score for the Gestalt survey metric was calculated. The range of scores for the Gestalt survey was a low of 4 to a high of 8.8, with the mean score falling right around 6.64.

**Professionalism.** The metric of professionalism is derived from the standard online evaluation that faculty complete about residents based on monthly rotations in the online New Innovations curriculum management program. The online tool is designed to capture many metrics including overall performance and performance on the core competencies (professionalism, communication, systems based practice, medical knowledge, and patient care) over the course of a rotation month. Faculty are requested to complete an evaluation on a resident with whom they had significant contact at least once per month but have the opportunity to complete a daily evaluation if they desire.

Professionalism is one of the core competencies as defined by the American Council for Graduate Medical Education (ACGME) and is evaluated via a standard Likert scale on the following three questions on the College’s standard assessment of residents:
PREDICTING SUCCESS IN RESIDENCY

1) Available, responsive, on time.
2) Demonstrates respectful, compassionate and ethical behavior. Accepts responsibility for actions.
3) Demonstrates sensitivity and responsiveness to individual patient needs. Puts patient above oneself (altruistic).

The professionalism score used as an independent variable in this study is the averaged professionalism score (across all questions and raters) for all scores given during each of the 3 quarters in the academic year. The professionalism score was evaluated for the first (Q1), second (Q2), and third (Q3) quarters of the first academic year. Professionalism scores are measured on a 1 to 5 scale with 5 being the highest and most desirable score. The mean professionalism score for all 3 quarters investigated was 4, with a range of 2.43 to 5. The standard deviation for the three quarters ranged from .47 (Q2) to .52 (Q1). Residents who scored two standard deviations or more below the mean were likely to be reviewed by the competency committee for possible probationary/remediation status in accordance with departmental policy.

**Crisis Management.** The metric of crisis management is another variable captured from New Innovations surveys that evaluate a residents’ performance each month. Crisis management is considered an essential attribute in Anesthesiology and while not considered a standard core competency, it falls within the domain of patient care and is regularly measured by the following question on the monthly tool, “Responds rapidly and appropriately to critical situations. Demonstrates sound clinical judgment.” The crisis management score used as an independent variable in this study is the averaged crisis management score (across all raters) for independent individual evaluation quarters.
The crisis management score was developed from averaged evaluation scores received for all three quarters (Q1), (Q2), and (Q3) of the first academic year and were added to the analysis as three separate independent variables. Crisis management scores are measured on a 1 to 5 scale; the mean crisis management score for all three quarters ranged from 3.65 to 4, with a range of lowest to highest scores for the quarters being 2.25 to 5. The standard deviation for the 3 quarters ranged, .48 (Q2), to .55 (Q1). Residents who scored two standard deviations or more below the mean were likely to be reviewed by the competency committee for possible probationary/remediation status in accordance with departmental policy.

**Overall Clinical Competence.** Overall clinical competence is the final metric that was extrapolated from the core competency New Innovations evaluations. This score was derived from one question on the evaluation form that asks about the overall performance of the resident on the rotation for which he/she was being evaluated. This metric consists of the average of all completed overall evaluations within the three quarters of the CA-1 academic year for each subject. The mean for overall clinical competency was 4.0, 4.0, and 3.9 (Q1, Q2, and Q3 respectively) and the range was from a low of 2 to a high of 5. The standard deviation for each of the 3 quarters ranged from .46 (Q2), to .60 (Q1). Residents who scored two standard deviations or more below the mean were likely to be reviewed by the competency committee for possible probationary/remediation status in accordance with Departmental policy.
Research Questions

Question 1

Linear regression was used to investigate the research question, which independent variables are associated with future test performance? The independent variables of sex, minority status, Step 1 and Step 2 scores, interview scores, and the Gestalt survey were investigated against the dependent variables AKT 1, AKT 6 and In-Training Examination. Due to their similarity to the dependent variables (standardized multiple choice knowledge tests) Step 1 and Step 2 scores were controlled for in each regression to determine if they were explaining too much variance; this was done by adding them into the regression after analysis of all other variables were completed.

Analysis was conducted as a series of linear regressions with each regression adding additional variables. The purpose for running the regressions this way was to determine the amount of variance for each independent variable. Regressions were run using Interview and Gestalt scores as the anchor variables, and then once again stepwise with all variables. Regressions were run by adding variables in the following order: interview/gestalt, sex/minority, and then Step 1 and 2.

Interview scores and Gestalt scores are variables that the department develops and administers so we were especially interested in reviewing their strength in the analysis. Thus the reason for conducting regressions first with the interview and gestalt independent variables followed by the addition of the descriptive variables of sex and minority status, and finally the test scores from the Step examinations.

In addition, a regression was run that included data that were collected during the first three months of residency (professionalism, crisis management and overall
performance score) and analyzed against the AKT 6 month and In-training exam. For this regression it was imperative to leave out the AKT 1 month exam as a dependent variable because that examination occurred prior to the collection of data for the additional independent variables and would confound results.

**Results of Analysis with Interview/Missing variable.** The following three tables (Tables 6-8) include the results of regressions run with the independent variables of interview, missing, sex, minority, and Step 1 and 2 and their relationship to the dependent variables of Step 1, Step 2 and In-Training exam scores.

**AKT 1 Month Examination.** The interview score, Step 1 and 2 scores or sex variables were not found to account for variance in any of the regressions for the AKT 1 month examination. However the variable of minority status did account for some of the variance when entered into the regression ($b = -28.50, p = .006$, Table 7, $b = -29.02, p = .008$, Table 8) indicating that URM individuals performed worse than non-minority residents on the examination

**AKT 6 Month Examination.** Interview scores were found to account for variance ($b = 6.17, p = .013$, Table 6, $b = 5.33, p = .03$, Table 7) for the AKT 6 month exam but lost significance when Step 1 and 2 scores were entered into the regression. Step 1 scores accounted for some of the variance in the regression as well ($b = -.47, p = .01$, Table 8) while seemingly reducing the impact of the interview score variable. Once again minority status continued to account for variance with the AKT 6 month examinations ($b = -27.04, p = .012$, Table 7, $b = -19.02, p = .05$, Table 8) indicating that under-represented minorities performed worse on the examination.
In-Training Examination. Only the Step 1 variable was able to account for any variance in the regressions with the dependent variable of In-Training examination and only once all independent variables were entered into the regression (b=.08, p=.03, Table 8). The remaining variables were unable to account for any variance.

Table 6

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variable of Interview Scores and Missing

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th></th>
<th>AKT 6 Month</th>
<th></th>
<th>In-Training Exam</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>49.841</td>
<td>17.349</td>
<td>12.978</td>
<td>17.88</td>
<td>22.668</td>
<td>3.26</td>
</tr>
<tr>
<td>Interview</td>
<td>3.227</td>
<td>2.342</td>
<td>.450</td>
<td>6.169</td>
<td>2.41**</td>
<td>.806</td>
</tr>
<tr>
<td>Missing</td>
<td>24.346</td>
<td>18.371</td>
<td>.433</td>
<td>42.147</td>
<td>18.94*</td>
<td>.702</td>
</tr>
<tr>
<td>R²</td>
<td>.029</td>
<td>.094</td>
<td>.023</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

Table 7

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variable of Interview, Missing, Sex, and Minority

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th></th>
<th>AKT 6 Month</th>
<th></th>
<th>In-Training Exam</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>61.21</td>
<td>17.63</td>
<td>23.049</td>
<td>18.38</td>
<td>24.35</td>
<td>3.44</td>
</tr>
<tr>
<td>Interview</td>
<td>2.39</td>
<td>2.29</td>
<td>.34</td>
<td>5.33</td>
<td>2.39*</td>
<td>.697</td>
</tr>
<tr>
<td>Missing</td>
<td>19.75</td>
<td>17.93</td>
<td>.35</td>
<td>37.35</td>
<td>18.68*</td>
<td>.62</td>
</tr>
<tr>
<td>Sex</td>
<td>-2.33</td>
<td>6.03</td>
<td>-.05</td>
<td>-1.38</td>
<td>6.28</td>
<td>-.026</td>
</tr>
<tr>
<td>Minority</td>
<td>-28.50</td>
<td>10.01**</td>
<td>-.34</td>
<td>-27.05</td>
<td>10.43*</td>
<td>-.30</td>
</tr>
<tr>
<td>R²</td>
<td>.144</td>
<td>.184</td>
<td>.083</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
Table 8

*Results of Analysis with Gestalt score variable.* The next sets of regressions (Tables 9, 10, 11) were run with the independent variables gestalt, sex, minority, and Steps 1 and 2; results of the regressions will be reported together by their corresponding dependent variable.

**AKT 1 Month Examination.** Gestalt scores were found to account for some of the variance for the AKT 1 month exam regression, but only when regressed independently (b=5.39, p=.031, Table 9). Minority Status also accounted for some variance when added into the regressions (b= -27.15 p=.007, Table 10, b=-28.10, p=.008, Table 11) but the negative b value indicates that minority students perform worse on the exam than their non-URM counterparts. Sex, Step 1 and Step 2 variables did not account for any variance in any regression with the AKT 1 month examination.

