Measuring Medical Student Anxiety Toward the Male Genitourinary Rectal Examination

BY

DEBORAH MOULTON ROONEY
B.A., Michigan State University, 1990
B.S., Michigan State University, 1991
M.A.M.S., University of Illinois at Chicago, 1997

THESIS

Submitted as partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Psychology in the Graduate College of the University of Illinois at Chicago, 2012

Chicago, Illinois

Defense Committee:

Everett V. Smith, Jr. Chair and Advisor
George Karabatsos
Kimberly A. Lawless
Carla M. Pugh, Northwestern University
Debra A. DaRosa, Northwestern University
ACKNOWLEDGEMENTS

I am indebted to a number of people, all who have contributed to this dissertation and my continued growth as an academic. I’d like to thank my advisor and dissertation committee chair, Dr. Everett Smith. I am particularly grateful for his generous time and patience throughout my studies and my dissertation research. His professional guidance, straightforward advice, and feedback have been crucial for my completion of this dissertation. I am also deeply grateful to the other members of my dissertation committee who have each provided invaluable assistance to me as I developed my dissertation —Drs. George Karabatsos, Kimberly Lawless, Debra DaRosa and Carla Pugh. My work is stronger because of their questions and insightful comments. I am thankful to Dr. Karabatsos for his expert advice on measurement model selection processes and continued patience with all things mathematical. I am thankful to Dr. Lawless for her expert advice on cognitive processes. Finally, I am thankful to Drs. DaRosa and Pugh for their expert advice as medical educators and their continuing mentorship they have graciously offered me.

Although not on my committee, I would also like to thank Dr. Carol Myford, who has continued to be encouraging and supportive throughout my dissertation process. I would like to also thank researchers who have committed effort in this area of interest. In particular, I’d like to thank Mr. Kevin Hayes, who graciously allowed me to include his work in my dissertation. And finally, I would finally like to thank my husband, Steve, and my daughters, Emily and Abby, for their understanding and unwavering support during my studies and dissertation process.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Background</td>
<td>1</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>4</td>
</tr>
<tr>
<td>Purpose of the Study</td>
<td>6</td>
</tr>
<tr>
<td>Research Questions</td>
<td>8</td>
</tr>
<tr>
<td>Significance of the Study</td>
<td>9</td>
</tr>
<tr>
<td>II. LITERATURE REVIEW</td>
<td>10</td>
</tr>
<tr>
<td>Defining Anxiety</td>
<td>10</td>
</tr>
<tr>
<td>Key Theories That Describe the Relationship of Anxiety, Personality, and Learning</td>
<td>12</td>
</tr>
<tr>
<td>Relationship of personality and anxiety</td>
<td>12</td>
</tr>
<tr>
<td>State-Trait Anxiety Theory</td>
<td>13</td>
</tr>
<tr>
<td>Attentional Control Theory</td>
<td>14</td>
</tr>
<tr>
<td>Considerations Relating to Anxiety, Learning, and Performance</td>
<td>20</td>
</tr>
<tr>
<td>Effects of Anxiety on Learning, Performance, and Performance Measures</td>
<td>20</td>
</tr>
<tr>
<td>Long-term effects of anxiety</td>
<td>23</td>
</tr>
<tr>
<td>Methods used to alleviate or minimize learner anxiety</td>
<td>25</td>
</tr>
<tr>
<td>Traditional Role of Anxiety in Medical Education</td>
<td>28</td>
</tr>
<tr>
<td>Medical student anxiety toward intimate examinations</td>
<td>28</td>
</tr>
<tr>
<td>Common teaching practices of intimate examination skills</td>
<td>29</td>
</tr>
<tr>
<td>Research on medical student anxiety during intimate examinations</td>
<td>31</td>
</tr>
<tr>
<td>Difficulties of measuring medical student anxiety during GUR examination</td>
<td>33</td>
</tr>
<tr>
<td>Summaries of Literature Review Explaining the Relationship of Medical Student Anxiety and Learning</td>
<td>34</td>
</tr>
<tr>
<td>Relationship of gender and medical students’ anxiety</td>
<td>35</td>
</tr>
<tr>
<td>Theories applied to performing the male genitourinary rectal examination</td>
<td>36</td>
</tr>
<tr>
<td>Application of Theories to Build the Construct</td>
<td>38</td>
</tr>
<tr>
<td>Validity Evidence</td>
<td>39</td>
</tr>
<tr>
<td>Defining Validity Using the Standards for Educational and Psychological Testing</td>
<td>42</td>
</tr>
<tr>
<td>Evidence based on test content</td>
<td>42</td>
</tr>
<tr>
<td>Evidence based on response processes</td>
<td>42</td>
</tr>
<tr>
<td>Evidence based on internal structure</td>
<td>42</td>
</tr>
<tr>
<td>Evidence based on relationship to other variables</td>
<td>43</td>
</tr>
<tr>
<td>Evidence based on consequences of testing</td>
<td>43</td>
</tr>
<tr>
<td>An Alternative Framework for Examining Validity Evidence</td>
<td>44</td>
</tr>
<tr>
<td>Classical test theory</td>
<td>44</td>
</tr>
<tr>
<td>The Rasch Measurement Models</td>
<td>46</td>
</tr>
<tr>
<td>Evidence relevant to the content aspect of validity</td>
<td>49</td>
</tr>
<tr>
<td>Evidence relevant to the substantive aspect of validity</td>
<td>50</td>
</tr>
<tr>
<td>Evidence relevant to the structural aspect of validity</td>
<td>53</td>
</tr>
<tr>
<td>Evidence relevant to the generalizability aspect of validity</td>
<td>57</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Male Genitourinary Rectal (GUR) Examination</td>
<td>234</td>
</tr>
<tr>
<td>B. Sample From STAI Form Y-1</td>
<td>237</td>
</tr>
<tr>
<td>C. Sample From STAI Form Y-2</td>
<td>238</td>
</tr>
<tr>
<td>D. Hamilton Anxiety Rating Scale (HAM-A)</td>
<td>239</td>
</tr>
<tr>
<td>E. Anxiety In Medical Students Questionnaire</td>
<td>243</td>
</tr>
<tr>
<td>F. M2 Student Evaluation Form 1</td>
<td>244</td>
</tr>
<tr>
<td>G. Student Evaluation Form 1</td>
<td>245</td>
</tr>
<tr>
<td>H. Student Evaluation Form I-GUR</td>
<td>246</td>
</tr>
<tr>
<td>I. 2010 Content Validity Rating Form A-GUR Instrument</td>
<td>247</td>
</tr>
<tr>
<td>J. 2010 Anxiety Toward Male GUR Exam Instrument</td>
<td>250</td>
</tr>
<tr>
<td>TABLE</td>
<td>PAGE</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>2. Validity Propositions, Research Questions, and Validity Evidence to Address Research Question 1: Does Evidence Based on the Content Aspect of Validity Support the Use of Student-Anxiety Scores to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?</td>
<td>124</td>
</tr>
<tr>
<td>3. List of Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 2: Does the Evidence Based on the Substantive Aspect of Validity Support the Use of Student Anxiety Scores to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?</td>
<td>126</td>
</tr>
<tr>
<td>4. List of Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 3: Does the Validity Evidence Based on the Structural Aspect of Validity Support the Use of Student Anxiety Scores to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?</td>
<td>127</td>
</tr>
<tr>
<td>5. Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 4: Does the Validity Evidence Based on the Generalizability Aspect of Validity Support the Use of Student Anxiety Scores of the New Scale to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?</td>
<td>128</td>
</tr>
<tr>
<td>6. Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 5: Does the Validity Evidence Based on the External Aspect of Validity Support the Use of Student Anxiety Scores of the New Scale to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?</td>
<td>129</td>
</tr>
<tr>
<td>8. Loading Pattern of Positively-Loaded Items on the First Extracted Component from the Principal Component Analysis of the Standardized Item Residuals (Items Sorted by Loadings)</td>
<td>178</td>
</tr>
<tr>
<td>9. Loading Pattern of Negatively-Loaded Items on the First Extracted Component from the Principal Component Analysis of the Standardized Item Residuals (Items Sorted by Loadings)</td>
<td>179</td>
</tr>
<tr>
<td>10. A Summary Of Deviances for the Two Model Options</td>
<td>182</td>
</tr>
<tr>
<td>11. The AIC and BIC of the Two Models Under Consideration for the Analysis of the A-GUR</td>
<td>183</td>
</tr>
<tr>
<td>12. The AIC and BIC of the Three Forms of the Facets Model Under Consideration for the Analysis of the A-GUR</td>
<td>186</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Construct model of Baddeley’s working memory model.</td>
</tr>
<tr>
<td>2.</td>
<td>Bandura’s model of triadic reciprocal causation.</td>
</tr>
<tr>
<td>3.</td>
<td>Cognitive process of Eysenck’s model of attentional control.</td>
</tr>
<tr>
<td>4.</td>
<td>The relationship of sources of medical student anxiety toward the male GUR examination.</td>
</tr>
<tr>
<td>5.</td>
<td>Construct map of measuring medical student anxiety toward the male GUR examination.</td>
</tr>
<tr>
<td>6.</td>
<td>Theoretical construct map of measuring medical student anxiety toward the male GUR examination.</td>
</tr>
<tr>
<td>7.</td>
<td>An empirical construct map showing the relationships among medical student anxiety, student gender, student ethnicity, and items’ difficulty to endorse.</td>
</tr>
<tr>
<td>9.</td>
<td>Theoretical Construct map of measuring medical student anxiety toward the male GUR examination.</td>
</tr>
<tr>
<td>10.</td>
<td>Modified construct map of measuring medical student anxiety toward the male GUR examination.</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGFI</td>
<td>Adjusted Goodness of Fit Indices</td>
</tr>
<tr>
<td>A-GUR</td>
<td>Anxiety Toward the Male Genitourinary Rectal</td>
</tr>
<tr>
<td>AERA</td>
<td>American Educational Research Association</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>APA</td>
<td>American Psychological Association</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike’s Information Criterion</td>
</tr>
<tr>
<td>BAI</td>
<td>Beck Anxiety Inventory®</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>CT</td>
<td>Cognitive therapy</td>
</tr>
<tr>
<td>DFF</td>
<td>Differential Facet Functioning</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>DSM</td>
<td>Diagnostic and Statistical Manual of Mental Disorders</td>
</tr>
<tr>
<td>GFI</td>
<td>Goodness of Fit</td>
</tr>
<tr>
<td>GHQ</td>
<td>General Health Questionnaire</td>
</tr>
<tr>
<td>GUR</td>
<td>Genitourinary Rectal</td>
</tr>
<tr>
<td>HADS</td>
<td>Hospital Anxiety and Depression Scale</td>
</tr>
<tr>
<td>HAM-A</td>
<td>Hamilton Anxiety Scale</td>
</tr>
<tr>
<td>IRF</td>
<td>Item Response Function</td>
</tr>
<tr>
<td>MAR</td>
<td>Missing at Random</td>
</tr>
<tr>
<td>MCAR</td>
<td>Missing Completely at Random</td>
</tr>
<tr>
<td>NCME</td>
<td>National Council on Measurement in Education</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>POMS</td>
<td>Profile of Mood States</td>
</tr>
<tr>
<td>PSS</td>
<td>Perceived Stress Scale</td>
</tr>
<tr>
<td>PSWQ</td>
<td>Penn State Worry Questionnaire</td>
</tr>
<tr>
<td>PI</td>
<td>Patient Instructor</td>
</tr>
<tr>
<td>RSM</td>
<td>Rational Emotive Behavior Theory</td>
</tr>
<tr>
<td>SAS</td>
<td>Self-Rated Anxiety Scale</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard Error of Measurement</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>SCL-90R</td>
<td>Symptom Checklist-90 Revised</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

Background

James, a first-year medical student, stood outside the examination room, clipboard in hand. Shoulders stooped, he was uncharacteristically quiet. On any other day he would have been lightly joking with one of his classmates to pass the time. Not today. Today was different. In a few minutes he and the other three students in his learning group would be prompted to enter the examination room. Today the students would be learning how to perform the male genitourinary rectal (GUR) examination. While he waited, James wiped his damp hands on his pants and tried to review the steps of the examination in his head to calm his nerves. First step was penis inspection and palpation, then testes, followed by hernia assessment, rectum inspection, and finally, the prostate examination. James felt his heart beating in his chest now. He wasn’t looking forward to this part of his medical education. What if he hurt the standardized patient, the person trained and paid to portray the patient? What if he fumbled or became flustered? He blushed at this prospect. His peers would notice, for sure.

Across the hall, Asha, a second-year student, was already in the examination room waiting her turn to perform the exam. As she watched one of her team members palpate the epididymis (the narrow tube that carries sperm from the testicle) she could hear Paul, the standardized patient, explaining the proper technique for this portion of the examination. In spite of his reassuring tone, Asha’s thoughts raced. Asha felt as if she had forgotten everything from last year’s introduction. What if she forgot to perform a step of the exam? What if she couldn’t identify a real pathology? Was she really cut out for medicine?
Although not actual incidents, these scenarios portray authentic student experiences that occur every year across the 140 medical schools in the United States as students are introduced to patient examinations. Learning the skills associated with patient examinations is naturally anxiety-evoking as students are introduced to new clinical experiences and patient interactions. Student anxiety is amplified while they learn and perform intimate examinations—examinations of clients’ breast, pelvic, prostate, and genitourinary systems. Intimate examinations, associated with a high amount of patient contact and the highest level of intimacy, continue to be a source of apprehension for most medical students (Abraham, 1996; Behrens et al., 1979; Campbell, McBean, Mandin, & Bryant, 1994; Ker, 2003; Popadiuk, Pottle, & Curran, 2002; Rochelson, Baker, Mann, Monheit, & Stone, 1985). For many students, learning the male GUR examination, as described in detail in Appendix A, may be the most anxiety-evoking of the intimate examinations.

Because early detection of prostate disease has historically been problematic, the need for skilled physicians that encourage their male patients to have regular genitourinary and rectal examinations must play a key role in maintaining men’s health. In order to produce such physicians, the undergraduate medical curriculum must identify potential student anxieties early in their training of the male GUR examination, and address these anxieties early in their education to minimize the potential adverse effects recent research has presented.

Historically, student anxiety has been considered by medical educators to be a normal element of initiation to becoming a practicing physician (Aktekin et al., 2001; Becker, Hughes, Geer, & Strauss, 1961; Haas & Shaffir, 1987; Moss & McManus, 1992; Pitkälä & Mäntyranta, 2004; Shuval, 1967). Student anxiety has been most closely associated with initial patient contact and understood to dissipate as students gain experience
(Becker et al., 1961; Haas & Shaffir, 1987; Shuval, 1967). In spite of this understanding—that a student’s anxieties dissipate as the student gains more practice—there is research that has suggested that anxiety’s influence on learning is largely inhibitive (Calvo & Carreiras, 1993; Phelps, 2006; Shors, 2006). In recent years researchers have acknowledged the negative impact of anxiety on medical students’ learning clinical skills (de Saintonge & Dunn, 2001; Stewart, Lam, Betson, Wong, & Wong, 1999). Although some researchers have suggested that there are learning benefits in the form of increased motivation and adrenergic response (Bryant & Harvey, 1995; Cassady & Johnson, 2002; Shors, 2006), a body of research supports the view that for medical students, the act of performing physical examinations creates a state of anxiety that may adversely affect learning and the performance of physical examinations (Huebner, Royer, & Moore, 1992; Mavis, 2001; Sarikayaya, Civander, & Kalaca, 2006).

As much as medical educators seem to understand the specific sources of persistent student anxieties while they learn and perform the pelvic and breast exam, little is known about student anxieties toward the male GUR exam. Furthermore, the same interventions that are meant to alleviate or minimize student anxiety toward the pelvic and breast exams may not be applicable when performing the male GUR examination. In order to improve the teaching programs for all of the intimate examinations, each form of intimate examination must be investigated and researchers need to identify specific sources of student anxiety for each type of intimate exam. With clear identification of student anxieties while they learn and perform the male GUR exam, educators will have a better understanding of the precise sources of anxiety during the exam and can modify the associated curriculum to better address and minimize student concerns early in their training.
Statement of the Problem

In spite of increased interest in the identification of student attitudes while they learn and perform intimate exams, most studies have focused on pelvic (Buchwald, 1979; Carr & Carmondy, 2004; Vontver et al., 1980; Wånggren, Pettersson, Csemiczky, & Gemzell-Danielsson, 2005) and breast examinations (Pugh & Salud, 2007). Fewer studies relate to student attitude and the process of learning and performing the male GUR exam (Hennigan, Franks, Hocken, & Allen-Mersh, 1991; Howley & Dickerson, 2003; Lawrentschuk & Bolton, 2004; Robins, Alexander, Dickens, Belville, & Zweifler, 1997). To date, only one study has attempted to identify specific factors that may influence students’ anxiety toward the male GUR exam (Robins et al., 1997). Regardless of the reasons that may explain the dearth of studies relating to the GUR exam, literature reveals no studies that have adequately examined the evidence supporting validity of the inferences of the data from instruments that identify and measure student anxiety toward the male GUR examination.

In spite of the lack of research in identifying and measuring the specific sources of student anxiety toward the male GUR examination, Sarikayaya et al. (2006) summarized a variety of studies that have presented strong evidence for the need to do so. For example, the authors stated that in one study (Sarikayaya et al., 2006) the rectal examination was deemed the highest source of anxiety of 11 clinical skills identified and reviewed by a group of medical students. In another study, the majority (97%) of students surveyed felt that the ability to perform a digital rectal exam (DRE) was a key component of the male GUR examination and an essential requirement for a medical practitioner. In spite of this, only 30% of the students felt their clinical school had been supportive in teaching them how to perform a DRE, and almost half the students (48%) lacked confidence to give an opinion based on their findings (Lawrentschuk & Bolton, 2004).
A number of researchers have suggested that the anxiety and lack of confidence present while learning to perform the male GUR exam may lead to profound, long-term, adverse effects, and influence the preclinical and clinical success of performing the GUR examination (Brenner, Hergenroeder, Kozinetz, & Kelder, 2003; Grover & Smith, 1981; Hennigan et al., 1991; Ker, 2003; McMahon, Marina, Kritek, & Katz, 2005). These effects may be so profound and residual that they may even affect a student’s decision to perform a male rectal exam long after their medical training is completed and well into clinical practice (Brenner et al., 2003; Turner & Brewster, 2000). These negative consequences of residual anxiety and lack of confidence on clinical decisions are most troubling. Brenner et al. (2003) posited that a lack of physician confidence may be a primary barrier to providing testicular self-examination training to patients. They examined the practices of residents of two pediatric residency programs in the United States. They suggested that, in spite of having the opinion that testicular cancer is a concern, the lack of confidence in performing a testicular examination was the most noted reason for not performing testicular exams, or teaching patients testicular self-exams. Turner and Brewster (2000) evaluated the reasons for decreased performance of the digital rectal examinations (DRE) by medical students and physicians. Outcomes suggested that decreased confidence in this clinical skill leads to decreased incidences of performing the physical examination as practicing physicians.

Most troubling, some researchers suggested that residual anxiety while learning intimate examinations as medical students, decreases incidences of performing intimate examinations as practicing physicians, even when clinically indicated (Brenner et al., 2003; Hennigan et al., 1991; Ker, 2003). These finding are disturbing because testicular cancer is the most common cancer among 15-35 year old males (Brenner et al., 2003). A third of
rectal cancers are palpable, and the omission of the rectal exam would delay surgical referral (Dixon, Thornton-Holmes, & Cheetham, 1990; Lawrentschuck & Bolton, 2004).

**Purpose of the Study**

In order to obtain a comprehensive understanding of the sources of medical students’ anxiety while they learn and perform the male GUR exam, the sources of anxiety and their potential influences on students’ performance of the male GUR examination were identified. Other elements that may influence student anxiety during the exam, such as student gender and personality, were also examined. An instrument that measures the variety of components that may influence a medical student’s level of anxiety toward the male GUR must be developed, and validity evidence for use for this population must be examined.