**AKT 6 Month Examination.** Gestalt scores, sex and Step 2 scores did not account for any variance when regressed against the AKT 6 month examination. On the other hand,

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th></th>
<th>AKT 6 Month</th>
<th></th>
<th>In-Training Exam</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>66.57</td>
<td>42.58</td>
<td>-125.0</td>
<td>39.03</td>
<td>-1.69</td>
<td>7.09</td>
</tr>
<tr>
<td>Interview</td>
<td>2.50</td>
<td>2.41</td>
<td>.35</td>
<td>3.41</td>
<td>2.21</td>
<td>.45</td>
</tr>
<tr>
<td>Missing</td>
<td>20.41</td>
<td>18.70</td>
<td>.36</td>
<td>21.78</td>
<td>17.14</td>
<td>.36</td>
</tr>
<tr>
<td>Sex</td>
<td>-2.28</td>
<td>6.15</td>
<td>-.05</td>
<td>-1.42</td>
<td>5.63</td>
<td>-.03</td>
</tr>
<tr>
<td>Minority</td>
<td>-29.02</td>
<td>10.59**</td>
<td>-.35</td>
<td>-19.02</td>
<td>9.71*</td>
<td>-.21</td>
</tr>
<tr>
<td>Step 1</td>
<td>0.00</td>
<td>.20</td>
<td>.00</td>
<td>0.47</td>
<td>.18**</td>
<td>.315</td>
</tr>
<tr>
<td>Step 2</td>
<td>-0.03</td>
<td>.19</td>
<td>-.02</td>
<td>0.27</td>
<td>.18</td>
<td>.20</td>
</tr>
<tr>
<td>R²</td>
<td>.144</td>
<td>.369</td>
<td>.285</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
hand, minority status accounted for variance in both regressions in which it was included (b= -29.98, p=.006, Table 10, b= -19.91, p=.05, Table 11). Finally Step 1 scores accounted for some variance when entered into the regression with b=.48, p=.014.

**In-Training Examination.** The Gestalt score variable accounted for some variance in all three regressions with the dependent variable of In-Training examination and seemed to gain strength the more variables entered into the regression (b=1.11 p=.016 , Table 9, b=.99 p=.033, Table 10 and b=1.03 p=.01, Table 11). The Step 1 variable accounted for some of the variance when entered into the regression (b=.10, p=.003, Table 11) and the low standard error indicates a lack of multicollinearity with other variables. Minority status, sex, or Step 2 variables were not found to account for variance for the In-training exam scores.

Table 9

*Regressions Run with Dependent Variables AKT 1, AKT 6, and In-Training Exam with Independent Variable of Gestalt*

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th>AKT 6 Month</th>
<th>In-Training Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>37.69</td>
<td>16.53</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>5.39</td>
<td>2.45*</td>
<td>.26</td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
Table 10

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt Scores, sex, minority

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th>AKT 6 Month</th>
<th>In-Training Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>46.83</td>
<td>16.52</td>
<td>.51</td>
</tr>
<tr>
<td>Gestalt</td>
<td>4.39</td>
<td>2.38</td>
<td>.22</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.20</td>
<td>5.81</td>
<td>-.00</td>
</tr>
<tr>
<td>Minority</td>
<td>-27.15</td>
<td>9.74**</td>
<td>-.32</td>
</tr>
</tbody>
</table>

R²         | .172        | .121        | .132            |

*Significant with a p value of ≤.05  **Significant with a p value of ≤.01

Table 11

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt Scores, sex, minority, Step 1 & 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th>AKT 6 Month</th>
<th>In-Training Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>42.71</td>
<td>43.31</td>
<td>121.70</td>
</tr>
<tr>
<td>Gestalt</td>
<td>4.85</td>
<td>2.53</td>
<td>.24</td>
</tr>
<tr>
<td>Sex</td>
<td>0.22</td>
<td>5.94</td>
<td>.00</td>
</tr>
<tr>
<td>Minority</td>
<td>-28.10</td>
<td>10.30**</td>
<td>-.34</td>
</tr>
<tr>
<td>Step 1</td>
<td>0.11</td>
<td>.20</td>
<td>.08</td>
</tr>
<tr>
<td>Step 2</td>
<td>-0.09</td>
<td>.19</td>
<td>-.08</td>
</tr>
</tbody>
</table>

R²         | .177        | .344        | .347            |

*Significant with a p value of ≤.05  **Significant with a p value of ≤.01

Results of Analysis with gestalt score and interview/missing variable. Tables, 12, 13 and 14 display the results of the regressions run with both the gestalt and interview variables together; results of the regressions are reported merged by their corresponding dependent variable.

AKT 1 Month Examination. The single variable that accounted in some variance in this set of regressions was minority status. Minority status, b=-26.85, p=.01, indicated that students who were URM’s were more likely to score poorly on the AKT 1 month
examination. It is important to note that while minority status showed several instances of significance in the regressions that the small n of URM limits the generalizability and usefulness of the findings; the minority variable findings have limited meaning given the n and should be considered as a control variable. The remaining variables of Gestalt score, sex, interview score, Step 1 and Step 2 accounted for no variance in prediction of success on the AKT 1 month exam.

**AKT 6 Month Examination.** Interview scores accounted for some variance when regressed with Gestalt scores \( (b=6.11, p=.03, \text{Table 12}) \) and again when the sex and minority variables were entered into the regression \( (b=5.49, p=.03, \text{Table 13}) \) but accounted for no variance when Step 1 and 2 scores were entered into the analysis. Minority status accounted for variance in the regressions \( (b=-19.03, p=.05, \text{Table 13}) \) and remained that way when Step 1 and 2 scores were included \( (b=-19.03, p=.05, \text{Table 14}) \), once again demonstrating (with the limited n caveat) that URM’s were likely to perform worse on the examination. The Step 1 variable accounted for some variance when it was entered into the regression \( (b=.47, p=.02, \text{Table 14}) \) and the low standard error suggests that it lacks multicollinearity with other independent variables. Gestalt scores, sex and Step 2 score variables accounted for no variance in the AKT 6 month analysis.

**In-Training Examination.** Gestalt scores accounted for variance in each of the regressions in this analysis \( (b=1.12, p=.03, \text{Table 12}, b=1.02, p=.01, \text{Table 13}, b=1.13, p=.01, \text{Table 14}) \) without losing significance as variables were entered into the analysis. The Step 1 score variable also accounted for some variance when entered into the regression \( (b=.09, p=.006, \text{Table 14}) \). The remaining variables of interview scores, sex, minority status or Step 2 did not account for any variance in this set of regressions.
PREDICTING SUCCESS IN RESIDENCY

Table 12

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt, Interview, and Missing

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th></th>
<th>AKT 6 Month</th>
<th></th>
<th>In-Training Exam</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>26.37</td>
<td>21.07</td>
<td></td>
<td>11.91</td>
<td>22.33</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>4.94</td>
<td>2.61</td>
<td>.24</td>
<td>0.23</td>
<td>2.77</td>
<td>.01</td>
</tr>
<tr>
<td>Interview</td>
<td>1.89</td>
<td>2.40</td>
<td>.26</td>
<td>6.11</td>
<td>2.55**</td>
<td>.80</td>
</tr>
<tr>
<td>Missing</td>
<td>16.46</td>
<td>18.49</td>
<td>.29</td>
<td>41.79</td>
<td>19.59*</td>
<td>.70</td>
</tr>
<tr>
<td>R²</td>
<td>.081</td>
<td></td>
<td></td>
<td>.094</td>
<td></td>
<td>.104</td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

Table 13

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam with independent variables of Gestalt, Interview, Missing, Sex, and Minority

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 1 Month</th>
<th></th>
<th>AKT 6 Month</th>
<th></th>
<th>In-Training Exam</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>38.66</td>
<td>20.74</td>
<td></td>
<td>24.49</td>
<td>22.08</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>4.18</td>
<td>2.55</td>
<td>.20</td>
<td>-0.58</td>
<td>2.71</td>
<td>-.03</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.93</td>
<td>6.01</td>
<td>-.02</td>
<td>-1.57</td>
<td>6.40</td>
<td>-.03</td>
</tr>
<tr>
<td>Minority</td>
<td>-26.85</td>
<td>9.93**</td>
<td>-.32</td>
<td>-27.28</td>
<td>10.56**</td>
<td>-.30</td>
</tr>
<tr>
<td>Interview</td>
<td>1.22</td>
<td>2.38</td>
<td>.17</td>
<td>5.49</td>
<td>2.53*</td>
<td>.72</td>
</tr>
<tr>
<td>Missing</td>
<td>12.68</td>
<td>18.20</td>
<td>.23</td>
<td>38.34</td>
<td>19.38*</td>
<td>.64</td>
</tr>
<tr>
<td>R²</td>
<td>.180</td>
<td></td>
<td></td>
<td>.185</td>
<td></td>
<td>.148</td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
Table 14

Regressions Run with Dependent Variables AKT 1, AKT 6, and IN-Training Exam and independent variables of Gestalt, Interview, Missing, Sex, Minority, Step 1, and Step 2

| Variable | AKT 1 Month | | | AKT 6 Month | | | In-Training Exam | |
|----------|-------------|------------------|------------------|------------------|------------------|------------------|
|          | B           | Std. Error       | β                | B               | Std. Error       | β                | B               | Std. Error       | β                |
| Constant | 40.41       | 43.99            | -126.07          | 41.29           | -8.16            | 7.11             |
| Gestalt  | 4.53        | 2.67             | .22              | -0.07           | 2.51             | -.00             | 1.13            | .44**            | .30              |
| Sex      | -0.57       | 6.14             | -.01             | -1.45           | 5.76             | -.03             | -0.00           | .99              | .00              |
| Minority | -27.95      | 10.45*           | -.33             | -19.03          | 9.81*            | -.21             | -1.93           | 1.69             | -.13             |
| Step 1   | 0.08        | .20              | .06              | 0.47            | .19*             | .31              | 0.09            | .03**            | .37              |
| Step 2   | -0.10       | .19              | -.08             | 0.27            | .18              | .20              | 0.04            | .03              | .17              |
| Interview| 1.35        | 2.47             | .19              | 3.43            | 2.32             | .45              | -0.28           | .41              | -.22             |
| Missing  | 12.88       | 18.94            | .23              | 21.89           | 17.78            | .36              | -1.15           | 3.12             | -.11             |
| R²       | .184        | .369             | .360             |                  |                  |                  |                  |                  |                  |

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

Results of Analysis with top 5 independent variables. As a final evaluation tool for question 1, I conducted an analysis of the top 5 independent variables after a variance inflation analysis was completed. The purpose of this regression was to determine if any additional information collected during the first year of residency could explain more variance in the regressions to predict performance on the dependent variables. However, because of the number of independent variables available for inclusion in this analysis, and to ensure a robust test, two steps were taken to ensure multicollinearity was limited in the analyses: 1) a correlation matrix was created and 2) a variance inflation analysis was conducted.