A review of current methods used to reduce student anxiety while students learn and perform intimate examinations is also important for the development of a more effective curriculum for the male GUR examination. Before improvements can be made to the male GUR skills curriculum, educators must identify and measure student anxiety toward the exam. The design of a new instrument that measures student anxiety toward the male GUR examination, and examination of evidence that supports or refutes validity arguments of the instrument is crucial to this process. This instrument may potentially identify specific characteristics of student anxiety toward the exam, and therefore better guide precise improvements to the curriculum.

This dissertation research will be used to provide a formal definition of anxiety, an extensive review of the study of anxiety in the context of medical education, and a review of the most commonly utilized anxiety scales. Based on this research, I will create a new instrument that may better identify the sources of student anxiety relating to the male GUR exam. Although prior studies (Robins et al., 1997; Howley & Dickerson, 2003) have
introduced instruments that identify student anxiety toward the digital rectal exam, an integral part of the GUR exam, these instruments lack the sensitivity to identify the sources of student anxiety toward the GUR exam. I propose that a new instrument, based on a strong theoretical foundation will offer improvements that will lead to a much more robust scale. Rasch measurement models (Rasch, 1960) can then be applied in order to review the validity evidence for inferences made based on data obtained from analysis of the new instrument. This research may inform educators about the specific characteristics of anxiety that influence students’ learning experience toward the male GUR examination.

Validation efforts as defined by Messick (1989, 1995) and the American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME; 1999) provided necessary evidence that supports credible inferences about the characteristics of student anxieties to medical educators and the Liaison Committee on Medical Education, the accrediting authority for medical-education programs. Examination of validity evidence based on the variety of aspects of validity was performed. This evidence may be used to support the use of student-anxiety scores to make inferences about medical students’ anxiety toward the male GUR examination. Validity evidence based on content aspect of validity may offer evidence that the new scale’s items fit the intended purpose of the scale, while evidence based on substantive aspect of validity may demonstrate anticipated response patterns. Additionally, evidence based on the structural aspect of validity may suggest that the relationship between items is consistent with intended structure of instrument, while evidence based on the generalizability aspect of validity may suggest that the meaning of the scale’s measures is maintained across contexts. Finally, evidence based on the external aspect of validity may indicate that student anxiety scores of the new scales correlate with the scores of the
short-form STAI and suggest that the new scale’s measures are consistent with the previously validated instrument.

**Research Questions**

The research question that I address in this study is: Does evidence based on the content, substantive, structural, generalizability, and external aspects of validity support the use of student-anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam? Broken down into specific areas, the research question becomes five areas of inquiry. These are:

1. Does evidence based on the *content* aspect of validity support the use of student-anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

2. Does evidence based on the *substantive* aspect of validity support the use of student anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

3. Does validity evidence based on the *structural* aspect of validity support the use of student anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

4. Does validity evidence based on the *generalizability* aspect of validity support the use of student anxiety scores of the new scale to make inferences about medical students’ level of anxiety toward the male GUR exam?

5. Does validity evidence based on the *external* aspect of validity support the use of student anxiety scores of the new scale to make inferences about medical students’ level of anxiety toward the male GUR exam?
Significance of the Study

The analysis provides validity evidence that will either support or refute the use of the newly developed scale to measure medical student anxiety toward the male GUR exam. The findings from the study will potentially help direct clerkship directors to modify the current curriculum in order to address and decrease controllable student anxieties while students learn and perform the male GUR exam. Improvements to the curriculum can lead to a more effective learning environment and increased learning for the medical student. A scale supported by empirical validity evidence can be utilized by other medical programs for identifying sources of medical student anxiety while they learn to perform the male GUR exam. Further, the development of a skills curriculum that supports a more comfortable learning environment for the medical student while they learn to perform the male GUR exam may lead to increased students’ skill and commitment to performing male GUR examinations as practicing clinicians. The goal is to produce generations of clinicians that are wholly comfortable performing male GUR exams and proactively encourage their male patients to have regular genitourinary and rectal examinations, and ultimately substantially decrease the national mortality rate from testicular cancer.
II. LITERATURE REVIEW

Defining Anxiety

The construct of anxiety in the context of medical education has been researched for the past five decades. In spite of the continuous interest by medical educators, there has been no standardized definition offered by the researchers in this area. Given the complexity of the anxiety construct, there are a number of approaches in which to operationally define anxiety, depending on the theoretical foundation. Although not referred to explicitly, there seems to be a trend in medical research to operationally treat anxiety in one of two ways. Some researchers in medical education appear to apply the construct and nomenclature offered by Seligman, Walker, and Rosenhan (2001). These authors explain anxiety as a state or mood characterized by four components or symptoms: the (a) cognitive, (b) physiological, (c) emotional, and (d) behavioral. The cognitive component of this theory reflects an individual’s anticipation of a vague danger. Cognitive symptoms may include decreased attention or focus and decreased ability to retrieve memory. Physiological symptoms are a body’s physical responses, such as nausea or stomachache, chest pain or increased heart rate, and increased respiration. Physiological indications of anxiety, called “somatic symptoms,” may include pallor (pale skin), sweating, trembling, and pupillary dilation. Emotional manifestations of anxiety include a sense of dread or panic, and physically cause nausea and chills. Behaviors, both voluntary and involuntary, compel a flight response, an attempt to avoid the source of anxiety. Manifestations of behavioral symptoms of anxiety include pacing, trembling, hand wringing, and finger tapping. These components combine to create the feelings that we typically recognize as fear that are
activated by an individual’s autonomic “fight or flight” response (Beck, Emery, & Greenberg, 1985). Externally, visible symptoms, called manifestations, or phenomena, can often be identified for the physiological, emotional and behavioral components of anxiety. Low levels of anxious mood leads to nervousness, while a high surge of anxiety may lead to panic. Not unlike the previous construct, the construct offered by Sieber, O’Neil and Tobias (1977) has been tailored to the learning setting. Here anxiety is also defined operationally in terms of four similar categories: (a) phenomenology of anxiety, (b) physiologic responses of anxiety, (c) task performance, and (d) the conditions that affect the anxiety process. Sieber et al. describe phenomenology of anxiety as “the individual’s conscious awareness of the anxiety-state reaction” (p. 27), while the physiologic responses are bodily reactions to anxiety, such as an increase in heart rate. Task performance can be affected by anxiety, and therefore is measured to evaluate the effectiveness of coping mechanisms or problem-solving skills. Finally, all internal and external conditions that affect the anxiety process are taken into consideration. For example, any thought processes, prior learning, or behaviors that may affect the anxiety process are included in this category. By comparison to the definition offered by Seligman et al., this construct acknowledges the four components (cognitive, physiological, emotional, and behavioral) of the prior construct while considering two additional aspects specific to the learning environment—the conditions that affect anxiety and anxiety’s relationship to task performance.

Anxiety has many connotations in general use. Even in medical education, many researchers have used the terms anxiety and stress almost interchangeably. For use in this study, anxiety needs to be differentiated from stress, and viewed as a distinct entity. Stress is a physical, chemical, or emotional stimulus that causes tension or is the result of such stimulus, while anxiety is a response to that stressor. For the purpose of this study, the
operational definition of anxiety is based on Eysenck, Derakshan, Santos, and Calvo’s (2007) definition: “An aversive emotional and motivational state occurring in threatening circumstances” (p. 336). Eysenck et al. based this definition of anxiety on Spielberger’s (1972) state–trait anxiety theory. Spielberger deconstructed the notion of anxiety into state anxiety, a temporary form of anxiety that relates to a particular situation or condition, and trait anxiety, a condition of personality that harbors worrisome or anxious thoughts. These operational definitions of state and trait anxiety will be referred to regularly throughout this paper. Further, I will describe the key theories that explain the relationship between anxiety, personality, and learning.

**Key Theories That Describe the Relationship of Anxiety, Personality, and Learning**

There are a vast number of schema and theories that exist to explain the relationship of anxiety, personality, and learning. For the purpose of this study, I have concentrated on specific definitions and contexts for the terms (a) state anxiety, (b) trait anxiety, (c) personality, and (d) learning. As presented earlier, the definitions of state and trait anxiety are offered by Spielberger’s State–Trait Anxiety Theory (Spielberger, 1972). Using State-Trait Anxiety Theory as foundation, I will describe the relationship of anxiety, personality and learning process as described by Eysenck et al. (2007).

**Relationship of personality and anxiety.** One widely-accepted working definition of personality as it relates to anxiety is based on Spielberger’s state–trait theory of anxiety (Spielberger & Gorsuch, 1966). The state–trait theory of anxiety differentiates between a person’s temporary emotional state and their stable disposition. A person’s likeliness to respond to an anxiety-evoking stimulus is described as a characteristic of an individual’s personality. A more recent theory from Eysenck et al. (2007) focused on the relationship of
state anxiety and cognitive processes during stressful encounters. A review of these
governing theories associated with situation-specific anxiety and its relationship with
cognitive processes has laid a theoretical foundation for this study that will attempt to
measure anxiety in medical students as they learn and perform the male GUR examination.

**State–Trait Anxiety Theory.** Spielberger (1972) based the state–trait theory on
concepts that were first introduced by Cattell (1966) and Cattell and Scheier (1961, 1963).
Spielberger described state and trait anxiety as “emotional reactions as expressions of
personality states” (Spielberger, 1983, p. 4). Further, he posited that emotional states exist at
any given moment and are associated with a specific level of intensity at that given time.
Spielberger defined anxiety and personality emotional states. The author characterized
anxiety states as subjective feelings of tension, apprehension, nervousness, and worry
(Spielberger, 1979). Anxiety states originate with the arousal of a person’s autonomic
nervous system—the same system that activates a person’s physiological responses.
Physiological responses are a person’s primitive, automatic, inborn response that prepares
the body to address or avoid perceived attack, harm, or threat to survival, resulting in what is
commonly known and described by Beck et al. (1985) as a “fight or flight” response.
Furthermore, an individual’s anxiety state, called *state anxiety* (S-Anxiety) by Spielberger,
is a transitory emotional state that is influenced by the individual’s relatively stable
disposition to respond to stressful situations.

Spielberger (1979) proposed that the disposition to respond to stressful situations is a
stable personality trait, which he labels *trait anxiety* (T-Anxiety), and based on the
constructs of Atkinson (1964). Further, Spielberger applied an analogy from physics to
explain the relationship of state and trait anxiety. He likened state anxiety to kinetic energy
that refers to a reaction at a given time with a given intensity, while trait anxiety was equivalent to potential energy, or an individual’s ability to react to a stressful situation.

**Attentional control theory.** The *attentional control theory* developed by Eysenck et al. (2007) identifies the complex relationship of state and trait anxiety on mental (cognitive) process and performance. Eysenck and Calvo (1992) suggested the existence of state anxiety is determined by an individual’s trait anxiety—one’s ability to react to stress, combined with a threatening situation. Their theory was based on the language from Spielberger’s state–trait anxiety model (Spielberger, 1972).

Eysenck et al. (2007) later posited that the individual in an anxious state is prone to worry about the threat to a current goal and will attempt to develop strategies to reduce their anxiety in order to achieve the desired goal. A number of researchers have agreed that uncontrolled anxiety during this type of scenario will likely have negative effects on the individual’s cognitive function or performance (Calvo & Alamo, 1987; Calvo & Carreiras, 1993; Darke, 1988; Eysenck & Byrne, 1992; Phelps, 2006; Shors, 2006).

Before applying the theory to the specific learning situation in this study, a review of the two assumptions and theoretical limitations of the theory must be made. First, we must consider the assumptions presented for the founding theory—the processing efficiency theory—as well as those for the attentional control theory, the basis for this study. There are two primary assumptions for the processing efficiency theory. The first assumption is that worry is the primary component of state anxiety that is responsible for the adverse effects anxiety has on performance effectiveness and efficiency. Defining *worry*, I relied on a commonly used definition from Borkovec, Robinson, Pruzinsky, and DePree (1983), “a chain of thoughts and images, negatively affect-laden, and relatively uncontrollable” (p. 10). Many studies have identified a consistent relationship between worry and anxiety. Kelly
(2004) noted a number of correlation studies (Constans, 2001; Davey, 1993; Wells, 1994) between the STAI and a variety of measures of worry. Kelly identified correlations ranging from .48 to .74, with an average of .61 in these studies.

Eysenck and his cohorts (Eysenck & Calvo, 1992; Eysenck et al., 2007) posited that worry’s effects on the learning process are twofold. The first is that worry-saturated thoughts tap into the attentional resources utilized by a working, short-term, memory with limited capacity, and ultimately uses available working memory for convergent task processes, potentially leading to adverse performance effects. Eysenck and Calvo suggested a dichotomy of sorts. High-anxiety individuals perceive threats to self-esteem in stressful situations that make them worry about negative consequences and personal inefficacy, ultimately decreasing efficacy. In some subjects the anticipation of undesirable outcomes may increase motivation to increase preparation effort in order to compensate for perceived weaknesses. Eysenck and Calvo’s (1992) processing efficiency theory suggests that when compensatory resources such as the ability to take more time to perform a task are unavailable, performance suffers. The attentional control theory is largely based on this concept. The foundation of the attentional control theory, the processing efficiency theory is based on the working memory model posited by Baddeley (1986). In its original form, Baddeley’s model of working memory (Baddeley & Hitch, 1974) was tripartite and composed of three primary elements (see Figure 1). These elements are named the *central executive*, the controlling element, and two short-term memory storage systems called *slave systems*. The slave systems control two perceptual domains: the phonologic loop (verbal domain) and visuo-spatial sketchpad (visuo-spatial domain).
Baddeley and Hitch (1974) suggested that the central executive is flexible, organizes information, coordinates the slave systems, and most important to the purposes of this study, is capable of shifting between tasks and selective attention and inhibition. The phonologic loop functions as a short-term phonological, or sound, store with aural memory traces and articulatory revival and rehearsal elements. The visuo-spatial sketchpad functions as a temporary storage of spatial and visual information, and manipulates spatial and visual information. There are concerns of the model presented by Jones, Macken, and Nicholls (2004), and Nairne (2002), but for the purpose of this study, they do not affect the proposed research questions so they will not be addressed here. Baddeley and Hitch posited that performance of two simultaneous tasks requiring the use of the two separate perceptual domains is nearly as efficient as performance of either task individually, but when an individual tries to carry out two tasks concurrently that use the same perceptual domain, performance is less efficient than when performing the tasks individually.

Additionally, Eysenck et al. (2007) pointed out the importance of distinguishing the two concepts that are the basics for the processing efficiency theory—effectiveness and
efficiency. *Effectiveness* refers to “the quality of task performance as indexed by standard behavior measures (generally response accuracy),” while *efficiency* refers to “the relationship between the effectiveness of performance and the effort or resources spent in task performance” (p. 336). Further, the authors explained that the effectiveness of performance declines as more resources (cognitive efforts) are devoted in order to attain a performance level (goal), resulting in a decrease of efficiency. Also, worry increases an individual’s motivation to minimize the anxiety state, and leads to enhanced efforts to minimize the anxiety state. In doing this, an individual must use available auxiliary processing resources (cognitive efforts) in order to minimize potential performance impairments. Using auxiliary processing resources may minimize adverse effects on performance, but generally at the cost of performance efficiency. In the event auxiliary resources are not available, the performance effectiveness suffers.

The second assumption presented by Eysenck et al. (2007) was that worry and associated anxiety affects performance and efficiency, particularly on tasks that require more from the processing and storage capacity of working memory. They further suggested that worrisome thoughts interfere with this function and tax the self-regulating mechanism and auxiliary processing used to inhibit these negative thoughts. The new attentional control theory addressed most of the limitations of the original theory, and identified an important distinction between processing efficiency and performance efficacy. A review of the constructs and propositions, as it relates to the distinction between processing efficiency and performance efficacy, will help clarify their relationship and explain the mechanism that controls anxiety effects, particularly as it relates to this application in medical education.

The assumptions of the attentional control theory help build the foundations on which my research questions are based. The fundamental assumption of the attentional
control theory is that worry is the component of anxiety that affects attentional processes, and ultimately affects performance. According to Eysenck et al. (2007), this ideal is founded on the assumptions of Power and Dalgleish (1997) that the state of anxiety is activated when a current goal is threatened. Secondly, detection of threat to the current goal increases an individual’s anxiety and causes the individual’s attention to be directed toward the source of anxiety and determine a response to that particular threat. Threat stimulus may be internal (negative or worrisome thoughts), or external (threatening distracters). Existing research in the field of attentional bias supports that anxious individuals are more likely to allocate attention to threat-related stimuli, or, more likely, have delayed disengagement from the stimuli, particularly in the presence of neutral stimuli (e.g., Egloff & Hock, 2001; Eysenck & Byrne, 1992; Fox, Russo, & Dutton, 2002; Mogg & Bradley, 1998; Mogg et al., 2000; Wilson & MacLeod, 2003). The third construct of the attentional control theory is that anxiety impairs the inhibitive function. An individual’s inhibitive function is based on their ability to properly distinguish between current goals and conspicuous stimuli.

Based on the research from Corbetta and Shulman (2002), Eysenck et al. (2007) suggested that an individual is influenced by two distinct attentional systems: the first, the goal-directed attentional system, is influenced by current goals, expectations and knowledge, and is involved in the “top-down,” or big-picture, control of attention. The opposing system is the stimulus-driven attentional system (the “bottom-up” attentional control system) and is active when behaviorally relevant, salient sensory events are detected. Under normal conditions the two systems interact and help direct the individual’s cognitive processing. Using this as framework, Eysenck et al. suggested that anxiety interrupts the flow between the two systems by increasing influence of the stimulus-driven attentional system via automatic response and decreasing influence of the goal-directed attentional system.
results in increased effects of the stimulus-driven attentional system. It seems reasonable that this occurs more frequently under stressful conditions.

This framework is applied to the central executive of Baddeley’s working memory model fairly seamlessly using a theoretical approach based on the work of Miyake et al., (2000). The work by Miyake et al. operationalized the basic control functions of the central executive relating to the goal-directed attention system. Eysenck et al. (2007) reviewed the three major functions of the central executive identified by Miyake et al.:

1. Inhibition: “One’s ability to deliberately inhibit dominant, automatic, or preponent responses when necessary” (Miyake et al., 2000, p. 57); this involves using attentional control to resist disruption or interference from task-irrelevant stimuli or responses (as quoted in Eysenck et al., 2007, p. 338).

2. Shifting: “Shifting back and forth between multiple tasks, operations, or mental sets” (Miyake et al., 2000, p. 56); this function involves adaptive changes in attentional control based on task demands (as quoted in Eysenck et al., 2007, pp. 338-339).

3. Updating: “Updating and monitoring of working memory representations” (Miyake et al., 2000, p. 56, as quoted in Eysenck et al., 2007, p. 339)

The fourth construct maintains that anxious individuals are more distracted by task-irrelevant stimuli, regardless of the source of the stimuli. Anxiety impairs processing efficiency because it reduces attentional control in the presence of internal (worrisome thoughts) and external (common distracters) anxiety-inducing stimuli. As a result, processing resources are more likely to be diverted from task relevant stimuli to task-irrelevant stimuli and involve the inhibition and/or shifting functions. Because the
attentional control theory is heavily founded on the principle that worry directly affects cognitive processes, this study also briefly discusses the complex relationship of worry, anxiety, and the instruments used to measure anxiety.

Considerations Relating to Anxiety, Learning, and Performance

Although discussion of the theoretical basis of the effects of anxiety on performance is helpful, it might be more informative to review any practical effects anxiety may have on learning and performance.

Effects of Anxiety on Learning, Performance, and Performance Measures

The effects of anxiety on learning have been examined and opinions vary. Although some researchers have posited that small amounts of anxiety may be beneficial to learning and performance (de Saintonge & Dunn, 2001; Ruebush, 1960; Yerkes & Dodson, 1908), as described earlier, some contemporary cognitive researchers (Calvo & Alamo, 1987; Calvo & Carreiras, 1993; Darke, 1988; Eysenck & Byrne, 1992; Phelps, 2006) have suggested that anxiety has detrimental performance effects. The following is a brief review of relevant research on the topic.