To test level of collinearity a correlation matrix was created and is illustrated in Appendix A. From these analyses, I can see that the interview variable is not significantly correlated with any other independent variable but that Gestalt scores are strongly correlated to quarters 1-3 crisis management, professionalism and overall competency.
scores. Furthermore minority status was found to have a strong association with performance on the second quarter crisis management and professionalism variables and a weaker one with Step 2 and second quarter overall competency scores. Finally, Step 1 scores were found to be correlated only with Step 2 scores, while the sex variable was not correlated with any other independent variable. Keeping these relationships in mind, to ensure that multicollinearity was limited in the analysis I conducted a variance inflation factor analysis (VIF) and only variables with VIFs of less than 5 were included in the analysis.

In this set of regressions a variance inflation factor analysis was completed to reduce the number of variables (from the possible variables of professionalism Q1, Q2, Q3, crisis management Q1, Q2, Q3, overall performance Q1, Q2, Q3, Step 1 & 2, minority status, interview scores, and Gestalt scores) to those with enough significance (variance inflation factor of less than 5) to regress. Variance inflation factor (VIF) is an analysis designed to assess the amount of multicollinearity in an ordinary least squares (OLS) type regression. VIF measures how much the variance, or the square of the estimates standard deviation, in a regression is increased because of collinearity. In general a VIF of >5 means that multicollinearity is high so for this analysis significance was determined as a VIF of less than 5.

The variables that had a VIF of less than 5, and were therefore included in this set of regressions were, interview scores, Step 1, Step 2, Gestalt, and 1st quarter professionalism against the dependent variables of AKT 6 month and 1 year in-training exam with the results reported in Table 15. Please note that the AKT 1 month could not be used in this analysis as a dependent variable because several independent variables
(such as first quarter professionalism, crisis management) were generated after the AKT 1 month exam was administered.

Table 15

*Regressions run with Dependent variable AKT6 and 5 independent variables with lowest Variance Inflation Factor (VIF)*

<table>
<thead>
<tr>
<th>Variable</th>
<th>AKT 6 Month</th>
<th></th>
<th></th>
<th></th>
<th>In-Training Exam</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
<td>VIF</td>
<td>B</td>
<td>Std. Error</td>
<td>β</td>
</tr>
<tr>
<td>Constant</td>
<td>-150.16</td>
<td>41.45</td>
<td>-11.39</td>
<td>7.13</td>
<td>-11.39</td>
<td>7.13</td>
<td></td>
</tr>
<tr>
<td>Interview Scores</td>
<td>.75</td>
<td>.85</td>
<td>.10</td>
<td>1.13</td>
<td>.13</td>
<td>.15</td>
<td>.10</td>
</tr>
<tr>
<td>Step 1</td>
<td>.45</td>
<td>.19*</td>
<td>.30</td>
<td>1.53</td>
<td>.09</td>
<td>.03**</td>
<td>.35</td>
</tr>
<tr>
<td>Step 2</td>
<td>.46</td>
<td>.18*</td>
<td>.34</td>
<td>1.65</td>
<td>.05</td>
<td>.03</td>
<td>.20</td>
</tr>
<tr>
<td>Gestalt</td>
<td>.12</td>
<td>2.57</td>
<td>.01</td>
<td>1.26</td>
<td>1.08</td>
<td>.45*</td>
<td>.29</td>
</tr>
<tr>
<td>Professionalism (1st Quarter)</td>
<td>.82</td>
<td>5.66</td>
<td>.02</td>
<td>1.14</td>
<td>.44</td>
<td>.97</td>
<td>.05</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.342</td>
<td></td>
<td></td>
<td>.342</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

**AKT 6 Month Examination.** In this analysis only the Step 1 and 2 scores were found to account for any variance in the regressions. Step 1 (b=.45, p=.025), and Step 2 (b=.46, p=.015) also had a low standard error indicating low multicollinearity and reinforcing the low VIF. The remaining variables of Interview scores, gestalt scores and 1st quarter professionalism did not account for any variance in this analysis.

**In-Training Examination.** Step 1 (b=.09, p=.01) and Gestalt scores (b=1.08, p=.02) accounted for variance for the prediction of performance of the In-training exam score in this analysis. Accounting for no variance were the remaining variables of interview scores, Step 2, Q1 professionalism.
Summary of Results from Question 1

Below is a summary of how each independent variable faired in the analyses used to investigate research question one as demonstrated in Tables 1-10.

**Step 1 Scores.** Step 1 scores were fairly consistently found to account for variance when regressed against the In-training exam scores with $p=.029$, $b=.08$ (Table 8) and $b=.10$ $p=.003$ (Table 11), and $b=.094$ $p=.006$ (Table 14). Step 1 scores were also found to be significant for the prediction of performance on the AKT 6 month exam with $p$ values of $.01$ $b=.47$ (Table 8), $b=.475$ $p=.014$ (Table 11), $b=.47$ $p=.02$ (Table 14) and a $p$ value of $.025$ $B=.446$ in Table 12. Step 1 scores were found to have no significance for the prediction of performance on the AKT 1 month examination. Overall, as an independent variable, Step 1 scores are valuable at predicting success on cognitive examinations within the first year of residency.

**Step 2 Scores.** Step 2 exam scores did not significantly predict performance on the AKT 1 month AKT 6 month or In-training exam scores. Step 2 was found to account for variance ($b=.459$ $p=.015$) only in the regression with the VIF independent variables (Table 15) where it was regressed with Step 1, Gestalt, Interview, first quarter professionalism, and interview scores. Step 2 scores, as an independent variable, offer little in the way of consistency when predicting performance on first year cognitive examinations.

**Gestalt Scores.** The Gestalt score significantly predicts ($b=5.39$ $p=.003$, Table 9) performance on the AKT 1 month when regressed as a single independent variable but has no significance in any other regression in this series. Furthermore, the Gestalt score was not found to have significance as an independent variable for the AKT 6 month in
any regression. As an independent variable regressed against the In-training examination the Gestalt score was repeatedly found to account for variance (b =1.11 p=. 016, Table 9, b=.99 p=.033, Table 10, b=1.03 p=.01, Table 11, b=1.12 p=.021, Table 12, b=1.02 p=.03, Table 13, b =1.13 p=.01, Table 14, b=1.08, p=.02, Table 15). As an independent variable the Gestalt score lacks the ability to account for variance with the cognitive examinations during the early months of residency (AKT 1 month and 6 month) but has consistent strength in the models for the In-training examination.

Sex. Sex was not found to be a significant predictor of performance for the AKT (1 or 6 month) or for the In-training exam. In all regressions run to analyze research question 1 whether or not you were a male or female held no significance in predicting outcomes.

Minority Status. Minority status was consistently found to be a predictor of success on the dependent variables of AKT 1 and 6 month exam, but not the In-training exam. Minority status was found to be significant at the .01 level against the AKT 1 month exam in all regressions in which it was included (b= -28.50 p=.006, Table 7, b= -29.02 p=.008, Table 8, b= -27.15 p=.007, Table 10, b= -28.10 p=.008, Table 11, b= -26.85 p=.009, Table 13, and b= -27.95 p=.01, Table 14). Minority status was found to have significance as a predictor for the AKT 6 month exam; demonstrating significance both at the 1% significance level (b= -27.05 p=.012, Table 7, b= -29.98 p=.007, Table 10, b =-27.28 p=.01, Table 13) and at a 5% significance level in the remaining tables (b= -19.02, Table 8, b= -19.91, Table 11, and b =-19.03, Table 14). Minority status, however, was not found to be a significant variable for predicting success on the In-training exam in any regression. Because the beta was always a negative in these regressions, all
significant variables indicated that minorities performed worse on the dependent variables of AKT 1 month and 6 month examinations; again this finding is severely limited by the small number of URM’s in the sample.

**Interview Scores.** Interview score was not significant in predicting success on the AKT 1 month examination in any regression. However, the interview scores were significant when regressed against the AKT 6 month exam with p values of .03 in Tables 4 b=5.33, and a b=6.11, p=.01 in Table 12. Interview scores were also significant for the 6 month AKT in Table 6 (b=6.17 p=.01). Interview scores were also unable to account for variance in any regression against the dependent variable of In-Training examination. Overall, interview scores as an independent variable were unable to consistently predict performance on most first year examinations.