There have been a number of researchers that have suggested there are learning and performance benefits of small amounts of anxiety. The Yerkes-Dodson Law (Yerkes & Dodson, 1908) suggested that a moderate amount of stress may be beneficial to performance, whereas too much is detrimental. Bryant and Harvey (1995), proponents of the Yerkes-Dodson law, have suggested that some anxiety may lead to increased motivation and adrenergic response. This position is supported by de Saintonge and Dunn (2001), who suggested that for male subjects, fear of negative evaluation seemed to cause anxiety that ultimately resulted in high achievement. The same study indicated that for females, anxiety
was inversely proportional to their achievement, indicating that anxiety may have negative effects on performance for female subjects.

Some researchers have found evidence that suggests acute anxiety may help recall information in experimental studies, particularly if the information itself is stress-inducing. For example, a study by Cahill, Gorski, and Le (2003) indicated an improvement in short-term memory recall in participants who were shown visually stimulating images while their arm was plunged in ice-cold water. The authors’ report suggested that the participants that experienced the stressor recalled more of the arousing pictures and in greater detail than did those that did not experience the stressor, while the acute stressor did not affect the recall of neutral words, implying that exposure to a relatively intense stressful experience in humans is associated with enhanced recall of information that is arousing in nature. A number of other studies that were limited to experimental conditions with adult male subjects indicated similar findings (Cahill et al., 2003; Domes, Heinrichs, Rimmere, Reichwald, & Hautzinger, 2004; Wolfe, Schommer, Hellhammer, Reischies, & Kirschbaum, 2002). Although this position has been retained by many experimental researchers, other researchers have developed new theories regarding the effects of anxiety on learning and performance.

Shors’ work is a blend of two positions (2006). Consistent with the experimental researchers’ opinion, Shors agreed that exposure to an acute stressful event tends to enhance learning of new information, if there is an effect at all. In the face of this position, Shors also suggested that exposure to a similar stressful event tends to impair the retrieval of information that has already been acquired, suggesting it is the retrieval process of learning that may be impaired by anxiety. Further, Shors suggested that the effects of stressful
experiences on subsequent learning are diverse and are attributed to the dynamic nature of the situation, and a variety of organismal properties.

Many cognitive researchers (Calvo & Alamo, 1987; Calvo & Carreiras, 1993; Darke, 1988; Eysenck & Byrne, 1992; Phelps, 2006, Ruebush, 1960) agreed that anxiety is associated with performance impairments, particularly in complex cognitive tasks. Eysenck and Calvo (1992) developed the processing efficiency theory which suggested there us a dichotomy: high-anxiety individuals perceive threats to self-esteem in test situations that make them worry about negative consequences and personal inefficacy, and ultimately decrease efficacy, while in some subjects the anticipation of undesirable outcomes may increase motivation to increase preparation effort in order to compensate for perceived weaknesses. Eysenck and Calvo further suggested that when compensatory resources, such as the ability to take more time to perform a task, are unavailable, performance suffers and ultimately leads to personal inefficacy. Calvo and Carreiras suggested that students with high anxiety have an increased awareness of intrusive thoughts that may reduce working-memory capacity used for task processing and ultimately weaken their performance. Calvo and Carreiras’ research indicated that for teaching situations that require complex cognitive function (difficult tasks), highly anxious learners who fear performance judgment may have impeded cognitive processes. This position was supported by a number of researchers who further suggested that during these evaluative situations, high test anxiety learners fail to concentrate on relevant tasks, but instead focus on negative thought processes and their heightened emotional arousal (Mandler & Watson, 1966; Meichenbaum, 1972; Saranson, 1975; Wine, 1971). Further work by Mandler and Watson (1966) has indicated that during evaluative situations, high test anxiety individuals are preoccupied with any combination of thought processes, including worrying about their performance, anticipation
of corrective action or loss of esteem, as well as increased awareness of personal physiologic responses. This process interferes with performance and negatively impacts performance measures. Although there are a number of opinions regarding anxiety’s potential andrological effects, we should also consider the potential long-term effects anxiety may have outside the academic setting.

**Long-term effects of anxiety.** To explain the long-term effects of anxiety on a learner, I will briefly review applicable work by Albert Bandura (1999). Although initially from the school of behaviorists, Bandura separated himself from the behavior theorists with the introduction of social cognitive theory (Bandura, 1986). In the context of this theory, Bandura describes people as, “…self-organizing, proactive, self-reflecting, and self-regulating, not just reactive organisms shaped and shepherded by external events [as in the behaviorist approach]” (Bandura, 1999, p. 154). More importantly, Bandura gave importance to the variety of cognitive processes that may affect a person’s behavior. Bandura suggested that while behaviors are a function of their consequences, they can be changed through cognitive mediation and social learning. Bandura considered three primary components that influence a person’s ability to adapt to a scenario or event. This opinion seems to be the basis of the concept of triadic reciprocal causation. Triadic reciprocal causation is a model that describes the relationship between three elements: one’s personality, behavior, and environment (Bandura, 1999). As depicted in Figure 2, Bandura presented the relationship between these elements that determine triadic reciprocal causation.

The personality element (P) includes such internal responses as cognitive (thoughts), affective (emotion), and biologic (physiology) responses, while the behavior element (B)
Figure 2. Bandura’s model of triadic reciprocal causation

represents the physical actions of an individual, and environmental element (E) represents environmental or situational events. All of these elements interact and influence each other bidirectionally. Although the figure may suggest that these elements’ influences are of equal force, they are not, and the influences of the elements on each other vary across different situations.

Further, Bandura proposed that although a unified theory, the concept of triadic reciprocal causation is much more complex in its foundation and is strongly rooted in sociostructural and psychological theories, such that “…the sociostructural and personal determinants are treated as interacting cofactors within a unified causal structure” (Bandura, 1999, p. 7). Many contemporary cognitive-behavioral theorists have applied similar schema in order to explain anxiety effects on procrastination. Some researchers have suggested that procrastination is rooted in two cognitive processes—irrational fears and self-criticism (Ellis & Knau, 1977; Solomon & Rothblum, 1984), while another has suggested that there is a correlation between anxiety and procrastination in the academic setting (Lay et al., as cited in Ferrari, Johnson, & McCown, 1995). Similar to Eysenck and his colleagues (2007) attentional control theory discussed earlier, these studies suggested that worrisome thoughts
and associated anxiety continue to tax an anxious person’s self-regulating mechanism and associated auxiliary processing used to inhibit these negative thoughts and, in anxious individuals, leads to task avoidance. Thusly, a vicious cycle is formed. Worrisome thoughts decrease an anxious individual’s learning efficacy and efficiency during threatening stimuli (the anxiety evoking learning scenario). If not addressed, the same worrisome thoughts lead to procrastination or complete task avoidance. Given the potential for long-term negative effects, learner anxiety can and should be addressed.

**Methods used to alleviate or minimize learner anxiety.** There are a variety of strategies used to alleviate or prevent anxiety, particularly for the adult learner. The techniques vary and are typically dependent on the either the behavioral or cognitive theoretical foundations. The largest family of techniques dependent on behaviorist research is dually founded on coping techniques that were introduced in 1971 (Goldfried, 1971; Suinn & Richardson, 1971). There are a number of techniques used to minimize learner anxiety that arose from Goldfried’s, and Suinn and Richardson’s research. For example, *applied relaxation* and *behavior intervention* techniques are two common strategies used in learning situations. Applied relaxation is a learned coping mechanism that allows the subject to recognize early signals of anxiety and learn to cope with anxiety by focusing on the physiological reactions encountered in a phobic situation (Chang-Liang & Denney, 1976). The family of behavior intervention techniques encompasses a broader range of strategies and applications. In the learning environment, strategies focus on minimizing a learner’s physiologic response to an anxiety-evoking situation, and are comprised of relaxation training exercises that may include muscle relaxation and deep breathing. One such technique was called the cue-controlled relaxation response technique by Russel and Sipich (1973). This technique focuses on controlled breathing with a purpose to create a
conditioning between self-instruction by the learner and the state of being relaxed. Like the majority of techniques rooted in behaviorist theory, this technique requires a number of weeks’ conditioning and rehearsal outside the anxiety-evoking environment in order to gain benefit in practice.

The most effective techniques and procedures used to alleviate anxiety are based on the theoretical foundation of cognitive behavior therapy (Rachman, 1993). Cognitive behavior therapy, introduced in 1985 by Aaron Beck, is sometimes called cognitive theory, and is based on the rational emotive behavior theory (REMBT) of Albert Ellis (Ellis, 1955). The contemporary approach based on cognitive behaviorism applies learning theory to mental events like thoughts and feelings. Currently termed cognitive therapy (CT), treatment for an anxious learner may help the learner identify how anxiety-inducing situations may affect their personal thoughts (cognition), emotions, physiologic responses, and behavior (Beck & Emery, 1985). The mechanism of alleviated or minimizing anxiety in cognitive therapy is such that the anxious person is made aware of their debilitating thought toward an anxiety-evoking stimulus and replace it with a more helpful and realistic one in actual situations. This approach using cognitive theory is widely considered most effective for subjects in a panic state and is successfully used in anxiety-evoking learning scenarios (Rachman, 1993). There are a number of techniques and procedures used in cognitive therapy in learning scenarios. They can be modified, based on the learner’s personality and needs. For example, one technique called coaching involves maintaining a positive association between the instructor and learner to help minimize a learner’s anxiety during an anxiety-evoking situation (Bandler & Grinder, 1986). Coaching is considered quite effective at minimizing anxiety in evaluative situations, as the instructor guides the learner through an anxiety-evoking situation, taking cues from the learner’s individual language or behavioral
norms (Reece, 1993). Similarly, another method, called modification of social interaction patterns is considered effective with learners that have low self-esteem or have high “trait-anxiety” and tend to be more anxious in general (Sieber et al., 1977). According to Sieber and her colleagues, modification of social interaction patterns may include nurturing and rewarding interactions where the instructor gives positive verbal reinforcement to the learner. Another tactic used to reduce learner anxiety is called redefinition of task situations (Saranson, 1975). The purpose of redefinition of task situations is to devise instruction that motivates the learner with low anxiety, yet redirects the attention of the high-anxiety learner through the emphasis of the intrinsic aspects of the task performance while minimizing the nature of the task (Sieber et al., 1977). In comparison to the aforementioned techniques founded on behaviorist theory, most of the techniques described here can be readily employed during an anxiety-evoking situation. All of these tactics are implemented to supplement and strengthen learners’ attentional learning and required cognitive processes in order help them overcome their fear of failure and improve performance by creating learning conditions where the learners are more likely to succeed. In learning situations, modification of a curriculum may be needed in order to reinforce attentional learning and performance skills. To the educator, this may mean offering opportunities to learners so that they practice or rehearse a skill until they overlearn the skills that are required. Any of these tactics reviewed may be used in any combination, depending on the specific needs of the learner and the nature of the situation. It makes sense that these tactics used to minimize or alleviate student anxiety during stressful learning events may be applied in medical education. It order to fully understand the need for addressing anxiety in medical education, it might be beneficial to review the past and current trends in treatment of anxiety in medical education.
Traditional Role of Anxiety in Medical Education

Student anxiety seems to have been propagated by the long-established teaching methods in medicine. Traditionally, medical education has been steeped in a culture that has governed teaching trends that have promoted such practices as the extended workweek and the “learning by doing” approach. Established teaching practices were seemingly tests of endurance, will, and confidence. Some researchers (Aktekin et al., 2001; Becker et al., 1961; Haas & Shaffir, 1987; Moss & McManus, 1992; Pitkälä & Mäntyranta, 2004; Shuval, 1967) suggested this is particularly true during medical students’ introduction to the intimate examinations where student anxiety has been considered a normal element of initiation to becoming a practicing physician. Anxiety was acknowledged, but considered lightly as a nuisance that dissipated as students gained experience (Becker et al., 1985; Haas & Shaffir, 1987; Shuval, 1967). In recent years, researchers have acknowledged the negative impact of anxiety on medical students’ learning (de Saintonge & Dunn, 2001; Linn & Zeppa, 1984; Stewart et al., 1999). In particular, a body of research (Huebner et al., 1992; Mavis, 2001; Sarikayaya et al., 2006) has suggested that the mere act of performing physical examinations creates a state of anxiety that may adversely affect learning and performance for medical students. Given this evidence, it makes sense that the traditional medical learning scenario described above would be most anxiety evoking for a medical student who is learning to perform an intimate exam.

Medical student anxiety toward intimate examinations. Learning the skills associated with patient examinations is naturally anxiety-evoking as students are introduced to new clinical experiences and patient interactions. A student’s anxiety is amplified while they learn and perform intimate examinations: examinations of clients’ breast, pelvic, prostate, and genitourinary systems. Many researchers (Abraham, 1996; Behrens et al.,
1979; Campbell et al., 1994; Ker, 2003; Popadiuk et al., 2002; Rochelson et al., 1985) supported the view that this type of examination is associated with the highest amount of patient contact and level of intimacy and continues to be a source of apprehension for most medical students.

**Common teaching practices of intimate examination skills.** A review of common intimate examination teaching practices may help understand the potential sources of medical students’ anxiety. Although most studies focus on the female pelvic exam, having a clearer understanding of current teaching methods of the female pelvic exam may identify some shared practice with the male GUR exam that may cause student anxiety.

There are five commonly implemented teaching practices for the female pelvic exam, ranging from clinical teaching opportunities to formal teaching sessions. The practices may be performed in isolated incidents, as single events, or combined to create a more comprehensive curriculum. These are:

1. Observation alone or student may be helped to acquire motor skills by examination of the anaesthetized or unconscious nonresponsive woman (Abraham, 1995).

2. Examination of a woman (who is a consenting patient) under general anesthesia, or examination of a woman (who is a consenting patient) at outpatient clinics. Both practices are under supervision by a doctor (Abraham, 1995).

3. Examination of surrogate patients (who are paid volunteers), supervised by a doctor. (Abraham, 1995; Godkins, Duffy, Greenwood & Stanhope, 1974; Perlmutter & Friedman, 1974).

4. Small group learning with one gynecological associate acting as teacher and patient, and student acting as patient. The gynecological associate acts as both
teacher and patient and a female student also acts as patient. (Abraham, 1998; Kleinman, Hage, Hoole, & Kowlowitz, 1996), and

5. Small-group learning consisting of two or three students and one or two trained women assistants, “gynecological associates” (Abraham, 1995) or “standardized patients” (Pugh & Salud, 2007; Pugh, Heinrichs, Parvati, Srivastava, & Krummel 2001). Both assistants act alternately as teacher or patient. The process includes an initial discussion about feelings, sensitivity, expectations, a pelvic-exam demonstration, guided instruction where the student examines one of the associates, and a feedback discussion with both associates completes the process (Abraham, 1995; Holzman, Bridgham, & Easterling, 1980; Kleinman et al., 1996). Each of these practices have unique contributing variables that may evoke anxiety that stems from a broad range of sources. Although reviewed in greater detail later in this paper, according to a sample of researchers, the sources of anxiety may include fear of hurting the patient, appearing inept, repulsion, and embarrassment. Because of ethical sensitivity of performing intimate examinations on patients (Coldicott, Pope, & Roberts, 2003), contemporary educators have mostly endorsed the latter option, as Vontver et al. (1980) posited, “The use of trained professional patients has been advocated as a means of reducing this anxiety while improving student performance” (p. 778). In recent years, the practice of utilizing standardized patients and patient instructors has become common practice for teaching all forms of the intimate exams. Interviews with second-year University B medical students supported this view, as 98% of the interviewed students stated their preference for a standardized patient over any other methodology, including actual patient practice, for
learning the male GUR examination skills. Additionally, students have reported role playing in the clinical scenario with a patient instructor as being very helpful in teaching communication skills and pointing out skills deficiencies (J. Tresley, personal communication, February 6, 2007).

In many curricula, the clinical experience with the standardized patient has been supplemented with other teaching resources and activities. For example, lecture materials (Rakestraw et al., 1983; Vontver et al., 1980), reading materials (Rakestraw et al., 1983), cadaveric prosections (Kamemoto, Kane, & Frattarelli, 2003), videotaped demonstration (Kamemoto et al., 2003; Pugh & Salud, 2007; Pugh, Obadina, & Aidoo, 2009; Rakestraw et al., 1983), inanimate models (Pugh & Salud; 2007; Pugh et al., 2009; Rakestraw et al., 1983; Vontver et al., 1980) and additional learning stations (Abraham, 1998; Pugh & Salud; 2007; Pugh et al., 2009) all have been utilized to supplement students’ learning experiences while they learn and perform the pelvic and breast examinations. Educators designed these additional resources to minimize student anxiety by increasing the students’ familiarity with the anatomy and the tasks involved with performing a particular exam. Most research relating to measuring student anxiety toward intimate examinations has focused on measuring anxiety effects of implementing these practices and materials.

**Research on medical student anxiety during intimate examinations.** Although there exists more than 40 years of research that has suggested that the presence of student anxiety while they learn and perform intimate examinations is normal and expected, only a handful of studies have attempted to identify the specific characteristics of anxiety while students learn and perform these examinations, and most have focused on the pelvic examination. As early as 1979, in an exploration of student anxieties of the pelvic examination, Buchwald (1979) identified six primary student anxieties: (a) fear of hurting
the patient, (b) being judged inept, (c) inability to recognize pathology, (d) sexual arousal, (e) finding examination unpleasant, and (f) disturbance of doctor-patient relationship. More recently, Kamemoto et al. (2003) have supported Buchwald’s findings, indicating medical students’ primary concern is that of inadvertently causing physical or emotional pain, as well as showing disrespect of the patient during the pelvic examination.

Pugh et al. (2009) also found that for the pelvic examination, the top rated student anxieties are (a) causing harm, (b) the intimate nature of the exam, and (c) general performance anxiety. They also identified the leading student anxieties for the breast exam. These are (a) fear of missing a lesion, (b) the intimate nature of the exam, and (c) general performance anxiety (Pugh et al., 2009). More notable, the same study also identified the primary concerns for students relating to the male GUR exam as (a) intimate nature of the exam, (b) causing harm, and (c) embarrassing the patient. This seems to be consistent with earlier works of Vontver et al. (1980), who suggested that “students are anxious to perform the examination in a professional manner, yet are unsure of what to look for, what to palpate, how to perform the technical maneuvers, and what to say to the patient in the process” (p. 778). These studies suggest that medical educators may have noticed the importance of the identification of student anxiety present during all intimate examinations: the pelvic, breast exams, and the male GUR exam. Although these studies have been instrumental in highlighting the characteristics of medical students’ anxiety associated with learning and performing intimate examination skills, few studies have attempted to identify specific characteristics of medical students’ anxiety toward the male GUR exam. Nieman, Kelliher, Sachdeva, and Cohen (1994) supported the existence of student anxiety toward the male GUR examination. Two other studies (Hennigan et al., 1991; Lawrentschuk & Bolton, 2004) attempted to measure students’ confidence at performing GUR-related tasks, while
two others attempted to capture students’ attitudes toward the use of standardized patients while learning to perform the male GUR exam (Howley & Dickerson, 2003; Robins et al., 1997).

**Difficulties of measuring medical student anxiety during GUR examination.** The lack of documented research on the identification of the sources and characteristics of student anxiety while they learn and perform the male GUR exam may not necessarily reflect a lack of interest. It seems as though educators may view student anxiety as more of a nuisance than a real concern. Another explanation is that medical educators may be inferring that the sources and characteristics of student anxiety for the male GUR exam mirror those of the other intimate patient examinations. There is no documented research that either supports or refutes this position. A recognizable danger of this philosophy may be that the interventions that are meant to alleviate or minimize student anxiety during one intimate examination may not be applicable to another intimate examination. Furthermore, measuring students’ anxieties toward each specific intimate examination must take into consideration not only the anxieties evoked by the training methodology, but the specific task being taught.