**Question 2**

An ordinary least squares analysis model (linear regression) was used to investigate the research question, what linear combination of factors, controlling for Step 1 and 2 examination scores, predicts success (not entering remediation and/or probationary status) during the first year of residency? The independent variables of sex, minority status, Step 1 and Step 2 scores, interview scores, and the Gestalt survey score were investigated. Step 1 and Step 2 scores were controlled for in each model to determine if they were contributing too much significance to individual analysis due to their similarity to each other and to the dependent variables (both multiple choice standardized examinations). Regressions were conducted with all independent variables and then again with the interview score and Gestalt score separately in order to identify which, if any, variables were contributing to the significance.
PREDICTING SUCCESS IN RESIDENCY

The following three tables (16-18) list the results of the three OLS regressions run on the data using the dependent variable remediation status and the independent variable of interview status, missing, sex, and minority status. Variables were entered into the regressions in a stepwise manner to determine their significance. All three regressions demonstrated that interview scores were significant (b=.13 p=.001, Table 16, b=.11 p=.004, Table 17, b=.12 p=.002 Table 18) when predicting the probability of remediation/probation status in the first year of residency. However the other variables of sex, minority status and Step 1 and 2 demonstrated no significance.

Table 16

Regression run with Dependent variable Remediation status and independent variables Interview/Missing

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.12</td>
<td>.27</td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>0.13</td>
<td>.04**</td>
<td>1.07</td>
</tr>
<tr>
<td>Missing</td>
<td>0.93</td>
<td>.28**</td>
<td>1.00</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.164</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

Table 17

Regression run with Dependent variable Remediation status and independent variables Interview/Missing, Sex, and Minority

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.006</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>0.106</td>
<td>.04**</td>
<td>.90</td>
</tr>
<tr>
<td>Sex</td>
<td>0.179</td>
<td>.09</td>
<td>.22</td>
</tr>
<tr>
<td>Minority</td>
<td>-0.264</td>
<td>.16</td>
<td>-.19</td>
</tr>
<tr>
<td>Missing</td>
<td>0.785</td>
<td>.28**</td>
<td>.85</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.242</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
Table 18

*Regression run with Dependent variable Remediation status and independent variables Interview/Missing, Sex, Minority, Step 1 and Step 2*

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.93</td>
<td>.64</td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>0.12</td>
<td>.04**</td>
<td>.99</td>
</tr>
<tr>
<td>Sex</td>
<td>0.18</td>
<td>.09</td>
<td>.21</td>
</tr>
<tr>
<td>Minority</td>
<td>-0.31</td>
<td>.16</td>
<td>-.22</td>
</tr>
<tr>
<td>Step 1</td>
<td>-0.00</td>
<td>.00</td>
<td>-.15</td>
</tr>
<tr>
<td>Step 2</td>
<td>-0.00</td>
<td>.00</td>
<td>-.06</td>
</tr>
<tr>
<td>Missing</td>
<td>0.88</td>
<td>.28**</td>
<td>.95</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.274</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

Tables 19, 20 and 21 display the results of the regression analysis conducted with the independent variables of Gestalt, sex, minority, and Step 1 and 2. In these regressions, Gestalt was consistently found to be significant at the 1% significance interval (b=.17 p=.00 Table 19, b=.17 p=.00 Table 20, b=.18 p=.00 Table 21). Sex was also found to be significant in Tables 17 and 18 with b=.26 p=.002, b=.28 p=.002 respectively. As sex was coded 0 for male and 1 for female and the b was positive, these regressions demonstrated that being a female led to a higher likelihood of being placed on remediation or probation in the first year of residency. The remaining variables in Tables 19-21, consisting of Step 1 and 2, and minority status, were not found to be significant.
### Table 19

*Regression run with Dependent variable Remediation status and independent variable Gestalt*

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.32</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>0.17</td>
<td>.04**</td>
<td>.50</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.251</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05  **Significant with a p value of ≤.01

### Table 20

*Regression run with Dependent variable Remediation status and independent variable Gestalt, Sex, Minority status*

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.39</td>
<td>.24</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>0.17</td>
<td>.03**</td>
<td>.50</td>
</tr>
<tr>
<td>Sex</td>
<td>0.27</td>
<td>.08**</td>
<td>.32</td>
</tr>
<tr>
<td>Minority</td>
<td>-0.23</td>
<td>.14</td>
<td>-.17</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.380</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05  **Significant with a p value of ≤.01

### Table 21

*Regression run with Dependent variable Remediation status and independent variable Gestalt, Sex, Minority status, Step 1 and Step 2*

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.21</td>
<td>.61</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>0.18</td>
<td>.04**</td>
<td>.53</td>
</tr>
<tr>
<td>Sex</td>
<td>0.28</td>
<td>.08**</td>
<td>.33</td>
</tr>
<tr>
<td>Minority</td>
<td>-0.29</td>
<td>.14</td>
<td>-.21</td>
</tr>
<tr>
<td>Step 1</td>
<td>0</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td>Step 2</td>
<td>-0.00</td>
<td>.00</td>
<td>-.15</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.399</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05  **Significant with a p value of ≤.01
In the last three tables (Tables 22-24), the results of the regressions with all the independent variables regressed together are displayed. Once again, Gestalt demonstrated a high level of significance, p=.01 (Table 22, 23, and 24) when regressed against remediation status, reinforcing its high level of significance in predicting the probability of being placed on remediation. Sex, or being a female, was also significant b=.23 p=.008 (Table 23), and b=.24 p=.006 (Table 24), illustrating that females were more likely to be placed on probation in their first year of training. The interview variable was also found to be significant in predicting remediation status (b=.09 p=.012, Table 22, and b=.08 p=.02 Table 24 respectively), but the remaining variables of minority status, Step 1, and Step 2 scores demonstrated no significance.

Table 22

Regression run with Dependent variable Remediation status and independent variable Gestalt, Interview, and Missing

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.81</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>0.14</td>
<td>.04**</td>
<td>.43</td>
</tr>
<tr>
<td>Interview</td>
<td>0.09</td>
<td>.03*</td>
<td>.74</td>
</tr>
<tr>
<td>Missing</td>
<td>0.7</td>
<td>.26</td>
<td>.75</td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>.328</td>
<td></td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
Table 23

*Regression run with Dependent variable Remediation status and independent variable Gestalt, Interview, Missing, Sex, and Minority*

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.75</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>0.15</td>
<td>.03**</td>
<td>.45</td>
</tr>
<tr>
<td>Sex</td>
<td>0.23</td>
<td>.08**</td>
<td>.28</td>
</tr>
<tr>
<td>Minority</td>
<td>-0.20</td>
<td>.14</td>
<td>-.15</td>
</tr>
<tr>
<td>Interview</td>
<td>0.06</td>
<td>.03</td>
<td>.54</td>
</tr>
<tr>
<td>Missing</td>
<td>0.52</td>
<td>.25*</td>
<td>.57</td>
</tr>
</tbody>
</table>

*R² = .421
*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

Table 24

*Regression run with Dependent variable Remediation status and independent variable Gestalt, Interview, Missing, Sex, Minority, Step 1 and Step 2*

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>Std. Error</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.10</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>0.16</td>
<td>.04**</td>
<td>.47</td>
</tr>
<tr>
<td>Sex</td>
<td>0.24</td>
<td>.08**</td>
<td>.29</td>
</tr>
<tr>
<td>Minority</td>
<td>-0.27</td>
<td>.14</td>
<td>-.20</td>
</tr>
<tr>
<td>Step 1</td>
<td>-0.00</td>
<td>.00</td>
<td>-.03</td>
</tr>
<tr>
<td>Step 2</td>
<td>-0.00</td>
<td>.00</td>
<td>-.18</td>
</tr>
<tr>
<td>Interview</td>
<td>0.08</td>
<td>.03*</td>
<td>.65</td>
</tr>
<tr>
<td>Missing</td>
<td>0.62</td>
<td>.26</td>
<td>.67</td>
</tr>
</tbody>
</table>

*R² = .453
*Significant with a p value of ≤.05 **Significant with a p value of ≤.01

**Summary of Question 2 Results**

Below is a summary of how the independent variables faired in their ability to predict whether an individual would be placed onto remediation/probation in their first year of residency.
Sex. As an independent variable, sex was found to be a significant in predicting an individual’s chances of ending up on probation and/or remediation when all variables were involved in the regressions (b=.24 p=.006 Table 24) and in the regressions run without the Interview variable (b=.27 p=.002, Table 20 and b=.23 p=.008, Table 23). The results of these analyses indicate that females are more likely to be placed into probationary status in their first year of residency training.

Minority Status. Minority status, that is being an underrepresented minority, was not found to be a significant predictor of likelihood to end up in remediation/probationary status.

Step 1 Scores. Step 1 scores were found to account for no variance in the analysis; therefore it seems as though a resident’s score on their Step 1 examination does not predict whether an individual ends up in probationary/remediation status.

Step 2 Scores. A resident’s score on their Step 2 examination did not predict whether an individual ended up in probationary/remediation status. Furthermore, Step 1 and 2 scores did not appear to influence the significance of other variables when added to the regressions.

Gestalt. The variable Gestalt had a strong predictive association with the remediation status variable, p value=.00 (Tables, 19, 20,21,22,23 and 24). Irrespective of variables in a regression and whether Step scores were regressed in the same equation, all regressions demonstrated that a low score on the Gestalt survey was highly predictive of an individual’s likelihood of being put into remediation and/or probationary status.

Interview Scores. Interview scores are continuous, coded as a discrete number of 1 through 10 and when they were regressed against the dependent variable of remediation
status they were found to account for variance: $b=.13 \ p=.001$ (Table 16) $b=.11 \ p=.004$ (Table 17), $b=.12 \ p=.002$ (Table 18), $b=.09 \ p=.012$ (Table 22), $b=.08, p=.02$ (Table 24).

This means that residents who received higher scores during the interview process were more likely to remain in good standing during the first year of residency. The interview variable did not lose its significance as more variables were added to the regressions, suggesting that it had little collinearity with other independent variables.

**Summary**

Several variables were found to be significant predictors of scores on tests of knowledge administered to residents during their first year of training. On one hand, variables such as Step 1 scores, Gestalt scores, and minority status were found to have repeated significance in the regressions independent of variables being assessed. On the other hand, the analysis of research question two demonstrated that the Gestalt score, sex, and interview scores of an individual were significantly strong in predicting whether or not an individual would be placed on remediation or probation in their first year of training. The next chapter will discuss the meaning of these results.
Summary and Discussion

In this chapter, I will review and discuss the results of my study. First I will summarize the research, along with the findings and study limitations, and then I will discuss the results in the context of what has already been demonstrated in the literature and what is new and novel.

This study was a post-hoc correlational analysis of three years of residents who interviewed and were accepted in an anesthesiology residency program. The study examined the first year of residency for metrics of success (as defined by not entering remediation/probation or scoring well on standardized assessment tests) for three cohorts from 2008-10. Regression analysis (OLS) was employed to complete the statistical evaluation for the following two research questions: 1) What are the independent variables that are associated with future AKT and In-Service Examination test performance of first year residents? and 2) What linear combination of factors, controlling for Step 1 and 2 examination scores, predicts success (not entering remediation and/or probationary status) during the first year of residency?

Summary of Analysis

When investigating the first research question (predicting performance on standardized exams), the analysis demonstrated that the independent variables of Step 1 and Gestalt scores were strongly correlated with performance on the three standardized examinations AKT 1, AKT6 and the In-training Exam (the Gestalt score was significant at predicting performance on the In-training examination only) in the first year of residency. Interview scores were also found to have a weak association with performance
on the AKT 6 month examination. Additionally, minority status, or being a minority, was found to be a predictor of performance on those same exams.

However, when investigating the second research question, whether or not any independent variables were significant in the prediction of remediation and/or probationary status in the first year of residency, several variables were found to be significant: Gestalt scores, interview scores, and sex were all found to be significant predictors of remediation status whereas minority status, and Step 1 and 2 scores were found to account for no variance.

**Limitations of Study**

A significant limitation of the study is the generalizability of the findings (or its external validity). Because the data that were analyzed were from one institution and a single anesthesiology program, it is difficult to extrapolate the results given that each program throughout the United States has the freedom to design and deliver curriculum and training as it sees fit within the confines of accreditation standards. Programs also have discretion on how they move individuals along the remediation/probation timeline and could choose to take a less/more aggressive approach than the study institution. A second major limitation of the study is the small n for the under-represented minority population. With only 6 URMs in the study cohort, with only 1 in each the first 2 years of the study, the n is too small to make generalizations about the findings. So even though minority status demonstrated some significance in the analysis the results will not be addressed in the discussions session.

However, in looking at the population alone, this study has good external reliability, for its population of residents will look similar to residents in other
anesthesiology programs throughout the nation. The study cohort will look similar in characteristics such as degree completion, minimal standards for entry, and completion of required examinations such as Step 1 and 2. There will, of course, be some differences in demographic features with the breakdown of male to female ratios or the percentage of underrepresented minorities. However, with the use of a blind match process for admittance to accredited residencies in the United States, individual programs do not get to choose their students independent of their participation the national matching/placement program.

Another limitation of the study was multicollinearity and repeated measure bias with several of the independent variables. Some of the independent variables are measuring overlapping skills and, therefore, several of them are highly correlated (Gestalt scores and crisis management scores for example). A correlation matrix was created and discussed in detail in Chapter 4, where I also explained how correlation was kept to a minimum by the inclusion of a VIF analysis. Repeated measure bias is also likely to exist in this study as there is a limited pool of evaluators for multiple evaluation independent variables. However, with the correlation matrix and the stepwise entry of the variables into the regressions these overlaps are evaluated and accounted for.

Another consideration is that it is very difficult to account for is the potential of bias. While I can account for multicollinearity of our variables, it is impossible to account for bias that may exist in some evaluation scores. Scores that include personal judgment be it from the interview process, the Gestalt survey, or from the new innovations on-line evaluations of performance could contain some inherent bias; such bias could be related to gender, personality type, or even bias that is carried over from a previous working
relationship. For example, a faculty member who has a negative working experience with a resident in an early month of their residency training and gives them a negative evaluation may carry some of that negative feeling over into future performance evaluations.

A further limitation that applies to question 2, the investigation of which variables are related to remediation/probationary status, lies in the fact that while all ACGME accredited programs must follow the same standards, leeway is given in how individual programs test the proficiency of their residents or decide their good standing/remediation or probationary status. While the study institution is fairly aggressive in its handling of underperforming residents, other programs may move on a different timeline with a different sequence of educational interventions.

Taking into account these limitations, it is still within reason to assume that individual institutions can use the independent variables they have on hand with the same dependent variables (AKT and In-training exams are mandatory across all US programs) to conduct a similar analysis to find variables of significance for their program.

**Internal Validity.** Even though the study employed a single group design, it still maintains internal validity. Internal validity was assured through the lack of selection bias (all residents in their respective cohorts were included) and no repeated testing or maturation issues needed to be addressed. Furthermore, to ensure minimal confounding variables were present, all independent variables were entered into the design in a step-by-step pattern in order to determine the significance of each independent variable.
Limitations Summary. Despite these limitations, I believe we can still find meaningful trends in predicting success in the first year of residency for all anesthesiology residency programs throughout the United States. Using the independent variables that all programs have on hand at entry, such as Step 1 and 2 Scores, interview scores, sex, minority status, along with a the creation of a 3 question Gestalt survey (as implemented in this study), there is the potential to identify variables that predict performance in similar programs.

Discussion of Findings for Question 1

Descriptive Variables

Sex. In this study the sex variable was found to have no significance in predicting test outcome in the first year of residency. Prior research has shown that there are some performance differences between the sexes (COGME, 1995; Hojat, Gonnella, & Xu, 1995; Hojat, Gonnella, Veloski, & Moses, 1990), especially in the non-cognitive domains, with males performing better than females, but that gap narrows as learners progress in their medical school training (Hojat et al., 1994). So, while sex differences may exist on test performance, it appears that one’s sex cannot predict by itself test outcome for this residency program; and the study program can assure itself that the sex of their resident population has little impact on test scores.

Test Performance. The reviewed literature demonstrated relationships between independent cognitive variables of success (test performance) and prediction of future success on cognitive examinations. For example, the literature confirmed that the MCAT was a strong predictor for performance on Step 1 and/or 2 examinations (Fleming, 2002;
Huff et al., 1999; Julian 2005) and that performance on those same Step exams were strong predictors of performance on Board examinations (Elam et al, 1994; Myles & Galvez-Myles, 2003). Previous research also indicated that the In-training examination was predictive of board performance (Bailey et al., 2000; Klein et al., 2004; McClintock & Gravlee, 2010). Based on this research, it seems as though previous performance on multiple choice type examinations should be significant in predicting future test performance; this generalization was confirmed in present research.

The results of this research demonstrated that the Step 1 exam was significant in predicting performance on the AKT 1 month, 6 month and the In-training exam. Step 2 scores were less significant in predicting performance on first year examinations. Step 2’s lack of predictive power could be explained by the nature of the exam itself; Step 1 is designed to test an individual’s ability to apply discrete concepts to the practice of medicine, whereas Step 2 is designed to assess critical thinking skills to the practice of medicine. Therefore it makes sense that Step 1 is a good predictor of the AKT exams (both tests of discrete knowledge) and that Step 2 is not a good predictor of those same tests because of the differences of test type (application of knowledge versus discrete knowledge). Step 1 as a stronger predictor of the ITE exam is somewhat puzzling, because aside from the tests being both nationally normed multiple choice type examinations, their nature is different (test of discrete knowledge in the AKT examinations versus test of applied knowledge for In-training exam). It would be logical if Step 2 were a stronger predictor of In-training examination scores but that was not borne out in this research. In my opinion, the Step 1 and Step 2 exam (with scores that are strongly correlated) are confounding variables whereby Step 1 pulls all of the variance in
PREDICTING SUCCESS IN RESIDENCY

the analysis. It would be prudent to rerun the analysis with Step 1 and Step 2 variables entered independently into the regression(s) to see if the results vary.

Non-Cognitive Predictors of Success

The literature review examined many non-cognitive variables and their relationship to academic success. We learned that residents that have a ‘broad’ approach to learning (one that involves the integration of many skills) are more successful than those with a ‘narrow’ (rote memorization) approach (Lindblom-Yanne, Lonka, & Leskinen, 1996). We also discovered that the integration of personality traits and interpersonal skills in prediction models increased the accuracy significantly (Schwartz et all, 1995; Shen & Comrey, 1997). While approach to learning and individual personality traits were not analyzed in this research, much of the essence of those variables was. For instance, the Gestalt score integrates personality traits and interpersonal skills into the overall assessment score and the New Innovations metric of communication skills is a measurement of interpersonal communication as applied to the clinical setting.