While medical educators have been able to minimize some negative effects of student anxiety while students learn and perform intimate examinations (Abraham, Chapman, Taylor, McBride, & Boyd, 2003; Pugh & Salud, 2007; Pugh et al., 2009), the complexity of the origins of anxiety has made full suppression impossible. It appears that student anxiety associated particularly with learning intimate examination skills involves an assortment of influences. According to Shors (2006), “The variability in the types of responses is attributable to organismal properties, such as age, sex differences, and species” (p. 56). Additionally, Shors suggested that students’ responses may be dependent on the
types of stressful event that occurs, and on the type of learning that is assessed. It becomes clear that due to the complexity of anxiety, there is no consistent or simple relationship between stress and learning. Clearly, in order to improve the teaching programs for all of the intimate examinations, each form of intimate examination must be investigated and the researchers need to identify specific sources of student anxiety for each type of intimate exam. With clear identification of students’ anxieties toward the male GUR exam, educators may have a better understanding of the precise sources of anxiety during the exam, and make improvements to the associated curriculum that address and minimize student concerns early in their training.

**Summaries of Literature Review Explaining the Relationship of Medical Student Anxiety and Learning**

**Anxiety and Learning**

Most researchers have attempted to identify the variety of factors that may influence a medical student’s level of anxiety. The most noted reasons include fear of appearing inept, discomfort and/or unfamiliarity with genitalia (Abraham, 1995, 1996), influences of learner gender (de Saintonge & Dunn, 2001; Dods & Treppa, 1978; Gledhill & Van der Merwe, 1989; Greenfield, Parle & Holder, 2001; Haist, Witze, Quinlivan, Murphy-Spencer, & Wilson, 2003; Hayes et al., 2004; Kwolek, Blue, Griffith, Wilson, & Haist, 1998; Toews et al., 1997), learner inexperience and age (Abraham, 1998; Aktekin et al., 2001), and finally, learner anxiety personality (Dods & Treppa; Firth, 1986; Henning, Ey, & Shaw, 1998). Quite often, these theories are competing and/or contradictory. Although there are a number of factors that may influence students’ anxiety toward the intimate exams, the focus of this study concentrates the literature review on the anxiety-evoking sources associated with the male GUR exam and their relationship with gender and anxiety personality.
**Relationship of gender and medical students’ anxiety.** A literature review of the existing studies that identify the possible relationship of gender and medical students’ anxiety revealed conflicting information. Two studies suggested that there were no gender differences in medical student anxiety in general (Henning et al., 1998; Stewert et al., 1999). Henning et al. further suggested that anxiety differences related more to personality than gender. The authors hypothesized that anxiety is most associated with perfectionism, labeled “imposter phenomenon” in medicine, which occurs when “high achieving individuals chronically question their abilities and fear them to be intellectual frauds” (Henning et al., 1998, p. 456).

Abraham et al.’s findings (2003) supported this position, and significant gender effects were not identified for medical student anxiety while they performed their first female pelvic exam. Contrary to these results, a number of studies spanning 40 years (Davidson, 1977; DePablo, Subriná, Martín, DeFlores, & Valdes, 1965; Dods & Treppa, 1978; Gledhill & Van Der Merwe, 1989; Pugh & Salud, 2007; Pugh et al., 2009; Saul & Kass, 1968; Walton, 1968; Whittle & Murdoch-Eaton, 2001) have identified that female medical students consistently reported their anxiety higher than their male peers.

Only one study has examined the relationship of gender and medical student anxiety while they learn and perform the male GUR examination (Howley & Dickerson, 2003). Findings supported that female medical students rated their state anxiety higher than their male cohorts. Statistical differences were found using a standardized measure, the State Trait Anxiety Inventory (STAI; Spielberger, Gorsuch & Lushene, 1970). These results seem to be the most credible, and further support other researchers’ hypotheses that female medical students report their anxiety toward performing other intimate examinations consistently higher than their male peers.
There is a variety of hypotheses that may explain why female medical students consistently rate their anxiety higher and confidence lower, than their male cohorts. Some researchers (Dods & Treppa, 1978; Hojat, Glaser, Xu, Velosky, & Christian, 1999) suggested that adult women are simply more likely to self-report anxiety than men. The National Institute of Mental Health (2006) reported that adult women are more susceptible than men to panic disorders, posttraumatic stress disorder, phobias, and generalized anxiety disorders. Some researchers have suggested that female medical students are more likely to self-report a negative affect, such as worry or anxiety, than their male counterparts (Dyrbye, Thomas, & Shanafelt, 2006). The influential factors are varied and may include social norms, social position, and hormonal influences (Jean-Murat, 2008). Although these factors should be considered when measuring students’ negative affect, these factors are outside the scope of this study.

Theories applied to performing the male genitourinary rectal examination. To understand the application of the state–trait anxiety theory, attentional control theory, and the relationship of gender and anxiety in practice, we can apply the language associated with the theories discussed earlier, and deconstruct a typical learning scenario of medical students as they learn to perform the male genitourinary exam on a standardized patient. To begin, the student, along with three or four peers, enters the examination room where the “patient,” a trained actor called a standardized patient, is waiting. During the introduction to the “patient,” a student’s state anxiety may begin to develop in anticipation of the task at hand. An observer may see symptoms of increased state anxiety as the student begins to demonstrate signs of physiological fight or flight responses, phenomena that may include increased respiration or a flushed face. Figure 3 shows the cognitive process of Eysenck’s model of attentional control during this process. First, invasive, worrisome, thoughts enter
the student’s mind, and attentional control shifts toward the threatening thought, “I hope I don’t hurt the patient.” The invasion of worrisome thoughts may be more prevalent in a female student, depending on her anxiety personality, while male students may be more concerned with fear of negative evaluation. In either case, the student’s other central executive function, inhibition, may be activated, as the student resists interference from the task-irrelevant stimuli (worrisome thoughts). The student may exert cognitive efforts in an internalized dialogue of “I can do this,” or by redirecting their efforts to a review of anatomy or clinical relevance.

During this time, the student’s working memory continues to be updated with new representations. Expending cognitive efforts on worrisome thoughts toward a task that also require cognitive effort results in decreased efficiency of the cognitive processing. Further, decreased efficiency may result in decreased performance efficacy that ultimately translates as a student’s decreased function of working memory. Some information fails to be incorporated into the working memory, and ultimately, never reaches long-term memory storages.

Not surprisingly, anecdotal evidence from interviews with medical students has suggested that many second-year medical students fail to remember what they had learned the previous year, even when reviewed under similar circumstances (J. D’Achille personal communication, November 30, 2006). Medical educators of nearly every medical curriculum are likely to observe this cognitive event every year during clinical-skills education. The effects of anxiety not only decrease efficiency of the students’ cognitive performance, but ultimately the educational program’s curriculum efficacy as well. In order to address and minimize medical student anxiety while they learn and perform the male GUR exam, a founding theory must be developed. This theory will become the basis for an
item structure that will effectively measure the sources and characteristics of medical student anxiety toward the male GUR examination.

Figure 3. Cognitive process of Eysenck’s model of attentional control.

Application of Theories to Build the Construct

In order to get the most informative representation of medical students’ anxiety toward the male GUR examination, it is important to recognize the variety of sources that may contribute to a learner’s anxiety while they learn and perform it. There are two primary sources of learner anxiety that will be examined in this study. The first source is the tasks that the students perform during a male GUR exam. For some, the mere review of the tasks may evoke a variety of anxiety phenomena. Appendix A offers a detailed description of each of the five tasks involved in performing the male GUR examination. The second source of
student anxiety is a combination of the personal elements that may influence a learner’s anxiety toward the specific tasks of the male GUR exam. The average student maintains a highly complex fusion of psychosocial resources, sexual tendencies, and memories from a variety of experiences. It is reasonable that this blend of the learner’s tendencies, experiences, gender, and personality will influence the anxiety level while learning to perform the examination.

The vast complexity of personal elements that may influence a learner’s anxiety during the GUR exam decreases the feasibility to isolate and measure all personal elements that a medical student may carry to their experience of performing the male GUR examination. Focusing the study to measure three personal elements: gender, ethnicity, and anxiety personality may result in a more robust measure. The result is the intersection of all sources of anxiety measured—the tasks that may evoke student anxiety, the personality and gender of the student that may influence the anxiety level, and the resulting intersection—measurable aspects, or items that quantify a student’s anxiety level. Figure 4 identifies the relationship of the sources of anxiety examined in this study.

Validity Evidence

Providing sufficient validity evidence for inferences made based on data obtained from analysis of the new instrument that measures student anxiety toward the male GUR examination is imperative. The contemporary presentation of validity evidence by Wolfe and Smith (2007a, 2007b) was founded on the unified concept of validity from the Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999) and the terminology and classification system proposed by Messick (1989, 1995). Wolfe and Smith combine established concepts with additional types of validity evidence described by the Medical Outcomes Trust (Medical Outcomes Trust Scientific Advisory Committee,
1995) to form a well-defined framework that identifies validity evidence produced using Rasch measurement models. The evidence will provide support to the results from the instrument and credible inferences about the characteristics of the student anxieties to medical educators and The Liaison Committee on Medical Education—the accrediting authority for medical-education programs.

![Diagram](image)

**Figure 4.** The relationship of sources of medical student anxiety toward the male GUR examination

Messick (1989) defined validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). The traditional classification of validity included three sources, or types of validity, including criterion-related, content, and construct validity. Messick believed that construct
validity alone should be the basis of validity evidence and test score interpretation. He argued that construct validity comprised adequate evidence and rationales supporting the trustworthiness of score interpretation. Although Messick’s framework was unitary, he defined six sources, or aspects of construct validity, all of which could be assessed to support or refute interpretation of test scores. These were (a) content, (b) substantive, (c) structural, (d) generalizability, (e) external, and (f) consequential validity (Messick, 1995).

Messick’s (1989) original conceptualization of validity became the foundation for the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), in which validity is described as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). Like Messick, the Standards for Educational and Psychological Testing (Standards) promoted that the process of validation involves “accumulating evidence to provide a sound scientific basis for the proposed score interpretations” (p. 9). Validity refers to the quality of the inferences made from test scores, and a sound validity argument includes multiple sources of validity evidence (AERA, APA, & NCME, 1999). Based on Messick’s principal of construct validity, the Standards identified construct validity as the primary body of validity evidence that housed other types of validity (AERA, APA, & NCME, 1999).

Types of validity evidence, rather than distinct types of validity, are used to support the intended interpretation of test scores. To clarify this point, a test cannot be deemed valid, but validity evidence can support or refute the inferences made by the test scores under specific conditions. The Standards (AERA, APA, & NCME, 1999) described five different types of validity evidence that are used to develop a validity argument for the interpretation of test scores or to contribute to construct validity. They included evidence from (a) test
content, (b) responses processes, (c) internal structure, (d) relationships to other variables, and (e) consequences of testing.

Defining Validity Using the Standards for Educational and Psychological Testing

Evidence based on test content. The Standards described test content or the content aspect or the content aspect of validity as “the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring” (AERA, APA, & NCME, 1999, p. 11). Evidence based on test content suggests that items represent the intended content domains. For the purpose of this study, two aspects of test content will be considered. These are items, or factors that may influence a medical student’s anxiety, and potentially anxiety evoking aspects associated with the male GUR examination. Test content and scale development will be overseen by an expert panel in order to ensure the items of the test reflect the intended content domain and justify eventual interpretations made from test scores.

Evidence based on response processes. Evidence based on response processes, or the substantive aspect of validity, refers primarily to individual responses and person fit. Person fit describes the fit between the performances or responses of the individual test or questionnaire takers and the construct being measured (Wolfe & Smith, 2007b). For the purpose of this study, validity evidence for inferences made will be based on analysis of individual responses from the examinees, the medical students.

Evidence based on internal structure. The Standards described the internal structure of a test, or structural aspect of validity, as “the degree to which the relationships among test items and test components conform to the construct on which the proposed test
score interpretations are based” (AERA, APA, & NCME, 1999, p. 13). Evidence based on the internal structure of the test will determine the extent to which it supports the intended construct. Researchers may produce validity evidence using a variety of psychometric analyses, including item analysis, rating scale analysis, reliability, dimensionality, and differential item functioning. This study will utilize most of these analysis methods in order to test internal structure evidence.

**Evidence based on relationship to other variables.** The Standards (AERA, APA, & NCME, 1999) based the external aspect of validity on the relationship with other variables, using four distinct types of validity evidence, (a) convergent, (b) discriminant, (c) test criterion, and (d) validity generalization. *Convergent* evidence indicates that the scores and measures of one test have a direct (positive) relationship with other scores and other measures that have been supported by sufficient validity evidence intended to assess the same (or similar) constructs. Contrary to this evidence, *discriminant* evidence indicates that the scores and measures of one test have an indirect (negative) relationship with other scores and other measures intended to assess different constructs.

**Evidence based on consequences of testing.** Validity evidence based on consequences of testing, or the consequential aspect of validity, refers to providing evidence based on the evaluation of any short- or long-term consequences of score interpretation and use. According to Messick (1994, 1995), the primary concern of adverse consequences of testing is that any negative impact on individuals or groups should not be derived from any source that may result in invalidity of the test. Additionally, potential adverse consequences must be weighed against potential benefits (Kane, 2001; Messick, 1989).
An Alternative Framework for Examining Validity Evidence

The alternative validity framework is Wolfe and Smith’s (2007a) product of supplementing the validation framework from Messick (1989) with added criteria submitted by the Medical Outcomes Trust (1995). The result was a framework that proposed seven types of validity evidence: (a) content, (b) substantive, (c) structural, (d) generalizability, (e) external, (f) consequential, and (g) interpretability (Wolfe & Smith, 2007b). Because evidence presented by Wolfe and Smith was based on the Rasch measurement model, I will briefly describe classical test theory, the Rasch measurement model, and discuss the latter’s strengths over the former. Finally, I will summarize the first five types of validity evidence that may be provided by the Rasch model and addressed by this study.

Classical Test Theory

Also known as true score theory, classical test theory is a psychometric theory that has been widely accepted in medical education. This theory suggests that the observed score, X, is equal to the sum of the true score, T, the averaged score of a subject for repeated testing, and the random error, E. This relationship is expressed mathematically as

\[ X = T + E. \] (Equation 1)

Simply put, the equation states that the true score is equal to the observed scores plus or minus the error terms (Novick, 1966). Traditionally, many researchers and educators alike have relied on an estimated reliability coefficient, most often Cronbach alpha, based on this theory.

Reliability is necessary for the validation of measures, because it demonstrates that for a particular group of subjects the scores are stable, giving evidence that the subjects’
measured behavior is stable. Insufficient reliability indicates resulting scores may contain a large component of random error, prohibiting researchers from making any meaningful interpretations, with clear consequences to validity.

The *Standards* (AERA, APA, & NCME, 1999) also recommends reporting the standard error of measurement (SEM). The SEM is generally reported in conjunction with the reliability coefficients discussed earlier, and estimates within-person inconsistency by calculating a reliability coefficient using the standard deviation (S) for a hypothetical set of repeated measurements on a single individual for that test, and the reliability coefficient for the test, r. This calculation is shown by the equation:

\[ SEM = S \sqrt{1 - r}. \]  

Although very popular, the use of these coefficients and the application of classical test theory to raw scores with parametric statistics are associated with a number of limitations, including

1. Difficulty generalizing scores to other samples due to sample dependency of person and item measures;
2. Difficulty predicting the interactions between a person and an item, given that persons and item scores are on different scales (Smith, Conrad, Chang, & Piazza, 2002);
3. Difficulty directly comparing measures of persons and items without complete responses;
4. Only a single SEM that applies to all scores in a particular population is typically provided;
5. Lack of methods to investigate aberrant response patterns (Smith et al., 2002);

6. Potential for invalid inferences made from the analysis when using parametric statistical analyses which assumes continuous scale (Merbitz, Morris, & Grip, 1989); and

7. Potential for inaccurate estimation of the internal consistency reliability of test scores (Embretson & Reise, 2000).

The Rasch Measurement Models

Rasch (1960) developed a measurement model that is capable of fully separating person and item parameters from each other, while predicting what will happen when a person encounters a specific item. The dichotomous model is written as

\[
P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)},
\]

and contains two parameters, the ability of the person \( n \) (\( \theta_n \)), and the difficulty of the item \( i \) (\( \delta_i \)), with \( P_{ni} \) being the probability of person \( n \) correctly responding to item \( i \) (Rasch, 1960). In full, the model states that the probability of person \( n \) correctly answering item \( i \) is a function of \((\theta_n - \delta_i)\). According to the Rasch model, the probability that person \( n \) will answer item \( i \) correctly is determined by the ability of that person \((\theta_n)\) and the difficulty of that item \((\delta_i)\). To explain the model in a practical application, if a person’s ability is greater than the difficulty of an item, that person will have more than .50 probability of answering it correctly. On the other hand, if the person’s ability is lower than the difficulty of an item, that person will have less than .50 probability of answering it correctly. Finally, if an item’s
difficulty perfectly matches a person’s ability, that person will have a .50 probability of answering it correctly.

There are benefits of the Rasch (1960) measurement model. First, assuming the data fit the model, the parameter estimates are neither sample nor test dependent. Because the person and item estimates are free from the sampling distribution of incidental parameters, the Rasch model can use observed item responses to estimate the ability of the examinees regardless of the difficulty distribution of the particular items administered. Concurrently, the model is able to estimate the difficulty of the items regardless of the ability distribution of the examinees who happen to have taken the items. Second, the estimates of a person’s ability and item difficulty are reported on the same interval scale, called the “logit” scale.

Other benefits of the Rasch model is that the Rasch model directly addresses the first two limitations of classical test theory and fulfills the requirements for a measurement model for construct modeling as defined by Wilson (2004, 2005). Wilson described construct modeling as a framework for developing an instrument that is comprised of four building blocks: (a) a construct map, (b) an items design, (c) an outcome space, and (d) a measurement model. A primary requirement of the measurement model mandates that the locations of persons and items on the measurement scale be independent of the sets of items or persons researchers will take into the consideration. Even in its most basic form, the Rasch model seems to surpass classical test theory for the construct modeling proposed in this study.

For the purpose of this study, which uses polytomous data, another form of the Rasch (1960) model may be appropriate. The two most widely used models for polytomous data are the rating scale and the partial credit models. The rating scale model was designed to accommodate more than two response categories per item (Andrich, 1978a, 1978b;
The model is expressed mathematically as

\[
P_{nik} = \frac{\exp\left(\sum_{k=0}^{k} (\theta_n - (\delta_i + \tau_s))\right)}{\sum_{t=0}^{m_i} \exp\left(\sum_{k=0}^{t} (\theta_n - (\delta_i + \tau_s))\right)}.
\] (4)

In this equation, the probability that person \(n\) will respond with \(k\) (the response category) to item \(i\) is \(P_{nik}\). Further, the variable \(k\) represents the response category for a rating scale that has \((m_i + 1)\) rating categories, \(k \in \{0, \ldots, m_i\}\), while \(\theta_n\) is the ability level of person \(n\). The variable \(\delta_i\), the difficulty of item \(i\), and \(\tau_s\) is the difficulty of the \(k^{th}\) threshold. A threshold parameter \((\tau_s)\) represents the relative difficulty between rating categories. The primary limitation of this model is that it uses a single set of thresholds across all items. Therefore, the rating-scale model cannot be used when the items of a scale have variable rating categories. When using rating scales that have different numbers of rating categories, different labels for the rating categories, or when the assumptions with a common set of thresholds are not appropriate, the partial credit model is used (Embretson & Reise, 2000; Masters & Wright, 1996). The partial credit model is mathematically expressed as

\[
P_{nik} = \frac{\exp\left(\sum_{k=0}^{k} (\theta_n - \delta_{is})\right)}{\sum_{t=0}^{m_i} \exp\left(\sum_{k=0}^{t} (\theta_n - \delta_{is})\right)}.
\] (5)
Unlike the rating scale model, the partial credit model further defines the threshold for each of the items \( (\delta_i) \). This allows the rating scale for each item to have a distinct threshold structure.