Considering the impact that personality and one’s approach to learning plays into academic success it is not surprising to discover that the Gestalt score variable was not significant with predicting success on the earlier examinations (AKT 1 and 6 month examinations) whilst repeatedly being found significant against the In-training examination scores. The Gestalt score is designed to be a quick measurement of an individual’s clinical skills, medical knowledge, and future leadership potential so it is not unexpected to discover that it was not significant at predicting performance on examinations that simply measured discrete knowledge. However, its significance at predicting success of the In-training examination is interesting. While the In-training
exam measures the application of knowledge, much like the future board examinations, it is still a multiple choice nationally normed standards examination. Perhaps it is because the In-training exam’s modality is one of application of knowledge (clinical skills ability plus medical knowledge) that gives significance to the Gestalt scores.

**Interview Scores.** Interview scores were a very interesting variable because, as an independent variable, they capture so much data. Interview scores come from a 20 minute snapshot of a meeting that took place about 1.5 years prior to starting the CA-1 year. They are composite scores that incorporate the results of medical studies performance, test performance, letters of recommendation, and an in-person interview that captures soft competencies (such as communication skills, non-verbal cues, critical thinking skills). Research has demonstrated that models that integrate cognitive and non-cognitive variables tend to be the most accurate at predicting future performance (Schwartz et al., 1995; Shen & Comrey, 1997), meaning that interview scores, which are a measure of both cognitive and non-cognitive attributes, should hold some predictive power. On their own, interview scores have been found to be reasonable predictors of performance in medical school as long as the interviews being employed were structured interviews (Donnon, Paolucci, & Violato, 2009; Reiter, Eva, Rosenfeld, & Norman, 2007; Peskun, Detsky & Shandling, 2007), but tend to lose their strength (or any predictive power) in the analysis when standardized test scores were included (Elam & Johnson, 1992; Meredith, Dunlap & Baker, 1982; Reiter, et al., 2007).

On the one hand, in this study, interview scores were found significant in predicting performance on the AKT 6 month examination in all analyses. On the other hand, the interview score was not consistently able to predict performance on the AKT 1
PREDICTING SUCCESS IN RESIDENCY

month examination or the In-training examination. It is difficult to explain the anomaly of
the interview score variable having predictive significance for only the AKT 6 month
dependent variable, considering that the AKT 1 month and 6 month examinations are
similar in content and have the same multiple choice format. According to the literature,
interview scores are fairly poor predictors of pure cognitive examinations (Elam &
Johnson, 1992; Meredith et al., 1982; Reiter et al., 2007), but gain strength against
dependent variables that measure some aspect of clinical performance. Interview scores
lacking strength against cognitive dependent variables explains the lack of power against
the AKT 1 month exam and even the In-training examination, but not why it is
consistently predictive of performance on the AKT 6 month examination. Further
analysis should be conducted to investigate whether there are some confounding or
extraneous variable contributing to this result.

Discussion of Findings for Question 2

Summary

Several interesting results were found for research question 2, whether any
independent variables were significant predictors of remediation or probation in the first
year of residency. First, unlike with question 1, sex was found to be a significant variable,
whereas minority status was not. In this set of analyses Step 1 and 2 scores had no
significance in predicting remediation status. In contrast, interview scores were found
frequently to predict remediation status, whereby residents with higher interview scores
were much less likely to be placed on remediation and/or probation. Another important
variable with predictive power was the Gestalt score; it was significant in predicting remediation status in every regression with a consistent p value of less than .01.

Variables of Significance

**Sex.** Sex was found to be a significant predictor of whether or not an individual would be placed into a remediation or probationary status in the first year of residency. This means that being a female was a significant predictor of being placed into remediation status. This is a surprising finding, but it could be confounded because of the relatively small number of women (23 females to 44 males) in the study cohort, or it could indicate concrete performance differences on standardized examinations or in the clinical setting. With the sample size of females being 48% smaller than the males, there is a possibility that having a larger n with an equal distribution of males and females could yield a different result. However, given that nationally there are fewer women than men in the field of anesthesiology, it is more likely that there is a significant difference in how females are performing, or being assessed, at the study institution.

Gender differences have been reported between men and women on both cognitive and non-cognitive domains with cognitive domains (test taking and knowledge acquisition) (Hojat et all, 1994, 1995), narrowing the more the individual advances in training. However non-cognitive domains, such as communication skills or professionalism (evaluated in discrete areas such as hours worked per week, professional satisfaction, personality types), tend to remain disparate (Gross, 1992; Nadelson, 1989; Tesch, Wood, Helwig, & Nottinger, 1995).

Analysis of research question 1 demonstrated no difference between males and females in their examination performance and the correlation matrix demonstrated no
overlap between sex and other independent variables but females were found to be placed into remediation/probationary status more frequently than their male counterparts. Therefore some other sex-related issue must be driving the results associated with “success” (as defined as not being placed on remediation/probation, not as success on examinations as in question 1). The gender differences in who is placed into remediation/probation could simply be chalked up to evaluation bias between genders but women being scored worse than men is contrary to the literature that illustrates that females are often scored more favorably than males on OSCE (global) rating evaluations (Carson, Peets, Grant, & McLaughlin, 2010). The faculty population, who consistently rated the residents on the independent variables, is just over half male (55%) and there is a strong possibility that males are rating females harder than their male counterparts. This gender bias, where an individual is perceived treated and evaluated differently based upon their sex has been demonstrated in the literature but is rarely addressed in clinical practice (MacNell, 2014). An example of gender bias that exists in evaluations are when the same action is perceived as assertiveness for a male, is considered abrasiveness for a female. In these instances, high performing women often receive more negative evaluations than their male counterparts (Snyder, 2014), a factor that could be in play in the performance metrics being analyzed.

In my opinion, however, it is likely that women who find themselves in a remediation/probationary status do so because of professionalism issues, and not, as in demonstrated in this research, academic reasons. It is those professionalism red flags that move an individual into remediation/probationary status quickly. In Chapter 1 the criteria to be placed on remediation/probation was outlined and instances of professional
misconduct are the fastest way to reach that status. While the study institution allows people to be put onto an informal counseling status when they get unsatisfactory test scores (at which point they are expected to create and individualized learning plan (ILP) and likely retake the same test to demonstrate competency), instances of professional misconduct lead to immediate placement into a remediation/probationary status. Therefore, it is very likely that people who are simply poor test takers arrive at remediation/probationary status later in training (after their first year of training and outside of the scope of this study).

**Interview Scores.** The better (higher) the interview score individuals had the less likely they were to be placed onto remediation and/or probation in their first year of residency. This trend has been borne out in the literature with significant correlation between interview performance and clinical performance (performance in the clinical setting, excluding academic test performance) (Brothers & Wetherholt, 2007; Donnon, Paolucci, & Violato, 2009; Gough, Hall, Harris, 1964; Meredith et al., 1982; Reiter et al., 2007). However research suggests that the predictive power of the interview score variable is reduced if clinical performance is measured simply as a standardized test (Elam & Johnson, 1992; Meredith et al, 1982; Reiter et al, 2007; Smith, 1991). In this research project, interview scores are a compilation of an individual’s academic history, test scores, letters of reference, personal statements, and communication skills, and personality derived from a 20 minute encounter with the interviewer, so it is not surprising they have significance in predicting remediation status. The robustness of the interview score, touching so many cognitive and non-cognitive domains, makes it a natural predictor of success during the first year. During the first year of residency
training, individuals will be navigating through diverse systems and people while continually demonstrating their academic and clinical skills to those who will be evaluating them. Therefore the competency areas evaluated during the first year of training and during the interview process overlap significantly.

One must not also forget that the interviewer and first year faculty evaluator could overlap in many instances; so a faculty member who was left with a positive interview experience with an individual could carry some of that first impression into early work encounters, conversely the opposite of a negative first impression could carry over.

**Gestalt Scores.** Global assessments, or holistic ratings, are most frequently used in objective structured clinical exams (OSCE) and have been demonstrated to have as much validity as more traditional analytical (selection of specific individual scoring components from a checklist) assessments (Goulden, 1994), but due to inter-rater differences are slightly less reliable than their checklist counterparts (Ilgen, Ma, Hatala & Cook, 2015; Slater & Boulet, 2001). Global rating scales incorporate the subjective “I know it when I see it” along with any casual rater judgment of proficiency whereas traditional checklist scales, in the attempt to be more objective and cover a wider array of topics, can often lose the big picture of competency in favor of specificity as the raters separate individual skill sets and proficiencies (Cunnington, Neville, & Norman, 1996; Hodges, Regehr, McNaughton, Tiberius & Hanson, 1999; van der Vleuten, Norman, & De Graaff, 1991).

Gestalt scores used for this analysis are a global performance assessment created by averaging the results of a simple 3 question survey administered to faculty one month into residency training. Gestalt scores are highly significant at predicting whether or not
an individual will end up on remediation status at the study institution. In fact, the Gestalt score accounted for almost all the variance in every regression in which it was included; the lower the score on the Gestalt survey, the more likely an individual was to be placed into a remediation/probationary status in their first year of residency. The significance of the variable is likely because the Gestalt score is a compilation of an observed individual’s performance on both the clinical skills domain (question 1 of the survey), along with a quick assessment of their medical knowledge (question 2 of the survey). The third question on the Gestalt survey is designed to assess the individual’s probability of becoming a chief resident and likely picks up on many of those soft competencies of communication skills, ability to work in a team, and the simple likeability factor. The Gestalt score variable is significantly correlated with quarters 1 to 3 New Innovations evaluations of overall clinical competency, crisis management, and professionalism; although because of timing issues (when the data were compiled and available for use), those variables were not used in the analysis to predict probation/remediation status.