Regardless of the specific model employed, the family of Rasch (1960) models offers advantages over the classical test theory for this study, and may help supply validity evidence that supports the inferences made from the measures of the new scale.

**Evidence relevant to the content aspect of validity.** Like the *Standards* (AERA, APA & NCME, 1999), Wolfe and Smith (2007b) discussed evidence relevant to the content aspect of validity. According to Wolfe and Smith, the content aspect relates directly to the “relevance and representativeness” of the test content (p. 245). For example, documented history of item development and examination of the technical quality of items is essential to evidence of validity relevant to content. Wolfe and Smith suggested that the technical quality of the items can be evaluated based on item-measure correlations and the item mean-square fit indices in the Rasch framework. The item-measure correlation, also called the point-measure correlation, provides a Pearson correlation between the vector of scores on an item and the ability estimates of the respondents. The correlation identifies the degree in which the scores on an item are consistent with the averaged scores of the remaining items. A positive point-measure correlation is ideal and indicates that the particular item contributed useful information to the construct measured by the test as a whole. In a practical description for this application, a positive point measure for a group of respondents indicates that the individuals with high-anxiety measures scored appropriately high on the item, while individuals with lower anxiety measures scored less anxiety on the item.

The mean-square item fit indices can also be utilized to offer validity evidence relevant to the content aspect. The unweighted mean-squared item fit statistic can help
identify problematic measurement conditions, such as multidimensionality and poorly written items, by indicating discrepancies between observed scores from expected values. The discrepancy is captured by the mean of the squared standardized residuals of the observed ratings from their expected value (Wright & Masters, 1982). Resulting indices are often transformed or standardized. These values are termed *standardized unweighted mean-squared fit indices*, and range from $-\infty$ to $+\infty$ (Wolfe & Smith, 2007b). A negative unweighted mean-squared fit index indicates less variability than the Rasch model predicted, while a positive value suggests more variability than expected (Adams, Wilson, & Wang, 1997). Smith, Schumacker, and Bush (1998) suggested that cutoff values for problematic items are $\pm 2.00$, $\pm 3.00$, and $\pm 4.00$ for small (less than 1,000), large (3,000), and very large (over 5,000) samples, respectively.

**Evidence relevant to the substantive aspect of validity.** Wolfe and Smith (2007b) describe the substantive aspect of validity in much the same way the *Standards* (AERA, APA & NCME, 1999) describes evidence related to response processes. They stated that “the substantive aspect of validity addresses the degree to which theoretical rationales relating to both item content and cognitive processing models adequately explain the observed consistencies among item responses” (Wolfe & Smith, 2007b, p. 248). Wolfe and Smith suggested that additional evidence of validity related to the response processes of examinees include analysis of rating scale function, analysis of Rasch person-fit statistics, and confirmation of the theoretical item hierarchy.

Describing the aforementioned indices relevant to the content and substantive aspects of validity, the indices are used to evaluate response patterns and give a clear picture of the functioning of the rating scale’s structure. Wolfe and Smith (2007b) suggested that
evaluation of the rating scale functioning can provide important information about the examinees’ use of the rating scale and determine if examinees’ responses are consistent with the intended use. Based on Linacre (1998b), Wolfe and Smith presented four essential guidelines used to ensure measurement accuracy and stability. These include:

1. Each rating scale should include a minimum of 10 total observations to ensure precision of thresholds.
2. Distribution of observations should be smooth and unimodal to support proper use of rating scale categories.
3. Average respondent measure of each category should be ordered, and increase with the values of the rating scale categories to offer evidence that the scoring model represents the construct.
4. The value of the unweighted mean-squared fit statistic, as described earlier, should be less than 2.0 to support responses of the rating categories are used in accordance with Rasch model expectations (Wolfe & Smith, 2007b).

Wolfe and Smith (2007b) presented four additional guidelines that can be used to evaluate measurement accuracy and stability. The first is that category ratings need to be consistent with model predictions. These model predictions should be consistent with ratings. Wolfe and Smith suggested that the consistency of model predictions with ratings can be measured by coherence statistics which reflect “the percentage of ratings that are in the same category as the model-based expected ratings and the percentage of model-based expected ratings that are consistent with the actual rating” (p. 252).

The second guideline involves the ordering of the category thresholds. If ordered, the category thresholds can verify that each is associated with a distinct and appropriate range of the underlying ability scale.
According to Wolfe and Smith (2007b), the third guideline suggests that adjacent
category thresholds should be a minimum distance apart based on the number of categories.
In this case, where used a 5-point scale, Wolfe and Smith suggested a difference of at least
.81 logits. Finally, the authors recommended a maximum of 5 logits between adjacent
category thresholds. When these conditions are satisfied, we may conclude that the rating
scale was utilized by respondents as intended by the item developers.

Similar to the item fit statistics relating to the content aspect of validity, the Rasch
person-fit statistic identifies responses of examinees who may be inconsistent or unexpected,
given the model expectations. Inconsistencies may highlight problematic response patterns
such as guessing, carelessness, or item bias that can interfere with the measurement of the
construct that is intended (Wolfe & Smith, 2007b). Unexpected response patterns challenge
the Rasch model’s position that higher-ability examinees have a higher probability of
responding correctly to harder items than the lower-ability examinees. If the data show
unexpected patterns of responses, these examinees should be examined further. Similar to
the item-fit statistics discussed in the prior section, Wolfe and Smith suggested that a cut
score of $\pm 2.00$ for the person fit-statistic is considered reasonable for identifying person
misfit (Wolfe & Smith, 2007b).

Evaluation for the theoretical item hierarchy can be used to provide additional
evidence relevant to the substantive aspect of validity. Wolfe and Smith (2007b) presented
two methods for acquiring evidence of theoretical item hierarchy. Both are based on a
feature of the Rasch model—its ability to have a unique ordering of items by difficulty along
a continuum. Because this item ordering is invariant over the entire ability continuum, test
developers can easily evaluate whether the intended item hierarchy is consistent with the
hierarchy produced by the examinee responses. Simply put, do the items intended to be
difficult actually require a higher ability for correct response than those that were intended to be easy? And inversely, do the items intended to be easy actually require a lesser ability for correct response than those that were intended to be difficult? Test developers may rely on an expert panel to rate item difficulty and create an item hierarchy plot that identifies the relationship between the values given by expert raters to the empirical item calibrations. If the item hierarchy is consistent with the theoretical construct of the test, test developers may consider this evidence relevant to the substantive aspect of validity.

**Evidence relevant to the structural aspect of validity.** Wolfe and Smith (2007b) presented two types of validity evidence relevant to the structural aspect of validity. First, consistent with the *Standards* (AERA, APA & NCME, 1999), they suggested that correlation analyses should be considered. In addition, Wolfe and Smith suggested that a Rasch dimensionality analysis can contribute additional evidence relevant to test structure. Rasch dimensionality analysis is used to evaluate whether the data are unidimensional. An assumption of the more commonly used Rasch models, the quality of unidimensionality demands that the test items measure a single dominant construct. According to Wright and Stone (1979), this assumption is a requirement of all meaningful measurement. Rasch dimensionality analysis can indicate the degree to which the items in a test measure a single, measurable trait. According to Wolfe and Smith (2007b), evaluation of unidimensionality is performed via four steps. The first step is to review point measure correlations. Although described earlier in this chapter, Wolfe and Smith (2007b) may best describe an appropriate application of point measures to “indicate the degree to which the scores on a particular item are consistent with the average score across the remaining items” (p. 246). Positive point measures are ideal, and indicate positive correlation of the item scores with the average
score on the remaining items. A negative value indicates negatively correlated items and potentially threatens unidimensionality.

Item fit indices, specifically the mean-square and standardized fit statistics, can also be used to evaluate unidimensionality. In particular, a standardized outfit statistic over +2.00 typically indicates unexpectedly high residuals. Items with fit statistics over the accepted range should be considered potential threats to unidimensionality and reviewed for quality and relevance to content.

The third step in evaluating unidimensionality involves principal component analysis (PCA) of the residuals. Based on Linacre (1998a, 2004b), Smith (2002) stated that if a principal component is found, then residual variance is due to common factors shared by the items. To perform PCA of the residuals, the principal component is extracted based on a standardized residuals matrix. Eigenvalues from the Rasch-scaled measures and the principal component analysis are then placed on a common scale, and their relationship examined in the resulting “scree plot” (Wolfe & Smith, 2007b, p. 258). Wolfe and Smith (2007b) described two scree plots that indicate a threat to unidimensionality. The first is a scree plot that shows a second component that accounts for more variance than is contributed by a single item, resulting in eigenvalues greater than 1.0. The second problematic scree plot of eigenvalues fails to reach asymptote at the second component. According to Wolfe and Smith, either case justifies further evaluation of dimensionality.

The fourth step in evaluating unidimensionality entails a comparison of the aforementioned PCA of residuals to a PCA of simulated data, described in detail by Smith (2002).

The assumption of unidimensionality can also be evaluated to some extent by reviewing both local independence and monotonicity. Because unidimensionality has such a
predictive relationship with both local independence and monotonicity, the conditions of
meeting the assumptions of local independence and monotonicity essentially support the
requirements of unidimensionality.

Local independence identifies whether the observed responses are independent of
each other, given an individual’s score on the latent variable, and determined by calculating
the residual correlation between all of the item pairs. In this context, local independence
means that the selected anxiety rating of a student on an item will not be related to the
student’s ratings on other items after controlling for ability. Local independence can be
evaluated using a variety of indices, including the G² (Bishop, Fienberg, & Holland, 1975;
Chen & Thissen, 1997), Yen’s Q₃ statistic (Yen, 1984), and Fisher’s Z index of item
dependency (Fisher, 1924). According to Chen and Thissen (1997), Yen’s Q₃ statistic is
generally more robust at identifying local dependence when compared to the G² (Chen &
Thissen, 1997). In spite of this praise, there are limitations of Yen’s Q₃ statistic, which are
reviewed in detail by Shen (1997). There are three primary steps associated with calculating
the Fisher’s Z statistic. First, I determined the standardized residuals for each rating of
student n on item i, using the formula:

\[ d_{ni} = \frac{(\text{observed rating} - \text{expected rating})}{SE_n}. \]  \hspace{1cm} (6)

Second, I correlated the standardized residuals, dₙᵢ, for all pairs of items i, j across all
students, thereby producing a residual correlation coefficient (rᵢⱼ) for each item pair. Finally,
using the Microsoft (2003) Excel computer program, I computed Fisher’s Z statistic, using
the formula:
\[ z_{ij} = \frac{1}{2} \ln \frac{1 + r_{ij}}{1 - r_{ij}} \]  

(7)

which has an expected value (mean) of

\[ E(z) = \frac{1}{2} \ln \frac{1 + E(Q_3)}{1 - E(Q_3)} = 1/2[\ln(1 + E(Q_3)) - \ln(1 - E(Q_3))] \]  

(8)

with \( E(Q3) \) equal to \(-1 / (I-1)\), and variance equal to \( 1 / (N-3) \), where \( Q_3 \) is Yen’s \( Q_3 \) statistic (Yen, 1984, 1993), a Pearson correlation coefficient between the residuals of a pair of items after partalling out the measured construct, and \( N \) is the sample size.

Standardized Fisher’s Z is then shown as:

\[ STDZ_{ij} = z_{ij} - E(z) / \sqrt{1/(N-3)} \]  

(9)

which has an approximate normal distribution. Typically, violations of local independence are identified if the item pair’s Fisher’s Z statistic is more than two standard deviations above the mean of Fisher’s Z or less than two standard deviations below the mean of Fisher’s Z (Kim, De Ayala, Ferdous, & Nering, 2007). Violation of the assumption of local independence indicates additional factors, beyond the primary Rasch dimension, may be contributing to highly correlated residuals on these two items. Common explanations for highly correlated residuals may include similarly worded items or items sharing common stimuli.
Evidence relevant to the generalizability aspect of validity. Traditionally, test reliability analyses common in classical test theory have been used to evaluate the degree to which measures maintain their meaning across time and contexts. Wolfe and Smith (2007b) have extended the evidence relevant to the generalizability aspect of validity to include invariance of person measures across time and contexts. According to Wolfe and Smith, person measure invariance “serves as an umbrella for several types of analyses designed to demonstrate that person measures maintain internal or external relationships across measurement contexts such as time, subsets of items, subpopulations, or administration contexts” (p. 260). They deconstructed generalizability evidence into five components, including (a) reliability, (b) item calibration invariance, (c) person measure invariance, (d) differential prediction, and (e) validity generalization. I will review two of the components most relevant to the purpose of this study: reliability and item calibration invariance.

Reliability, as discussed earlier in this chapter, has traditionally been addressed via classical test theory. Although indices can be determined via traditional methods, the Rasch model allows for more accurate computation of many reliability estimates. More specifically, after meeting the assumption that the data fit the model, the Rasch model offers a more precise estimate of a person’s ability and an item’s difficulty because the individual standard error (SE) is utilized instead of a sample or averaged standard error of measurement (SEM). According to Smith (2001) the individual standard error provides a more accurate estimate by providing a quantifiable precision for every person measure and item difficulty and describing the range within which each person’s ability or item’s difficulty falls.
Person separation reliability is analogous to the traditional Cronbach alpha or Kuder-Richardson Formula 20 (KR-20) and is considered the index of reliability of measures for the Rasch family of models that reflects an estimate of the ratio of true score variance to observed score variance (Wright & Masters, 1982).

Person separation reliability ($R_p$) is described mathematically as

$$R_p = \frac{SD_p^2 - MSE_p}{SD_p^2},$$  \hspace{1cm} (10)

where $SD_p^2$ is the observed variance of measures, and $MSE_p$ is the mean-square of the measurement error for person ability. To estimate how well persons separate items, the same equation can be used to calculate item separation reliability by simple substitution of observed variance of the items ($SD_i^2$) and the mean square measurement error of item difficulty ($MSE_i$). Separation reliability values can range from 0 to 1, and a higher value indicates more separation.

**Evidence relevant to the external aspect of validity.** The external aspect of validity has been most closely associated with the traditional form of validity evidence coined “criterion validity.” Wolfe and Smith (2007b) described appropriate evidence being reported as “correlations between sets of measures or hypothesis test statistics and effect sizes” (p. 268). According to Wolfe and Smith, the correlations between sets of measures are most often reported in the form of a multitrait–multimethod matrix that identifies correlations between multiple measures of same, similar, and different constructs. Developed by Campbell and Fiske (1959), the matrices identify correlations between theoretical constructs that may be expected to be related or unrelated. These relationships have been traditionally
termed *convergent evidence*, for positive (related) correlations, and *divergent*, or *discriminant, evidence* for negative (unrelated) correlations. Wolfe and Smith provided four additional forms of validity evidence that are obtained through the use of the Rasch model, including examining group and within-individual comparisons, responsiveness, person-item maps and strata, and confirmation of alternative item hierarchies for known group differences. For the purpose of this study, I will focus on the latter two forms of evidence relevant to external aspect of validity.

Person-item maps and strata are used to determine whether an instrument is capable of detecting change in person measures over time or following an intervention. This quality is also known as responsiveness (Medical Outcomes Trust Scientific Advisory Committee, 1995) and is typically called sensitivity, particularly in medicine (Wolfe & Smith, 2007b). A clear benefit of the Rasch model and outcomes of Rasch analysis are person-item maps, which visually identify the relationship of a group of person abilities and a set of item difficulties along a continuum. The person strata index identifies how many different levels of person measures are statistically distinguishable for a set of items. In order to offer evidence relevant to the external aspect of validity, the person strata must be capable of distinguishing between at least two levels of anxiety (Wolfe & Smith, 2007b). If the item difficulties for a set of items are highly variable and are well dispersed along the previously mentioned continuum, the resulting person strata will be higher than if items’ difficulties were more homogeneous. A high person strata demonstrates high responsiveness and indicates the test items are capable of detecting change in examinee ability over time or following an intervention, and offers support for the external aspect of validity.

The concept of item calibration hierarchies and invariance of item calibrations across groups has been previously discussed as evidence relevant to the substantive and
generalizability aspects of validity, respectively. The same methods can be used to offer
evidence relevant to the external aspect of validity. For example, expected differences in
groups’ target construct can be hypothesized, and tested via DIF analysis. As described
earlier, SAI can be utilized to identify differences between the item calibration values of
two groups. Doing so may offer potential evidence relevant to the external aspect of
validity.

In combination, the amalgam of philosophies based on the Standards, Medical
Outcomes Trust (1995), and the alternative framework from Wolfe and Smith (2007b) offer
five sources of validity evidence, providing the basis of a sound validity argument that
justifies the interpretation of scores or measures obtained from tests. As the Standards
(2001) suggested, a sound validity argument is established using “a set of propositions that
support the proposed interpretation for the particular purpose of testing” (p. 9). The validity
propositions should describe the degree to which existing evidence and theory support the
intended interpretation of test scores for that particular use. The alternative framework
presented by Wolfe and Smith (2007a, 2007b) projects traditionally accepted types of
validity evidence onto a contemporary framework, supported by the practical application of
the Rasch model. This alternative framework of validity evidence is crucial to the
development of a new instrument that measures the intended construct.

Psychometric Properties and Validity Evidence of Instruments

For Measuring Clinical Anxiety

The development of a new instrument that measures anxiety requires a thorough
review of existing instruments used to measure anxiety. A review of the strengths and
weaknesses of available instruments, combined with evaluations of their psychometric
properties using the alternative framework of validity evidence will help guide the creation
of a robust instrument that measures anxiety in medical students toward the male GUR examination. Although not exhaustive, I have reviewed instruments most relevant to this study.

**State–Trait Anxiety Inventory**

The STAI (Spielberger et al., 1970) consists of two separate self-report scales that measure state and trait anxiety. The STAI has been adapted in more than 30 languages for cross-cultural research and clinical practice (Spielberger & Diaz-Guerrero, 1983). To date, more than 3,000 studies and reviews of the standard, full-length scales of the STAI have appeared in the research literature in many fields, including psychology, social sciences, education, and medicine. Researchers have also utilized and evaluated shortened versions of the state anxiety scale of the STAI for use in the medical setting (Chlan et al, 2003; Marteau & Bekker, 1992; Tluczek et al, 2009).

The current version of the standard form of the STAI from Spielberger et al. (1983) contains two discrete scales—the state scale, STAI Form Y-1, consists of 20 statements that evaluate how respondents feel *right now, at this moment*. The trait scale, STAI Form Y-2, consists of 20 statements that assess how people *generally* feel. The principal qualities evaluated by the STAI are feelings of apprehension, tension, nervousness, and worry. Spielberger et al. (1970) noted in the original STAI manual that the state scale may also be used to evaluate how a subject felt at a particular time in the recent past, how they anticipate they will feel in a specific future situation, and in a variety of hypothetical situations.

Regardless of the subject’s ability, examinees must read the directions to the subjects in order to emphasize the difference between the state and trait scales and maintain construct validity of the overall scale. Also, following standard practice, Spielberger et al. (1970) recommended that the state scale be administered first, followed by the trait scale in order to
avoid potential influences the administration of trait scale may have on state anxiety measure. Responses to the scales are rated on a 4-point rating scale, although the categories differ for the scales. The anxiety scale item categories describe the intensity of a subject’s feelings and are labeled 1 (Not at all), 2 (Somewhat), 3 (Moderately so), and 4 (Very much so). The trait anxiety scale item categories describe the frequency of a subject’s feelings and are labeled 1 (Almost never), 2 (Sometimes), 3 (Often), and 4 (Almost always).

Scoring is summative for both scales of the STAI. Spielberger et al. (1970, p. 12) explained the scoring rubric in the STAI manual as a rating of 4 indicates the presence of a high level of anxiety for 10 S-Anxiety items and 11 T-Anxiety items (e.g., “I feel frightened,” “I feel upset”). A high rating indicates the absence of anxiety for the remaining 10 S-Anxiety items and 9 T-Anxiety items (e.g., “I feel calm,” “I feel relaxed”). The scoring weights for the anxiety-present items are simply the same as the selected category numbers on the test form. The scoring weights for the anxiety-absent items are reversed, so that responses marked 1, 2, 3, or 4 are scored 4, 3, 2, or 1, respectively. The anxiety-absent items for which the scoring weights are reversed for the state scale are items 1, 2, 5, 8, 10, 11, 15, 16, 19, and 20. The anxiety-absent items for which the scoring weights are reversed for the trait scale are items 21, 23, 26, 27, 30, 33, 34, 36, and 39.

The STAI manual supplies scoring keys for the scales to aid in hand scoring. The manual also supplies normative data for high school students \((N = 424)\), university students \((N = 855)\), working adults \((N = 1,838)\), and military recruits (2 studies, \(N = 1,701, N = 263\)) for the current form (Form Y), and male neuropsychiatric patients \((N = 461)\), general medical and surgical patients \((N = 161)\), and young prisoners \((N = 212)\).

Reliability data for both Forms X and Y are presented as evidence of internal consistency in the STAI manual (1970). The stability coefficients for Form Y are based on
two groups of high school students testing in classroom settings. The test-retest coefficients (Cronbach index of internal consistency) are identified for male \( (N = 173) \) and female \( (N = 178) \) students for the trait scale at 30-day and 60-day intervals. For male students, coefficients were \(.71\) and \(.68\) for the 30-day and 60-day interval, respectively. For female students the resulting coefficients were \(.75\) and \(.65\) for the 30-day and 60-day interval, respectively. These results indicate a moderate-to-high correlation, and reasonable reliability for the trait scale for high school students.

For the state scale, the test-retest coefficients for male \( (N = 178) \) and female \( (N = 179) \) students at 30-day and 60-day intervals indicate something different. For male students coefficients were \(.62\) and \(.51\) for the 30-day and 60-day interval, respectively, while for female students, the resulting coefficients were much lower: \(.34\) and \(.36\) for the 30-day and 60-day interval, respectively. These results indicate a moderate correlation for the male students, and low correlation for the female students. Given the construct of the state scale, the changes reflected in the female scores likely reflect the transitory nature of anxiety, the defined purpose of the scale.

Although Spielberger et al. (1970) discussed traditionally described construct validity, a large amount of evidence presented by the authors supports the external aspect of validity for the state and trait scales. I reviewed the evidence supplied by the authors in the current STAI manual. I also introduced contemporary studies that have reviewed the current forms of the state and trait scales, including the shortened versions of the state anxiety scale of the STAI.

**Evidence relevant to the external aspect of validity.** Spielberger et al. (1970) considered the external aspect of validity as evidence for validation for use of the STAI to measure state and trait anxiety in a variety of sample populations. The STAI manual
presents validity evidence that, in large, is based on data resulting from form X. Because Spielberger et al. offered high correlations (ranging from .96 to .98) between forms X and Y, I reviewed the evidence presented in the STAI manual.

The first evidence is based on a comparison of anxiety from contrasting groups. For the state scale, the authors suggested that on comparison of mean scores of military recruits to college students, high school students, and working adults, the recruits’ mean scores would reflect higher level of stress during their rigorous training program. Male military recruits \((M = 44.05, SD = 12.18)\) was found to be higher than male college students \((M = 36.47, SD = 10.02)\), male high school students \((M = 39.45, SD = 9.74)\), and male working adult males \((M = 35.72, SD = 10.40)\). A similar trend was found for the female subjects. This prediction was supported by the consistency of the mean state and trait scores for the other groups, which indicated lower emotional stress for those groups.

Other evidence relevant to external validity presented by the authors includes correlations between the state and trait scales for the same groups of subjects—working adults, college students, high school students, and military recruits. The authors suggested that subjects with high trait anxiety scores tend to have high state anxiety scores, even in times of neutral anxiety. The authors added that, based on the trait-state theory (Spielberger, 1972), correlations between the two scales will likely be lower in high threat situations, such as those characterized by physical danger, while correlations will be higher during threats to self-esteem, such as social evaluation conditions. Given this construct, the data suggest that the higher correlation, as seen in the working adult group (males, \(r = .75\), females, \(r = .70\)) may indicate that this group is faced with less physical threat, yet more evaluative situations. At the other extreme, the lower correlation between the state and trait anxiety scales may indicate a higher perceived physical threat and decreased threat to self-esteem.
More evidence relevant to the external aspect of validity is presented in the form of traditionally coined “convergent validity.” A correlation between the STAI trait anxiety scale and the other available personality tests at the time is presented in the STAI manual (Spielberger, 1972). The most widely used scales that measured trait anxiety were the IPAT Anxiety Scale (IPAT; Cattell & Scheier, 1963) and the Taylor Manifest Anxiety Scale (TMAS; Taylor, 1953). Correlations between the STAI trait scale, the IPAT and the TMAS were moderately high, ranging from .73 to .85. Although correlations were reasonable, Spielberger suggested that the new STAI trait scale might be a more robust scale at measuring trait anxiety than the other two scales. They indicated that the IPAT items might be related more closely to anger, with the example item, “Often, I get angry with people too quickly” and the TMAS might have reflected a construct of depression rather than anxiety, with items such as “I cry easily,” and “At times I think I am no good at all.”

Another relevant study presents more evidence relevant to external aspect of validity in the form of both convergent and discriminant validity. Spielberger (1972) examined correlations of the state and trait scores with scores from the Personality Research Form (PRF; Jackson, 1967) a self-report scale that was designed to assess personality traits broadly relevant to the functioning of individuals in a wide variety of situations for young adults and adults. The study included 124 university students who sought counseling for emotional and educational–vocational problems at the center associated with the university. Spielberger reported that the mean state anxiety scores were significantly higher for students with emotional problems ($M = 40.37, SD = 9.34, N = 38$) than those with educational–vocational problems ($M = 36.68, SD = 8.49, N = 124$). Similar results were reported for the trait anxiety scores. These results indicated that the STAI scale was able to differentiate between higher emotional stress in the subjects. The authors of the STAI presented evidence
relevant to the external aspect of validity of the scale to measure trait and state anxiety as defined by the model originally posited by Spielberger and Gorsuch (1966). The data from the original analyses presented by the STAI manual are somewhat convincing. Although the study primarily examined the relationships of the original STAI (Form X) to other anxiety scales available in the early 1970s, an updated review of validity analyses with the current STAI (Form Y) and contemporary scales would offer more information. Most of the subsequent studies have focused on evaluating evidence relevant to the structural aspect of validity. I will review these now.

Evidence relevant to the structural aspect of validity. The study by Kendall, Finch, Auerbach, Hooke, and Mikulka (1976) examined evidence relevant to the structural aspect of validity using principal-components factor analysis on 140 university students. Responses were subjected to principal components factor analysis using an indirect oblimen oblique rotation. Results indicated three primary factors that accounted for 80.6% of the total variance. The first factor, A-Trait (cognitive anxiety), was represented by 14 items, all from the STAI trait scale. The second factor, A-State (negative descriptors), was represented by 6 of the 10 negatively worded items from the STAI state scale and related to worry, rumination, and disappointment. The third factor, A-State (positive descriptors), was composed of 8 of the 10 positive items from the STAI state scale that related to feelings of self-confidence, and happiness. These findings support the trait-state anxiety theory originally presented by Spielberger (1972) and offers evidence relevant to structural aspect of validity for the STAI to discriminate between state and trait anxieties.

The study by Bieling, Antony, and Swinson (1998) examined the evidence relevant to structural aspect of validity of the trait scale for STAI form Y-2 on a group of adult patients diagnosed with psychiatric disorders ($N = 212$) defined by the *DSM-IV* (American
Psychiatric Association, 1994), and a group of nonclinical volunteers ($N = 49$). Using exploratory factor analysis, the factor structure for the patient group indicated a 2-factor solution, consistent with the 2-factor structure identified by Spielberger (1989).

The authors then tested three possible construct models. These included a trait anxiety model that represented a single construct underlying the 20 items, a second model with two uncorrelated factors, anxiety and depression, and a third model that represented a tripartite, hierarchical model that included general negative affect, anxiety, and depression factors. Items were subjected to confirmatory factor analysis using maximum likelihood extraction. Resulting goodness of fit (GFI) and adjusted goodness of fit indices (AGFI) suggested that the hierarchical solution (GFI = .84, AGFI = .80) provided a superior fit to the data than the 1-factor model (GFI = .77, AGFI = .72) and the 2-factor model (GFI = .82, AGFI = .78). This study substantiated results from an earlier study by Clark and Watson (1991).

Although different from Spielberger et al.’s original construct of the trait scale of the STAI and of that posited by Kendall et al. (1976), the study of Bieling et al. (1998) may not necessarily contradict the prior studies. This study may have merely pointed out the complexity of the relationship between anxiety and depression, and closely linking depression and ego to Spielberger’s definition of trait anxiety.

Research by Tenenbaum, Furst, and Weingarten (1985) is particularly relevant to this study. The authors evaluated the construct of the STAI scales by examining item responses using the Rasch model (Rasch, 1960/1980). The authors examined item fit and items’ capability for measuring and differentiating among a variety of persons with different anxiety levels using the trait and state STAI scales with athletes ($N = 100$) before competition.
Results for the trait scale identified two misfitting items: 14—Comfortable, and 1–Calm, which were recommended to be removed. Six items elicited odd responses, including 8–Rested, 16–Content, 20–Pleasant, 6–Upset, 1–Calm, and 14–Comfortable.

Several items were located at the same point in the continuum, indicating either the scale’s inability to differentiate the athletes’ ability to endorse items, or the redundancy of items. Most problematic, the distribution of items along the continuum was rather homogenous, making it difficult to identify the athletes with low or high trait anxiety. The authors did not reveal person separation reliability so I am unable to discuss this in detail.

Tenenbaum et al. (1985) performed a similar analysis for the STAI state scale. Six misfitting items were indicated: 4–Regretful, 17–Worried, 7–Worrying (misfortune), 2–Secure, 5–At ease, and 1–Calm. Parallel to the trait anxiety scale, the items for the state anxiety scale were placed within a single logit range along the continuum. Given the homogeneity of the person measures, differentiation of the athletes with low state anxiety from those with high state anxiety was difficult.

In summary, the authors reported that the STAI anxiety and trait scale were not robust enough to differentiate levels of trait and state anxiety. In spite of these findings, this report does not contradict Spielberger’s original construct. Spielberger’s primary concern was not with measuring intensity of anxiety, but measuring presence of anxiety.
Six-Item Short Form of the State STAI

Because I utilized a shortened form of the STAI state scale in this study, a review of
the relevant validity evidence that has been evaluated by the founding research associated
with the 6-item short form of the STAI state scale is warranted.

Evidence relevant to the structural aspect of validity. Tluczek, Henriques, and
Brown (2009) evaluated two versions of shortened, 6-item scales developed by Marteau and
Bekker (1992) and Chlan et al., (2003). In this study, a sample of 288 parents with infants
were split in four groups and designated as either (a) parents of infants with cystic fibrosis
(\(n = 26\)), (b) parents of infants with congenital hypothyroidism (\(n = 39\)), (c) parents of
infants identified as cystic fibrosis carriers (\(n = 45\)), or (d) parents of healthy infants
(\(n = 40\)). The short forms of the state scale of the STAI were completed at three intervals,
with the infants approximately 2, 6 and 12 months of age.

Evidence relevant to the generalizability aspect of validity. The aforementioned
study by Marteau and Bekker (1992) also compared the scores of the four groups of
examinees across the 4-, 6-, and 20-item forms. Internal consistency was found to be lowest
for the 4-item form (\(\alpha = .77\)), and higher for the 6-item (\(\alpha = .82\)) and 20-item (\(\alpha = .91\))
forms. The moderate reliability for the 6-item STAI form that included anxiety-present items
(original items 3, 6, and 17) and anxiety-absent items (original items 1, 15, and 16) suggests
evidence relevant to the generalizability aspect of validity.

Evidence relevant to the generalizability aspect of validity was also evaluated by the
previously discussed study by Tluczek et al. (2009). In this study, the authors analyzed the
internal consistency of the two 6-item forms of the STAI state scale developed by Marteau
and Bekker (1992) and Chlan and colleagues (2003). Tluczek et al. evaluated the internal
consistencies across three time points, using Cronbach alpha, and identified that the scale
developed by Marteau and Bekker had marginally higher internal consistency ($r = .79$ at
Times 1 and 2 and $r = .81$ at Time 3) than that of Chlan et al. ($r = .79$ at Times 1 and 3 and
$r = .75$ at Time 2) offering evidence relevant to the generalizability aspect of validity. In
spite of these differences, the validity evidence may suggest that shortened versions of the
STAI may be a viable alternative to the long version of the STAI for a number of scenarios.

**Evidence relevant to the external aspect of validity.** The first study to evaluate
evidence relevant to the external aspect of validity was by Marteau and Bekker (1992). The
authors created five shortened forms (2-, 4-, 6-, 8-, and 10-item versions) of the state scale
of the STAI. The shortened forms were administered to four examinee groups consisting of
medical students ($n = 38$), nursing students ($n = 45$), pregnant women ($n = 200$), and
pregnant women with abnormal prenatal test results ($n = 23$). The authors determined that all
but the 2-item form had high correlations ($r > .90$) with the original 20-item STAI state
scale. The high correlation of the scores from the short forms supported convergent validity
and offered evidence of external aspect of validity.

In a similar study, Chlan et al. (2003) developed a different 6-item, short-form state
scale of the STAI and evaluated validity evidence for use with critically-ill patients. Unlike
the short scale developed by Marteau and Bekker that is focused on the cognitive
phenomena of anxiety, this scale is focused on the somatic/physiologic phenomena of
anxiety. Like the scale developed by Marteau and Bekker (1992), this scale was 2-factored,
but shared only one item (original item 17). The scale developed by Chlan et al. included
anxiety-present (original items 9, 12, and 17) and anxiety-absent items (original items 5, 10,
and 20). Exploratory factor analysis identified the 6 items that most strongly correlated with
the original 20-item state scale. The resulting 6-item state STAI developed by Chlan et al.
had high correlation with the original state STAI ($r = .92$), offering evidence relevant to the external aspect of validity.

Tluczek et al. (2009) also evaluated evidence relevant to the external aspect of validity. In this study—discussed earlier in greater detail—the authors evaluated correlations of the 6-item scales developed by Marteau and Bekker (1992), and Chlan et al. (2003) with the original 20-item STAI. Correlations indicated high convergence ($r > .93$) for both versions of 6-item scale across all three time points offering evidence relevant to the external aspect of validity.

Criticism of the STAI has been primarily directed toward the trait scale and its potentially unstable construct. Kelly (2004) examined the relationship between worry and trait anxiety by examining correlational data between the STAI trait scale and the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990) in a sample of university students ($N = 218$). Using principal-component factor analysis with a varimax rotation, he identified four factors that accounted for 57.8% of the variance. He labeled the factors (a) Negative affect, (b) Worry, (c) Fatigue, and (d) Avoidance. Three items with depressive experience did not load on any of these factors. Correlational data suggested that the PSWQ correlated significantly ($p < .0001$) with each factor: Worry ($r = .74$), Negative affect ($r = .61$), Fatigue ($r = .74$) and Avoidance ($r = .43$). Controlling for Worry, the partial correlation between the PSWQ and the remaining STAI factors, Negative affect, Fatigue, and Avoidance were significant ($p < .0001$), $r = .38$. This result indicated that the STAI trait scale likely continues to predict worry independent of the worry factor. More importantly, this study is evidence that should caution against use of the STAI trait scale without examining the potential variables that seem to influence the factor structure of the scale.
The Beck Anxiety Inventory®

The Beck Anxiety Inventory® (BAI; Beck, Brown, Steer, Eidelson, & Riskind, 1987) is a self-report rating scale that measures the severity of anxiety in adults. The scale originated from three earlier self-report instruments that measure a variety of aspects of anxiety. These include the Anxiety Checklist (Beck et al., 1985) used to evaluate the severity of anxiety in clinically depressed patients, the PDR Check List (Beck, 1978), which measures the side effects of antianxiety and antidepressant medications, and finally, the Situational Anxiety Check List (Beck, 1982), used to evaluate the severity of somatic and cognitive symptoms of anxiety in general and two specific situations—public speaking and some other anxiety-evoking situation designated by the patient.

The BAI contains 21 descriptive statements of anxiety symptoms that are rated on a 4-point scale. Items were developed according to diagnoses defined by either the American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders, (DSM-III; 1980) or American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders, (DSM-III-R; 1987), suggesting that BAI findings be applicable to inpatients diagnosed as having DSM-III or DSM-III-R disorders.

Instructions for administration include both self-administration and oral administration, with expected administration times between 5 and 10 minutes. In addition to space for the person’s name and testing date, the BAI record form includes the following statement in bolded type:

Below is a list of common symptoms of anxiety. Please carefully read each item in the list. Indicate how much you have been bothered by each symptom during the PAST WEEK, INCLUDING TODAY, by placing an X in the corresponding space in the column next to each symptom.

Available categories are 0 (Not at all); 1 (Mildly, it did not bother me much); 2 (Moderately, it was very unpleasant but I could stand it); and 3 (Severely, I could barely stand it).
Scoring guidelines of BAI have changed since its introduction. The latest version of the BAI Manual (Beck & Steer, 1993) recommends interpreting the intensity of self-reported anxiety based on a total score. Total scores from 0 to 7 points reflect a minimal level of anxiety, while a score between 8 and 15 indicates a mild form of anxiety, scores of 16 to 25 reflect moderate anxiety, and scores of over 26 indicate severe anxiety with a maximum score of 63 points.

Although originally developed to be used for inpatients diagnosed with DSM-III or DSM-III-R disorders, evidence for validity of the BAI to measure anxiety in psychiatric, clinical, and nonclinical populations has made this scale accessible for a variety of applications. The BAI Manual (Beck & Steer, 1993) presented data from a clinical sample of 393 outpatients evaluated by the authors between 1985 and 1989. The sample included 236 women (60.1%), and 157 men (39.9%). The mean age was 37.1 (SD = 11.3) with the diagnostic composition of 95 (24.2%) panic with agoraphobia, 93 (23.7%) panic without agoraphobia, 4 (1.0%) agoraphobia without panic attacks, 44 (11.2%) social phobia, 19 (4.8%) simple phobia, 26 (6.6%) obsessive-compulsive, 2 (0.5%) posttraumatic stress disorders, 90 (22.9%) generalized anxiety, and 20 (5.1%) not otherwise specified anxiety disorder. Evidence relevant to structural, external, and generalizability aspects of validity were presented by Beck.

Evidence relevant to the structural aspect of validity. To support evidence relevant to the structural aspect of validity, Beck et al. (1988) presented factorial validity evidence based on a principal-factor analysis with promax rotation for 393 outpatients with anxiety disorders. The authors reported that the BAI yielded two correlated ($r = .56, p < .001$) dimensions. The first dimension, representing somatic aspects of anxiety contained symptoms such as numbness, feeling hot, and sweating. The second factor
represented subjective and panic-related anxiety with such symptoms as fear of the worst happening and fear of losing control. After using principal-factor analysis with centroid cluster analysis and controlling for two confounding factors, age and gender, four clusters reflected neurophysiological, subjective, panic, and autonomic domains. The physiological domain consisted of seven items: 1–Numbness, 3–Wobbliness in the legs, 6–Dizzy or lightheadedness, 8–Unsteady, 12–Hands trembling, 13–Shaky, and 19–Faint. The subjective domain consisted of six items: 4–Unable to relax, 5–Fear of the worst happening, 9–Terrified, 10–Nervous, 14–Fear of losing control, and 17–Scared. The panic domain included 4 items: 7–Heart pounding or racing, 11–Feelings of choking, 15–Difficulty breathing, and 16–Fear of dying. The final domain, autonomic, also included 4 items: 2–Feeling hot, 18–Indigestion or discomfort in the abdomen, 20–Face flushed, and 21–Sweating (not due to heat). No other studies have refuted or supported these findings.