There is no doubt that there is some overlap in those faculty who complete the Gestalt survey, faculty who are involved in the interview process, and even those faculty who complete evaluations on the residents later which may explain in part the significance of this independent variable. With a ‘gut’ feeling of how someone will perform throughout the remainder of the residency, we ask the evaluator to project future competency based upon a very discreet experience (less than one month of working together) and there is a great likelihood of transference between that first experience with a resident and their future evaluations from that faculty member. We can see this overlap when we look at the relationship between the Gestalt score and the 1st to 3rd quarters new
innovations scores in the correlation matrix in Appendix A, as the variables are highly correlated. Some studies show that the problem with global ratings is that faculty might tend to overemphasize technical competencies and undervalue humanistic skills (Domingues, Amaral, & Zeferino, 2009); however, because the Gestalt score does not account for variance at predicting exam performance, but instead is significant only in predicting remediation/probationary status, it does not seem to be occurring in this instance. Instead, the low level of correlation with Step 1 and Step 2 exams and high correlation with New Innovation performance on crisis management, communication, and overall clinical performance indicates the Gestalt score accurately captures the assessment of non-cognitive skills.

**Conclusion**

With the variables that we have available to us when residents are matched with residency programs and those that we can collect with their first few months of training, we are able to predict with a strong degree of accuracy performance on cognitive examinations and whether or not a resident will be placed into promotion/remediation during their first year of training at our training institution. Variables such as interview scores, Step 1 scores, and Gestalt interview scores are strong predictors of performance and/or success in residency. Our program should use these data we have available to implement early identification of academic difficulties and to design intervention programs for at-risk residents and use what we have learned to inform like programs of our success and failures.
Implications of Findings

The results of this study are important to the field of anesthesiology education and I believe to the greater field of graduate medical education. From the analysis, it is evident that residents come into training programs with information that has the potential for significance in predicting future test performance. Furthermore, a simple 3 question Gestalt survey, such as the one outlined previously and administered early in residency, allows the possibility of early identification of those individuals likely to end up in substantial academic distress.

At its most simple level, we want to assume that there is linearity in an individual’s performance over time, where each individual travels a fairly straight line of performance from the start of the training program to the finish. If trainees followed a linear progression programs would simply have to recruit specific residents to ensure success within specific institutions. This linearity holds true when we look at academic success and see that past test performance can often predict future performance (Step 1 is a strong predictor of the AKT 6 month and In-training exam) but is confounded when we add in demographic variables such as minority status and sex. To add to this linearity problem are the variables of professionalism, communication style, and perceived competency (captured by the Gestalt survey tool) that are significant predictors of success when defined in non-cognitive assessment terms (remediation/probation). Creating a prediction tool that accurately predicts success in both ‘success’ domains means that one has to accept that performance is not linear and certain variables are not modifiable (sex, minority status).
Programs, independent of specialty, should analyze the data they have on hand (from the residency application process and from early required examinations) to look for variables that are significant to their trainee population to aid in the timely identification of residents in academic distress. Early identification of these residents can ensure that programs can employ interventions early enough to positively impact the resident’s future and lead to the reduced likelihood of his/her leaving the program (forcibly or voluntarily) and an increased likelihood of on-time graduation. Employing effective, and proven, assessment tools and/or intervention tools will have a positive impact on the practice of anesthesiology as a whole; leading to timely graduation of residents and an efficient training timetable that matches with fellowship and employment opportunities. In theory, early identification and interventions would decrease the attrition rates for programs leading to a more satisfied trainee population and less disruption to the training institution.

It will be imperative to continue to do this type of research within academic training programs to ensure that the programs are able to easily identify those in academic distress early on in training. Once identified, programs will then have the opportunity to employ early interventions and increase the potential for resident retention and graduation of a competent clinician. However one cannot merely trust that analysis on the cohort of residents from 2008-11 will remain valid going forward; tests change along with required residency curricula. Repeated analysis that includes incoming descriptive information analyzed along with any early measurement tool will be the best way to ensure that programs can successfully graduate every matched resident. I believe it would be prudent for every residency, independent of discipline, to employ statistical
analysis to look for ways to identify residents with potential to enter academic distress. This analysis used variables that are easily accessible either through the match process or through performance records (that are required by the residency board committees) and would require very little time to collect and collate. If residencies do not keep records of their interview process, I would strongly suggest they start because those data have the potential to be a valuable predictor of success during the first year of residency.

**Future Direction for Research**

Exploring the relationship of the sex of an individual learner and their performance during first year of residency is an area with much opportunity for further research. Sex is a fixed variable and it cannot be modified, yet it was found to be a powerful variable in the present study for predicting whether or not an individual would end up on remediation or probationary status. I believe that further investigation into the performance files of the individuals who entered remediation status is warranted. Coding the reason an individual was placed into remediation status (academic/professional/other) would be useful into understand why females were more likely to be placed into that status.

There is also a substantial opportunity to do further research on the minority status variable. A multi-institutional study might provide a clearer picture whether minority status has any significant and meaningful relationship with performance on cognitive examinations. Because the URM sample size at my institution is so small a larger n could remove interactions between demographic variables and provide a more detailed picture of whether or not being a minority is truly predictive of success during anesthesiology residency.
PREDICTING SUCCESS IN RESIDENCY

Milestones, have recently replaced the ACGME mandated core competency evaluations, as a required reporting measure. Milestones are 25 discreet evaluation points, specialty specific, upon which an individual is measured on throughout their residency designed to evaluate along a continuum rather than a generalized Likert scale of performance. With milestones trainees are evaluated in relation to their level of training with the caveat that it is expected for them to ebb and flow in performance on different domains throughout their three year residency. For example, it is within the realm of possibility that a first year trainee has second year communication skills while scoring at the intern level on managing the airway. These dynamic and fluctuating evaluation metrics now being collected in combination with the existing independent variables analyzed in this study, offer opportunities to investigate many other possible prediction models.

There are also some variables that were not analyzed but could be considered predictors, such as rank order (rank that an individual was placed on the program list for match with the program) or applicant age. Fryer, Corcoran, George, Wand and DaRosa (2012) investigated rank and success in a surgical residency and found that raw rank scores (scores that were not yet adjusted due to committee input) were good predictors of future resident performance. Urology also found similar results with the relationship between higher rank position and future success in residency (Grewal, Yeung and Brandes, 2013). At the study institution, rank order is determined as a weighted rank from a combination of interview score, Step 1 performance, and committee input. Given that rank order is a global assessment variable, much like the Gestalt score, there is a great potential for it to have some power of prediction.
As a final investigation, it would be prudent to examine the interview variable more closely. Are some interviewers better at predicting success in residency than others? Are male interviewers more likely to be more generous score givers than females (Lumb, Homer, & Miller, 2010), or are they more likely to rate females more negatively (MacNell, 2014)? Is there consistency across interviewers or is there a negative interview effect where an interviewer has bias against some interviewees (Elam, Andrykowski, 1991; Lumb, et al., 2010)? Breaking down specific subsections of the interview could also lead to improved predictive validity. For example Brothers and Wetherholt (2007) study demonstrated that the interview subsections of personal characteristics and letters of reference were predictive of future performance on the ACGME core competencies. My institution already collects these detailed components of the interview aggregate score, so it would be prudent to include them in future research. Furthermore, it would be prudent to conduct further investigation into why the interview variable is significant at predicting performance on the AKT 6 month examination but not for the AKT 1 month or the In-Training exam; is this an anomaly or are there confounding variables involved in the analysis?

Altogether, the evidence gathered from this research project is fairly consistent with previous analyses. The research demonstrated that individuals have a fairly linear path of performance when it comes to predicting success on cognitive variables but that predicting success as defined by the probability of being placed into remediation (analyzing both their academic and clinical performance) is a much less linear relationship. Holistic global assessments (such as interview or gestalt scores) had much more power of prediction when it came to predicting overall program success.
Demographic variables that we cannot change were found to have several associations with performance during the first year of residency. Given these findings, it is important that future research explore these variables in more depth. Future researchers should continue to investigate the variables that predict success at their institution but at the same time not forget the institutional impact on some of the performance variables. For example, using the AKT 6 month exam as a dependent variable is only useful if you accurately prepare your trainees for that exam. Failing to look at the trees in the forest (not teaching the materials at the appropriate time or design the curriculum effectively) could negatively impact the results and the strength of an analysis such as that used in the present study.
### Appendix A