**Evidence relevant to the external aspect of validity.** Documented evidence relevant to external aspect of validity seems to be inconsistent. Evidence of convergent validity with other anxiety measures has been offered by the authors of the BAI. For example, Beck et al. (1987) reported a moderate \( r = .51 \) correlation of the BAI with the Hamilton Anxiety Rating Scale, suggesting reasonable evidence of convergence with the scale. Further, Beck et al. examined the correlations between the BAI and the Hamilton Anxiety Rating Scale (Hamilton, 1959), as restructured by Riskind, Beck, Brown and Steer (1987). They found the correlation to be \( .51 \) \((p < .001)\). In spite of this finding, one study (Creamer, Foran, & Bell, 1995) criticized Beck et al. for presenting such data from the Hamilton clinician-rated anxiety scale, yet failed to compare them with other available self-
rated anxiety scales, such as the STAI and the Zung Self-Rated Anxiety Scale (Zung, 1971). Although Beck et al. failed to perform a comparison with other self-rated anxiety scales, they compared the BAI with another clinician-rated scale, the Cognition Check List (Beck et al., 1987), a scale that measured the frequencies of dysfunctional anxiety-relating cognitions. Similar to the comparison with the Hamilton Anxiety Rating Scale, a comparison of the BAI with the CCL resulted in a correlation of .51 (p < .001).

An independent study did examine the relationship of the BAI with the aforementioned STAI, Form Y (Spielberger et al., 1983). Fydrich, Dowdall, and Chambless (1990) reported moderate correlations with the state (r = .47, p < .01) and trait (r = .58, p < .001) scale of the STAI, even with Bonferroni adjustments to minimize familywise type-I error. The same study reported significant correlations with both the 7-day anxiety rating (r = .54, p < .001) of the Weekly Record of Anxiety and Depression (WRAD; Barlow & Craske, 1994). In spite of offering convergence with other anxiety scales, the study by Fydrich et al. (1990) also reported the BAI was significantly correlated with two depression scales, the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Beck, Rush, Shaw, & Emery, 1979), r = .50, p < .001, and the WRAD (r = .38, p < .01). This contradicted earlier findings that suggested discriminant validity presented years earlier by Beck et al (1961). In support of Fydrich et al.’s findings, another study (Dent & Salkovskis, 1986) supported significant correlations of the BAI with the BDI (r = .61, p < 0.01), indicating evidence of convergence with the depression measure. Possibly in response to these findings suggesting convergence with depressive measures, some discriminant evidence was later presented by the authors of the BAI (Riskind, et al., 1987) in a clinical setting. The authors presented the scale’s moderate correlation with the BDI (Beck & Steer, 1987; Beck et al., 1961; Beck, Rush, Shaw, & Emery, 1979), r = .48
(\(p < .001\)), the depression scale of the Cognition Check List (CCL; Beck et al, 1987), 
\(r = .22 (p < .05)\), and the revised Hamilton Rating Scale for Depression (HRSD-R; 
Hamilton, 1960), as restructured by Riskind et al. (1987), \(r = .25 (p < .05)\) as evidence of 
discriminant validity.

The BAI authors’ claim of strong validity evidence to measure anxiety in clinical and 
nonclinical settings for patients with a variety of anxiety-related diagnoses has not gone 
unnoticed and has drawn responses from other researchers. It seems that most researchers 
agree the scale’s discriminatory power between anxiety and depression in patients in the 
clinical environment is better than other available scales (Creamer et al., 1995). In spite of 
this praise, there remains debate over data. Although the original factor analysis by Beck 
identified two factors, somatic symptoms and subjective anxiety and panic symptoms, 
subsequent studies (Beck et al., 1988; Borden, Peterson, & Jackson, 1991) have not been 
able to replicate this structure. The study that most closely replicated a 2-factor structure 
similar to the one presented by Beck et al. (1988) used a computer version of the BAI with a 
mixed inpatient sample (Steer, Rissmiller, Ranieri, & Beck 1993). Borden, Peterson, and 
Jackson (1991) identified a 5-factor principal component structure that accounted for 60% of 
variance. These results most closely resembled the study by Beck and Steer (1993) that 
identified a 4-factor structure with anxious outpatients. More recently, Creamer et al. (1995) 
compared the BAI with the STAI in nonstressful and stressful conditions in a nonclinical 
setting. The authors showed the BAI as having good psychometric properties with a high 
level of internal consistency (Cronbach alpha of .91 and .90 for repeated trials). In spite of 
this evidence, the authors questioned the evidence of construct validity offered by Beck 
et al.
The BAI demonstrated similar correlations with the trait \((r = .57 \text{ and } .68 \text{ at Times 1 and Times 2, respectively})\) and state \((r = .56 \text{ and } .64)\) scales of the STAI. The test-retest correlation for the BAI after a 7-week interval was moderate at .62, which was comparable to the state scale of the STAI, indicating that the BAI may measure a construct more closely related to state anxiety, that which is a measure of current psychopathology, and not a more stable personality characteristic. Further, Creamer et al. (1995) suggested that the discriminatory power of the BAI scale originates from its strong focus on cognitive and physiological symptoms of anxiety that most clearly differentiate anxiety from depression, and is therefore a “cleaner” measure of anxiety (p. 484). This may explain the low correlation of the BAI with the STAI, as the STAI generally focuses more on affective descriptors that typically begin with “I feel…”. Creamer et al. also suggested that the defined construct of the BAI seems to vary across situations, as the factor analysis revealed a unifactorial solution on the first, low stress administration, but a 2-factor solution similar to that proposed initially by Beck et al. at the second, more stressful administration.

**The Hospital Anxiety and Depression Scale (HADS)**

The Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983) is a self-report measure that identifies anxiety disorders and depression among medical patients in nonpsychiatric hospital clinics. In addition to its use for adults, validity evidence has supported the use of the scale for older adults (Snaith, 2003), and adolescents (White, Leach, Sims & Cottrell, 1999). Widely accepted and used in a number of languages, the HADS has been often cited. One review paper noted more than 747 HADS-referring papers identified in Medline, ISI and PsycINFO indexed journals (Bjelland, Dahl, Haug, & Neckelmann, 2002). The scale is actually composed of two scales—anxiety (HADS-A) and depression (HADS-D).
Unlike the STAI (Spielberger et al, 1983), which measures state and trait anxiety, the HADS limited construct focuses on present state anxiety, as indicated by instructions directing subjects to respond to a timeframe “over the last few days.” The HADS may be readministered at weekly (or longer) intervals to provide a useful record of progress. Because the intended construct of the scale is limited to measuring anxiety and depression in nonpsychiatric settings, the scale avoids measuring symptoms relating to serious mental disorders. Unlike the broad-spectrum of the construct measured by one scale (reviewed later in this dissertation), the Hamilton Anxiety Rating Scale (HAM-A; Hamilton, 1959, 1969), this scale avoids measuring symptoms associated with physical disorders, including dizziness, headaches, insomnia, anergia (lack of energy), and fatigue, in order to minimize “noise” from somatic disorders.

The current HADS scale consists of 14 items: 7 measure anxiety and 7 measure depression. Anxiety items reflect the typical wording psychological anxiety measures, while the depression items focus on assessing positive affectivity. Each item is scored using a 4-point scale (0 - 3). The categories vary in language and direction, with an equal amount of negatively worded questions for anxiety and depression items. Scoring of the scales has been established on four score ranges. For either scale, a score below 8 is considered normal, a score between 8 and 10 identifies a mild disorder, while a score above 11 indicates caseness, or moderate disorder, and a score over 15 indicates severe disorder. The maximum possible score for both scales is 21.

Zigmond and Snaith (1983) offered evidence relevant to the structural aspect of validity and asserted that internal consistency of the HADS has been established on the current version of the scale with cancer patients (N = 568; Moorey et al., 1991) The internal consistency of the anxiety scale was based on Cronbach alpha, and determined to be .93,
while for the depression scale internal consistency was .90. The authors of the scale claim that test-retest reliability was established using healthy respondents and was .89 and .92 for the anxiety and depression scales, respectively. A literature review by Bjelland et al. (2002) indicated that internal consistency of the HADS was tested in numerous studies. The minimum recommended Cronbach coefficient alpha is .60 for self-reported measures (Nunnally & Burnstein, 1994). In every study reviewed by Bjelland et al., internal consistency was determined to be adequate.

Snaith and Zigmond (1994) pointed to the study by Moorey et al. (1991) to affirm their position that the HADS is 2-factored, offering more evidence relevant to structural aspect of validity. Moorey et al. identified two independent factors, anxiety and depression, accounting for 53% of the variance. Lisspers, Nygren, and Söderman (1997) and Spinhoven et al. (1997) have since tested and supported the 2-factor structure of the HADS. Using principal component analysis, both studies determined that a 2-factor structure was stable for general population (Bjelland et al., 2002). Lisspers et al. found the same 2-factor structure for both males and females. Independently, Spinhoven et al. found that the 2-factor solution was stable across different age groups from the general population and in different clinical samples (general practice, medical outpatients with unexplained somatic symptoms, and psychiatric outpatients).

Zigmond and Snaith (1983) also presented evidence relevant to the external aspect of validity. The HADS authors considered “concurrent,” or convergent, and construct validity as evidence for validation for use of the HADS to measure anxiety and depression in medical patient samples. The review of studies examining convergent validity by Bjelland et al. (2002) indicated reasonable evidence of convergent validity for both the anxiety and depression scales. Evidence of convergent validity for the HADS anxiety scale included
studies reporting correlations between the HADS and other anxiety measuring scales. Two studies reported adequate correlations between the HADS and General Health Questionnaire (GHQ-28; Goldberg, 1972), which were .50 (Caplan, 1994), and .68 (Chandarana, Eals, Steingart, Bellamy, & Allen, 1987). Two other studies reported higher correlations between the HADS and Clinical Anxiety Scale (CAS; Snaith, Baugh, Clayden, Husain, & Sipple, 1982), which were .69 (Snaith & Taylor, 1985), and .75 (Upadhyaya & Stanley, 1993). Five studies reported moderate correlations between the HADS anxiety scale and the STAI, ranging from .64 to .81 (Elliott, 1993; Herrmann, Schlotz & Kreuzer, 1991; Lisspers et al., 1997; Millar, Jelicic, Bonke, & Asbury, 1995; Savard, Laberge, Gauthier, Ivers, & Bergeron, 1998). The same studies identified moderate correlations between the HADS depression scale and the trait scale of STAI, ranging from .68 to .71. Finally, one study by Lepine, Godchau, Brun, and Lemperiere (1985) found low correlations (.34 to .44) between the HAM-A and the anxiety scale of the HADS, indicating the measures may have different constructs. Overall, the correlational data between the HADS and a number of other anxiety and depression self-reports act as reasonable evidence of convergent validity, and bolster the evidence relevant to the external aspect of validity.

Despite the existence of studies that have tested and supported the 2-factor (anxiety and depression) structure of the HADS for application with a variety of population samples, there have been studies that suggested a contradicting dimensionality. Clark and Watson (1991) identified a 3-factor model, based on the tripartite theory of anxiety and depression. The authors posited a 3-factor model, comprised of general distress (negative affect), physiological hyperarousal (anxiety), and anhedonia (depression, or lack of positive affect). More recently, Dunbar, Ford, Hunt, and Der (2000) utilized confirmatory factor analysis with data from community-based participants and tested different factor models using
confirmatory factor analysis samples of three different age groups (aged approximately 18, 39, and 58 years) from the general population ($N = 2547$). The study supported Clark and Watson’s 3-factor model (1991), indicating that their model was superior when compared to other models that resulted from a weaker theoretical foundation. Dunbar et al. also tested the 2-factor model achieved by Moorey et al. (1991) in the same study. They showed measures of goodness of fit relatively close to the 3-factor model (comparative fit index 0.93 vs. 0.97 and root mean square error of approximation 0.06 versus 0.04). According to Bjelland et al. (2002), Dunbar et al.’s (2000) evidence suggested that the HADS is a bidimensional measure, indicating both anxiety and depression domains. They also posited that the findings were not absolute with loadings too low to confirm differentiated scales. Further, Dunbar et al. warned users of high correlations between the domains, suggesting a lack of discriminative power of the HADS. The authors suggested a variety of possible explanations for the lack of discriminative power of the HADS, including blaming overfitting data from large sample sizes, and referenced Clark and Watson (1991), a study that indicated that there has consistently been a high degree of overlap between the anxiety and depression constructs, particularly for studies examining the general population samples. Dunbar et al. also reminded us that this is not a criticism unique to the HADS; it is a problem that continues to be an issue for many self-reported anxiety and depression measures. Regardless of explanation, it is evident that factorial evidence does not indicate a solid construct as claimed by its authors. It seems that brevity of the scale that has generated high acceptability in the medical field has also created instability in the construct. As Dunbar et al. suggested, the HADS should be supplemented for better measuring of negative affect and autonomic anxiety constructs.
In spite of the construct inconsistencies, the HADS has been used by numerous researchers to measure anxiety and depression in medical settings (Brandberg, Bolund, Sugurdardottir, Sjoden, & Sullivan, 1992; Lundqvist Sjosteen, Blomstrand, Lind, & Sullivan, 1991; Shiell & Shiell, 1991; Thompson & Meddis, 1990). Many examples include studies that measure anxiety and depression in patients with cancer or other somatic conditions (Bjelland et al., 2002). Bjelland pointed out the psychometric properties of HADS were “seldom the main issue in these studies, the sample sizes were mostly relatively small \((n < 250)\), and the results were frequently given without further discussion” (Bjelland, p. 70). A review of existing studies’ analyses of convergent validity may help gain more information regarding the validity of the HADS to measure anxiety and depression in medical and general population samples.

**The Hamilton Anxiety Scale (HAM-A)**

The Hamilton Anxiety Scale (HAM-A; Hamilton, 1959, 1969), also known as the Hamilton Rating Scale for Anxiety (HRSA) and the Hamilton Anxiety Scale (HAS), was introduced by Hamilton as one of the first clinical rating scales used to measure the variety of symptoms associated with anxiety in adults and children. As reviewed in the opening of this chapter, anxiety symptoms in the psychiatric setting may include cognitive, psychic, behavioral, and somatic phenomena. These phenomena are common to the eight Anxiety Disorders as defined by the fourth major revision of *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2000).

Originally designed to measure a patient’s “neurotic anxiety state,” the HAM-A has become one of the most widely used and accepted outcome measure for the evaluation of anxiety symptoms. Common applications of the HAM-A include evaluation of efficacy of antianxiety drugs (McEnvoy et al., 1980; Tyrer, Gardner, Lambourn, & Whitford, 1980).
and measurement of anxiety states in patients diagnosed with both depressive (Gjerris et al., 1983) and anxiety (Sheehan, Sheehan, Raj, & Janavs, 2007) neuroses. In its current state, the Hamilton Anxiety Scale seen in Appendix D consists of 14 items, each representing a grouping of manifestations of anxiety (and associated phenomena), including mood (worry, apprehension), fear (specific phobias), insomnia, somatic symptoms (muscle fatigue, hot/cold flashes), cognitive dysfunction (concentration), and behavioral symptoms (restlessness, tremors). Of all existing anxiety measures, the HAM-A covers the broadest spectrum of anxiety phenomena. The items are rated by an interviewing clinician on a 5-point scale ranging from 0 (Not present) to 4 (Severe). The patient’s anxiety level is determined by a total score; a score of less than 17 identifies a mild anxiety, a score between 18 and 24 is mild to moderate, while a score between 25 and 30 indicates a moderate to severe level of anxiety.

In spite of the instrument’s wide use, studies that have evaluated validity evidence of the HAM-A to measure anxiety in clinical use have been limited. Validity evidence of the HAM-A has been mostly relevant to the external aspect of validity in the form of discriminant validity with its sister scale that measures clinical depressive symptoms, the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960). Hamilton (1967) reported a correlation of .25 between the single-item ratings on the HAM-A (anxious mood) and the total score (global depression scale) on the HRSD for a heterogeneous sample. Later, a similar study by Mowbray (1972) indicated a correlation of .60 between the anxious mood and global rating for depression, revealing a weaker discriminant power. To date, validity evidence of HAM-A has been limited to the clinical setting to measure anxiety severity in patients who have been diagnosed with anxiety neurosis. There has been no evidence that
supports validity for measuring anxiety severity in nonmedical subjects for the HAM-A and thus fails to offer evidence relevant to the generalizability aspect of validity.

In spite of the extensive use of the HAM-A in the field of psychiatry, there are clear limitations of the scale for measuring anxiety in medical students. First, the scale has been developed for use in psychiatry only. Some opponents assert that the HAM-A lacks sensitivity to measure anxiety in subjects without true neuroses (Hamilton Anxiety Scale, 2007). This means that subjects having lower levels of clinical anxiety will not be identified, and the HAM-A fails to support evidence relevant to the generalizability aspect of validity. Additionally, the scale has been developed as an interviewer-rated scale. This format is impractical with a large sample and launches concerns over the potential decrease in interrater reliability. Some researchers (Bruss, Gruenberg, Goldstein, & Barber, 1994; Clark & Watson, 1991) have posited that interrater reliability for the HAM-A is unpredictable and weakens evidence relevant to the structural aspect of validity. This belief was formed, in part, by a lack of established reliability aids that include administration and instructions for the original scale. The scale also lacked scripted questions for the interviewers. These deficiencies most definitely lead to inconsistent use of the scale (Clark, 1989) and decrease accuracy of the measured outcomes, particularly in cross-site or cross-rater comparisons (Bruss et al., 1994), and ultimately threaten evidence relevant to content aspect of validity. It makes sense that higher reliability is observed with heterogeneous samples, and better trained (more consistent) clinical raters. When these conditions falter, even slightly, the interrater reliability declines and threatens evidence relevant to the structural aspect of validity. Researchers have attempted to rectify the deficiencies of the original scale and create structured interview guides that include instructions for standardized administration and anchors for assigning severity ratings (Bruss et al., 1994; Shear et al., 2001).
In spite of attempts at improving the administration of the HAM-A, the most concerning limitation of the scale is the aforementioned existence of high correlation with depression scales, which continues to threaten the structural aspect of validity. Because there is a tendency for overlap of anxiety and depressive symptoms (Clark, 1989; Watson & Kendall, 1989), particularly in depressed subjects (Watts, 1966), it is not possible to develop an anxiety scale that adequately measures all components of anxiety, yet avoids convergent measure of a depression construct.

**Other Scales Used to Measure Anxiety**

Although discussion has focused on the four previous scales, there are a number of available scales used to measure anxiety in a variety of settings and applications. I have not focused on these scales due to their lack of documented use or limited practical value for this application. Despite this, a brief overview of the most notable of these scales is warranted.

The Profile of Mood States measures (POMS; McNair, Lorr & Droppleman, 1992) has been described as the most commonly used measure of anxious and depressed mood (Clark & Watson, 1991). According to the authors of the POMS, it measures a variety of mood states, including one positive and five negative mood state domains (McNair et al., 1992). The positive domain measured is vigor-activity, while the negative state domains include (a) depression-dejection, (b) anger-hostility, (c) fatigue-inertia, (d) confusion-bewilderment, and finally, the applicable mood state, (e) tension-anxiety. Rating is performed via interviewers. Via interview, respondents rate 65 items of the six domain adjectives that describe emotional and cognitive mood states on a 5-point intensity scale. The items are presented in terms of how the respondent has been feeling in the past week, ranging from 0 *(Not at all)*, to 4 *(Extremely*. The total mood disturbance score is then derived
by subtracting the positive score (vigor-activity) from the total of the five negative mood scores. A higher score indicates greater distress. McNair et al. presented evidence relevant to the structural aspect of validity and suggested that internal consistency for all items was .90 or above. Test-retest reliability ranged between .65 for vigor-activity and .74 for depression-dejection, presenting evidence relevant to structural aspect of validity. The authors posit that the measures have been standardized for psychiatric outpatients and college students when administered together (McNair et al., 1992). Although the POMS measures have been heavily used in psychiatric and medical applications, there is confirmation that the six POMS measures have been collectively utilized in medical education (Kirsling, Kochar, & Chan, 1989).