#### Correlation Matrix of Independent variables

<table>
<thead>
<tr>
<th></th>
<th>Step1</th>
<th>Step2</th>
<th>Minority</th>
<th>Sex</th>
<th>Interview</th>
<th>Missing</th>
<th>Gestalt</th>
<th>Q3 Crisis mgmt</th>
<th>Q3 Prof.</th>
<th>Q3 Clinical Co.</th>
<th>Q2 Prof.</th>
<th>Q2 Clinical Co.</th>
<th>Q1 Crisis mgmt</th>
<th>Q1 Prof.</th>
<th>Q1 Clinical Comp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step1</td>
<td>1</td>
<td>1</td>
<td>.52</td>
<td>.00</td>
<td>.06</td>
<td>.057</td>
<td>.125</td>
<td>-.08</td>
<td>.09</td>
<td>-.11</td>
<td>.12</td>
<td>-.16</td>
<td>.09</td>
<td>.10</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.72</td>
<td></td>
<td>.128</td>
<td></td>
<td>.08</td>
<td>.02</td>
<td></td>
<td>.11</td>
<td>.3</td>
<td>.16</td>
<td>.09</td>
<td>.12</td>
<td>.123</td>
</tr>
<tr>
<td></td>
<td>.00</td>
<td>.303</td>
<td>.96</td>
<td>.647</td>
<td>.312</td>
<td>.51</td>
<td>.45</td>
<td>.37</td>
<td>.55</td>
<td>.33</td>
<td>.17</td>
<td>.46</td>
<td>.40</td>
<td>.31</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step2</td>
<td>.52</td>
<td>.08</td>
<td>.12</td>
<td>.31</td>
<td>.08</td>
<td>.194</td>
<td>.21</td>
<td>.09</td>
<td>.21</td>
<td>.14</td>
<td>.33</td>
<td>.27</td>
<td>.33</td>
<td>.14</td>
<td>.19</td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>.30</td>
<td>.1</td>
<td>.01</td>
<td>.00</td>
<td>.115</td>
<td>.468</td>
<td>.08</td>
<td>.43</td>
<td>.24</td>
<td>.00</td>
<td>.02</td>
<td>.00</td>
<td>.02</td>
<td>.297</td>
</tr>
<tr>
<td></td>
<td>.00</td>
<td>.011</td>
<td>.49</td>
<td>.1</td>
<td>.115</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minority</td>
<td>.12</td>
<td>.00</td>
<td>.12</td>
<td>.31</td>
<td>.08</td>
<td>.122</td>
<td>.07</td>
<td>.09</td>
<td>.13</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
<td>.27</td>
<td>.14</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>.8</td>
<td>.30</td>
<td>.95</td>
<td>.327</td>
<td>.576</td>
<td>.22</td>
<td>.43</td>
<td>.29</td>
<td>.77</td>
<td>.00</td>
<td>.00</td>
<td>.02</td>
<td>.25</td>
<td>.42</td>
<td>.852</td>
</tr>
<tr>
<td></td>
<td>.01</td>
<td>.08</td>
<td>.007</td>
<td>.1</td>
<td>.031</td>
<td>.037</td>
<td>.08</td>
<td>.00</td>
<td>.15</td>
<td>.11</td>
<td>.13</td>
<td>.06</td>
<td>.13</td>
<td>.06</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>.00</td>
<td>.49</td>
<td>.49</td>
<td>.95</td>
<td>.958</td>
<td>.804</td>
<td>.764</td>
<td>.48</td>
<td>.95</td>
<td>.55</td>
<td>.34</td>
<td>.57</td>
<td>.29</td>
<td>.61</td>
<td>.835</td>
</tr>
<tr>
<td></td>
<td>.9</td>
<td></td>
<td>.1</td>
<td>.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>.05</td>
<td>.19</td>
<td>.19</td>
<td>.03</td>
<td>.1</td>
<td>.926</td>
<td>.23</td>
<td>.15</td>
<td>.14</td>
<td>.15</td>
<td>.08</td>
<td>.14</td>
<td>.04</td>
<td>.03</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>.64</td>
<td>.11</td>
<td>.327</td>
<td>.80</td>
<td>.000</td>
<td>.05</td>
<td>.22</td>
<td>.48</td>
<td>.22</td>
<td>.50</td>
<td>.22</td>
<td>.74</td>
<td>.79</td>
<td>.548</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>.12</td>
<td>.07</td>
<td>.03</td>
<td>.07</td>
<td>.926</td>
<td>.000</td>
<td>.26</td>
<td>.75</td>
<td>.74</td>
<td>.99</td>
<td>.55</td>
<td>.85</td>
<td>.37</td>
<td>.49</td>
<td>.687</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.31</td>
<td>.46</td>
<td>.576</td>
<td>.76</td>
<td>.000</td>
<td>.06</td>
<td>.75</td>
<td>.74</td>
<td>.99</td>
<td>.55</td>
<td>.85</td>
<td>.37</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td>Gestalt</td>
<td>.08</td>
<td>.21</td>
<td>.21</td>
<td>.150</td>
<td>.237</td>
<td>.137</td>
<td>.51</td>
<td>.54</td>
<td>.39</td>
<td>.60</td>
<td>.51</td>
<td>.67</td>
<td>.43</td>
<td>.29</td>
<td>.355</td>
</tr>
<tr>
<td></td>
<td>.2</td>
<td>.51</td>
<td>.08</td>
<td>.226</td>
<td>.48</td>
<td>.053</td>
<td>.269</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.01</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3 Crisis mgmt</td>
<td>.09</td>
<td>.097</td>
<td>.00</td>
<td>.152</td>
<td>.038</td>
<td>.51</td>
<td>1</td>
<td>.87</td>
<td>.39</td>
<td>.35</td>
<td>.43</td>
<td>.24</td>
<td>.292</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>.45</td>
<td>.95</td>
<td>.221</td>
<td>.759</td>
<td>.00</td>
<td>0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Q3 Prof.</td>
<td>.11</td>
<td>.21</td>
<td>.00</td>
<td>.149</td>
<td>.041</td>
<td>.54</td>
<td>1</td>
<td>.79</td>
<td>.46</td>
<td>.47</td>
<td>.47</td>
<td>.24</td>
<td>.17</td>
<td>.200</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.1</td>
<td>.37</td>
<td>.97</td>
<td>.228</td>
<td>.742</td>
<td>.00</td>
<td>0</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.05</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td>Q3 Clinical Co.</td>
<td>.07</td>
<td>.14</td>
<td>.07</td>
<td>.087</td>
<td>.017</td>
<td>.39</td>
<td>1</td>
<td>.32</td>
<td>.24</td>
<td>.42</td>
<td>.32</td>
<td>.23</td>
<td>.395</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

107
### Predicting Success in Residency

<table>
<thead>
<tr>
<th>Q2 Crisis mgm t.</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.55</td>
<td>.24</td>
<td>.772</td>
<td>.55</td>
<td>.483</td>
<td>.894</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.04</td>
<td>.00</td>
</tr>
<tr>
<td>Q2 Prof.</td>
<td>.33</td>
<td>.00</td>
<td>.006</td>
<td>.22</td>
<td>.224</td>
<td>.993</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Q2 Clinical Com pt.</td>
<td>.33</td>
<td>.17</td>
<td>.02</td>
<td>.006</td>
<td>.34</td>
<td>.504</td>
<td>.956</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Q1 Crisis mgm t.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.46</td>
<td>.00</td>
<td>.002</td>
<td>.027</td>
<td>.57</td>
<td>.229</td>
<td>.959</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Q1 Prof.</td>
<td>.12</td>
<td>.19</td>
<td>.002</td>
<td>.032</td>
<td>.098</td>
<td>.17</td>
<td>.23</td>
<td>.37</td>
<td>.34</td>
<td>.40</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Q1 Clinical Com pt.</td>
<td>.33</td>
<td>.12</td>
<td>.02</td>
<td>.002</td>
<td>.024</td>
<td>.06</td>
<td>.07</td>
<td>.31</td>
<td>.31</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

*Significant with a p value of ≤.05 **Significant with a p value of ≤.01
BIBLIOGRAPHY


Chen, C., Campbell, V., and Suleiman, A. (2001). *Predicting student performances at a*
PREDICTING SUCCESS IN RESIDENCY


PREDICTING SUCCESS IN RESIDENCY


admission interview ratings of medical school matriculants and non-matriculants. 

*Academic Medicine, 67*(12), 865.


*Graduate Medical Education Data Resource Book.* Retrieved February 9, 2012 from
PREDICTING SUCCESS IN RESIDENCY


checklists do not capture increasing levels of expertise. *Academic Medicine*, 74, 112—34.


Ko, M., Heslin, K., Edelstein, R., Grumbach, K. (2007). The role of medical education in
PREDICTING SUCCESS IN RESIDENCY

reducing health care disparities: the first ten years of the UCLA/Drew Medical Education Program. *Journal of General Internal Medicine, 22*(5), 625-631.


PREDICTING SUCCESS IN RESIDENCY


Noble, J., Sawyer, R. (2002). *Predicting different levels of academic success in College using high school GPA and ACT composite score*. ACT research report series. Iowa City, IA.

Noble, J. (2003). *The effects of using ACT composite score and high school average on College admission decisions for racial/ethnic groups*. ACT research report series. Iowa City, IA.


PREDICTING SUCCESS IN RESIDENCY

culture of academic medicine: Gender matters, a national study. *Journal of General Internal Medicine*, 28(2), 201-207.


Shen, S., Comrey, A.L. (1997) Predicting medical residents' academic performances by
their cognitive abilities and personality characteristics. *Academic Medicine*, 72(9), 781-786.


PREDICTING SUCCESS IN RESIDENCY

Examination (USMLE) Examination Results to Predict Later In-Training Examination Performance Among General Surgery Residents. *Journal of Surgical Education, 67*(6), 452-6.


PREDICTING SUCCESS IN RESIDENCY


USMLE (2009) 2010 Bulletin of Information:. A publication of the FSMB and the NBME.


VITA

Chris Ashley Fox was born in Canada and resided there until after she completed her Bachelors of Arts degree at Mount Allison University. She continued her education by pursuing a Masters of Linguistics at Iowa State University and then ultimately a PhD in Educational Leadership and Policy Analysis at the University of Missouri-Columbia. She has extensive experience working in medical education and faculty development having worked in several Higher Education institutions, in both undergraduate and graduate education programs, across the United States for the past 15 years.