In spite of this, only one New Zealand study indicated that the tension-anxiety measure had been administered individually (Barker-Collo, 2003). The population consisted of university students, although they were primarily Pekaha, Maori, and Asian. This lent little validity evidence for use of the POMS tension-anxiety measure with the average medical student in the United States. The lack of evidence relevant to the generalizability aspect of validity is compounded by the scale’s impractical interview-led administration, and ultimately eliminates it from the list of viable options to measure the anxiety of several medical students at once.

The Symptom Checklist-90 Revised (SCL-90R; Derogatis, 1983) is a 90-item, self-report questionnaire that was designed to measure global psychological symptoms of subjects at both the global and symptom level dimensions. As stated by Zare’ et al. (2004), there are nine dimensions measured, including physiological, depression, anxiety, interpersonal sensitivity, hostility, obsessive-compulsive, phobic anxiety, paranoid ideation, and finally, psychoticism. The 90 items are presented in terms of how the respondent has
been feeling in the last seven days, ranging from 0 (Not at all) to 4 (Extremely). Scoring is determined by the subjects’ responses on a 5-point Likert scale that range from 0 (Not at all) to 4 (Extremely). General psychological distress is measured by the global severity index (GSI), and calculated by dividing the summed total score by the number of items. Caseness of specific diagnostic syndromes is evaluated using the raw scores from each of the nine symptom dimension scores. The raw score is averaged across the respective items of that particular dimension, and then referred to gender-appropriate norms (Zare’ et al., 2004).

Derogatis (1983) reported the internal consistency of the scales range from .77 to .90, offering evidence relevant to the structural aspect of validity. Test-retest correlations ranged from .78 to .90, supporting evidence relevant to the generalizability aspect of validity. Zare’ et al. (2004) suggested the use of the SCL-90R has been accepted in clinical practices (Craig & Abeloff, 1974; Derogatis et al., 1983), and in research as a screening tool and outcomes measures (Pottenger et al., 1978; Steinglass, 1980). Recently, the SCL-90R has been implemented in medical education (Zare’ et al., 2004). Although the scale measures of the SCL-90R show promise in graduate medical education with surgical residents, the scales have yet to be implemented in undergraduate medical education with medical students. Because of the lack of documented use with medical students, combined with the relatively high cost of the SCL-90R, the scales were not considered a viable option for this study.

The Self-Rated Anxiety Scale (SAS; Zung, 1971) is a 20-item self-report questionnaire designed to measure the severity of anxiety in the clinical setting. Items are scored on a 4-point Likert scale. Categories include 1 (None or a little of the time), 2 (Some of the time), 3 (Good part of the time), and 4 (Most or all of the time). Wording of 15 of the items indicates increasing anxiety levels, while wording of 5 items indicates decreasing
anxiety levels. Items appear to measure relevant anxiety symptoms including cognitive, somatic, and emotional symptomatology. Raw scores are summed to identify anxiety severity. The scoring ranges from 20 to 80 points. Individuals that score between 20 and 44 points are considered to have normal level of anxiety, while scores between 45 and 59 identify mild to moderate anxiety levels. Scores of 60 to 74 indicate marked to severe anxiety levels, and finally, 75 to 80 indicate extreme anxiety levels. The SAS has been utilized to assess generalized anxiety, and treatment effects (Saletu-Zyhlarz, Anderer, Arnold, Saletu, 2003; Sramek, Frackiewicz, & Cutlet, 1997).

The SAS has also been utilized in medical education (Liu, Oda, Peng, & Asai 1997). Although interesting outcomes were reported in the study, the sample was limited to Chinese medical students, which limits the relevance to medical education in the United States. In spite of easy accessibility to the SAS, there is little research that utilizes the scale in Western medical education. For this reason, the SAS is excluded from consideration in this study.

The Perceived Stress Scale (PSS; Cohen, Kamarck, & Mermelstein, 1983) was originally designed as a 14-item self-report questionnaire that measures “the degree to which respondents find their lives unpredictable, uncontrollable, and overloading” (Cohen, 1986, p. 717) in adult subjects. The items are presented in terms of how the respondent has been feeling in the last month, while respondents are requested to indicate how often they felt a certain way. Scoring is determined by the subjects’ responses on a 5-point Likert scale. Categories include 0 (Never), 1 (Almost never), 2 (Sometimes), 4 (Fairly often), and 5 (Very often). Since its introduction, abbreviated, 10- and 4-item PSS scales have become available, and used in a variety of settings including clinical (Mitchell, Crane & Kim, 2008), research
Regardless of the version, criticism of the PSS exists. For example, Lazarus and Folkman (1986) contend that the 4-item PSS fails to adequately identify the interdependent variables, the “multiple systems” (p. 718) of the stress processes that include environmental context of the demands, and the constraints, coping resources, and personal agendas of the individual.

Criticism aside, we must also identify the relevance of stress, and ultimately, the relevance of the PSS, to this study. I am a proponent of the universally accepted belief that stress is directly associated with anxiety (Finkelstein, Brownstein, Scott & Lan, 2007; Spielberger, 1979, 1980; Stewart et al., 1999). Although this position may lend potential value to the PSS measures, we must review the relationship of stress and anxiety as defined for the purposes of this study. Earlier in this paper I differentiated the two concepts: stress is defined as a physical, chemical, or emotional stimulus that causes tension or the result of a stimulus, while anxiety is a response to that stressor. Simply put, stress is an entity that produces a response, anxiety. Although I agree stress directly influences anxiety, the purpose of the study is the measure of anxiety. Therefore, the items of the PSS will not be considered in this study.

**Summary of the Instruments Used to Measure Clinical Anxiety**

Reviewing the strengths and limitations of the numerous available scales that are used to measure anxiety will help determine the best possible measure that should be considered for use in this study. In particular, strengths and weaknesses specific to validity evidence and practical application will be the focus of my consideration. The Profile of Mood States measures (POMS; McNair et al., 1992) appear to be relevant to the study and
considered a robust and commonly-used measure of anxious and depressed mood (Kirsling et al., 1989). In spite of this praise, only one New Zealand study has administered the tension-anxiety measure individually (Barker-Collo, 2003). This limitation of the measure is compounded by the scale’s impractical interview-led administration, and ultimately eliminates it from the list of viable options to measure the anxiety of several medical students at once.

Supporting evidence of the Symptom Checklist-90 Revised (SCL-90R; Derogatis, 1983) includes high acceptance across a variety of settings, including medical education. Additionally, relevant studies have offered validity evidence relevant to structural and generalizability aspects of validity. Although the utilization of the global severity index and scale measures of the SCL-90R show promise in graduate medical education with surgical residents, they have yet to be implemented in undergraduate medical education with medical students. Given the lack of evidence relevant to the generalizability aspect of validity, combined with the scale’s relatively high cost, I excluded the SCL-90R for use in this study.

Benefits of the Self-Rated Anxiety Scale (SAS; Zung, 1971) include its design of measuring three of the intended four relevant anxiety symptomology. Additionally, the simple scoring rubric offered by the SAS and its well-defined categories are appealing. In spite of these benefits, studies that have utilized the SAS offer diminished relevance to medical education in the United States. Moreover, the lack of statistical information and validity evidence for the use of the SAS in medical education means that the scale will be excluded from consideration in this study.

Although commonly accepted and utilized in a variety of settings and applications, the PSS will not be considered for the purpose of this study. In spite of evidence of value of the PSS to measure stress in adults, we must revisit the primary objective of the study, to
measure anxiety in medical students. For the purpose of this study, based on definition, the PSS will not be considered in this study because it is designed to measure stress and not anxiety.

The following four scales were considered more strongly than the previous scales, based on a longer and more diverse application history. Additionally, studies that utilized and implemented the scales consistently offered more robust item analysis with compelling reliability estimates and validity evidence for use in medical education, the intended population of this study.

The BAI boasts discriminatory power between anxiety and depression in patients in the clinical environment (Creamer et al., 1995). Additionally, the manual for BAI includes five types of validity evidence, although there remains discourse concerning the psychometric properties of the scale. A number of studies that have examined the relationship of the BAI with other anxiety scales contradict those reported by Beck et al. in the developmental study. Only one study (Steer et al., 1993) produced a 2-factor structure of the BAI, similar to that claimed by the authors of the BAI, while recent studies have not been able to replicate this structure. Furthermore, Creamer et al. (1995) suggested that the defined construct of the BAI varies across situations. In spite of the benefits of its potential discriminant power, the BAI is largely limited by the threat of instability of the construct structure under varied conditions of stress. Because of this, it has been eliminated from consideration in this study.

The Hamilton Anxiety Scale (HAM-A; Hamilton, 1959, 1969), also known as the Hamilton Rating Scale for Anxiety (HRSA) and the Hamilton Anxiety Scale (HAS), has become the most widely used and accepted outcome measure for the evaluation of a patient’s anxiety. Of all existing anxiety measures, the HAM-A covers the broadest
spectrum of anxiety phenomena. The scoring rubric is simple to understand, and based on a total score, with cutoff criteria offered in order to identify patients’ level of anxiety neurosis. Restrictions of the HAM-A include validity evidence of the scale that is limited to supporting the measure of anxiety severity in patients that have been diagnosed with anxiety neurosis. To date, there has been no evidence that supports validity for measuring anxiety severity in nonmedical subjects for the HAM-A. Additionally, some opponents assert that the HAM-A lacks sensitivity to measure anxiety in subjects without true neuroses (Hamilton Anxiety Scale, 2007). This means that subjects having lower levels of clinical anxiety will not be identified.

Another limitation of the scale is the impractical interviewer-rated administration, which has launched concerns over the potential decrease in interrater reliability and threat to structural aspect of validity. The threat of unpredictable interrater reliability for the HAM-A was formed, in part, by a lack of established reliability aids that include administration and instructions for the original scale and scripted questions for the interviewers. In spite of attempts at improving the administration of the HAM-A, the most concerning limitation is the existence of high correlation with depression scales (Clark, 1989; Watson & Kendall, 1989). Snaith et al. (1982) have attempted to improve the specificity of the HAM-A, resulting in a HAM-A derived scale, the Clinical Anxiety Scale (CAS; Snaith et al., 1982). In spite of the development of an improved scale, this limitation remains with the HAM-A, and the lack of validity evidence that supports the use of the scale in a nonclinical setting, combined with the threat of unpredictable interrater reliability, compels me to consider alternative measures for this study.

The HADS (Zigmond & Snaith, 1983) is widely accepted and used to measure anxiety and depression in nonpsychiatric settings. Unlike the broad-spectrum of the
construct measured by the aforementioned HAM-A, the HADS avoids measuring symptoms associated with physical disorders in order to minimize “noise” from somatic disorders. The simple scoring rubric is easy to use, while caseness ratings make this scale even more accessible. Other positive features of the HADS have been discussed in detail earlier in this chapter. For example, internal consistency of the HADS has been tested and adequately established in numerous studies (Bjelland et al., 2002; Moorey et al., 1991), supporting evidence relevant to structural aspect of validity.

Bjelland et al. (2002) also indicated reasonable evidence of convergent validity for both the anxiety and depression scales. A number of studies (Elliott, 1993; Goldberg, 1972; Herrmann et al., 1991; Millar, Jelicic, Bonke, & Asbury, 1995; Lisspers et al., 1997; Savard et al., 1998; Snaith et al., 1982; Upadhyaya & Stanley, 1993) presented evidence of convergent validity for the HADS scales, while correlational data between the HADS and a number of other anxiety and depression self-reports acted as reasonable evidence relevant to the external aspect of validity.

In spite of the evidence that adequately supports convergent validity, evidence relevant to the structural aspect of validity has been inconsistent. The 2-factored construct presented by Zigmond and Snaith (1983) has been substantiated (Lisspers, 1997; Moorey et al., 1991; Spinhoven et al., 1997), while another study (Clark & Watson, 1991) identified a 3-factor model. Dunbar et al. (2000) supported this structure and warned users of the decreased discriminative power of the HADS. This position was supported by Clark and Watson, who were more concerned about the HADS’ lack of discriminative power for studies examining general population samples. We should remember that this criticism is not unique to the HADS. This limitation is a problem that continues to be an issue for many self-reported anxiety and depression measures. Regardless of explanation, it is evident that
the factorial evidence does not indicate a solid construct as originally claimed by its authors. It seems that the brevity of the scale that has generated high acceptability in the medical field, has succeeded at the cost of construct stability. For this reason, the anxiety scale of the HADS will not be used in this study.

There are many advantages of the STAI, beginning with the easily accessible STAI manual that offers well-documented administration directions and simple scoring rubric, and supplies normative data from a variety of samples. The STAI boasts a long history of enduring critical evaluations from a variety of sources, and a large amount of supporting evidence relevant to the external aspect of validity in the form of construct, discriminant, and convergent validity. Additionally, evidence relevant to the structural, generalizability, and external aspects of validity for the 6-item, short-form STAI offers promise for use in this study.

Limitations of the STAI have been directed at the trait scale. Originally presented by Shedletsky and Endler (1974) and supported by Kendall et al. (1976), the most significant limitation is the trait scale’s limited ability to assess only one dimension of the trait anxiety—the cognitive dimension of ego threat. Kendall et al. further posited that the unidimensionality of the trait scale may threaten the stability of the construct of the measure across a variety of situations and weaken evidence relevant to the generalizability aspect of validity. Another limitation to consider is the complicated relationship trait anxiety has with other negative mood states. As discussed earlier, the evaluation of factor structure of the STAI trait scale by Bieling et al. (1998) suggested a hierarchical model representing anxiety and depression at the same level loading on a higher-order factor.

The relationship of worry and trait anxiety has also been examined. As discussed earlier, Kelly (2004) examined correlational data between the STAI trait scale and the Penn
State Worry Questionnaire in a sample of university students. The result of the study indicated that the STAI trait scale predicts worry independent of the worry factor. These studies could be used as evidence that cautions against use of the STAI trait scale without examining the potential variables that seem to influence the factor structure of the scale. While this caution is justifiable, a review of scales that measure anxiety has revealed that confounding relationships with anxiety to other mood states is not unique to the STAI, but is apparent with many of the anxiety measures whose evidence of construct validity were tested.

A careful review of the most documented anxiety scales used in nonclinical settings has helped isolate the most appropriate scale to measure anxiety in medical students. The strengths and weaknesses of these scales are presented in Table 1.

Table 1

*A Review of Most Commonly Utilized Anxiety Measures, Identifying Their Strengths and Weaknesses for Measuring Anxiety in Medical Students*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Author(s) / Date</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beck Anxiety Inventory® (BAI)</td>
<td>Beck, Epstein, Brown, &amp; Steer / 1988</td>
<td>1- Flexible administration 2- Strong evidence relevant to structural validity 3- Best discriminatory power between anxiety and depression</td>
<td>1- Inconsistent evidence of convergence with outside measures with same construct (external aspect of validity)</td>
</tr>
<tr>
<td>Profile of Mood States (POMS)</td>
<td>McNair, Lorr &amp; Droppleman / 1992</td>
<td>1- Practical, well-defined scoring rubric</td>
<td>1- Very limited use in medical education 2- Impractical interview administration 3- Little validity evidence presented</td>
</tr>
</tbody>
</table>
Table 1 (continued)

*A Review of Most Commonly Utilized Anxiety Measures, Identifying Their Strengths and Weaknesses for Measuring Anxiety in Medical Students*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Author(s) / Date</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>State–Trait Anxiety Inventory (STAI)</td>
<td>Spielberger, Gorsuch, &amp; Lushene / 1970</td>
<td>1- Well accepted for use in variety of settings</td>
<td>1- Rasch analysis indicates scales’ inability to differentiate person ability in current state</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2- Practical administration/scoring</td>
<td>2- Trait scale found to be convergent with worry</td>
</tr>
<tr>
<td>State–Trait Anxiety Inventory (STAI)</td>
<td></td>
<td>3- Strong evidence relevant to external aspect of validity</td>
<td>scales (possible threat to external aspect of validity)</td>
</tr>
</tbody>
</table>
Psychometric Development and Validation of Instruments

Used to Measure Medical Student Anxiety

A review of the existing research that has examined medical students’ anxiety helps identify the need for an instrument that is supported by sufficient validity evidence that measures student anxiety in this particular setting. In spite of the broad agreement that anxiety is prevalent in medical education, until recently most research examining the mental state of medical students has been directed toward depression and global mental health, with few studies examining stress and anxiety. A recent literature review of anxiety and depression and other indicators of distress among medical students in the United States and Canada indicated 13 studies that were focused on depression, 8 on general mood disorder, 10 were combination studies, and one focused on anxiety and depression (Dyrbye et al., 2006). This review has supported the existence of a variety of studies that measured anxiety in medical education. There is a broad range of settings and applications, from examining generalized stress and anxiety in medical students and residents (Aktekin et al., 2001; Carmel & Bernstein, 1987; Coburn & Jovaisas, 1975; Conrad, 1988; Dyrbye et al., 2006; Leopold et al., 2005; Stewart et al., 1999) to examining medical students’ anxiety while they learn and perform anxiety-evoking tasks of physical examinations (Abraham et al., 2003; LeBlanc & Bandiera, 2007; Moss & McManus, 1992; Pugh & Salud, 2007; Pugh et al., 2009; Sarikayaya et al., 2006). It will be helpful to review the studies that examine medical students’ anxiety and the measures that have been used in these studies. Direction from previous work in the area will help guide the development of the instrument that will be used to measure medical students’ anxiety toward the male GUR examination.
Measurement of Medical Students’ Generalized Stress and Anxiety

This section begins with studies that have examined generalized stress in medical students (e.g., Aktekin et al., 2003; Firth, 1986; Niemi & Vainiomäki, 2006; Toews, Lockyer, Dobson, & Brownell, 1993, Toews et al., 1997). Firth used the Generalized Health Questionnaire (GHQ; Goldberg, 1972), supported by open-ended questions to measure stress in fourth-year British medical students ($n = 318$). There were no attempts to evaluate validity evidence of the GHQ for use among medical students, and analysis was limited to reporting frequencies. Firth’s work was most notable as one of the first to acknowledge stress among medical students and opened the door for further exploration of measuring anxiety among medical students. More studies that measured medical student stress and generalized anxiety followed Firth’s study. While there were some instruments supported by validity evidence for use in medical education, the selection of instruments was most likely based on availability and ease of use. In two separate studies, Toews et al. (1993, 1997) used the Symptom Check List and the revised Symptom Check List (SCL-90; Derogatis, Rickels, & Rock, 1976) to measure first through fourth-year medical students’ general stress levels. Findings indicated that medical students rated their psychological stress higher than population norms. Other studies (Henning et al., 1998; Lloyd & Gartrell, 1984; Vitaliano, Russo, Carr, & Heerwagen, 1984) that utilized the SCL-90 have been consistent with these findings.

More recently, educational researchers outside North America have taken interest in the potential negative effects of anxiety on academic performance of students. One study with University of Hong-Kong medical students (Stewart et al., 1999) investigated the relationship between stress-related measures and academic performance using personality and depression scales, and the STAI state anxiety scale. Bivariate correlations indicated
there were significant negative correlation between stress and academic performance. Two other studies outside North America examined the prevalence of generalized anxiety for medical students. Aktekin et al. (2003) used three measures that have been supported by sufficient validity evidence for use in education setting: the GHQ, the STAI and the commonly used BDI (Beck, et al. 1961) to assess psychological changes in Turkish medical students during their undergraduate education. Their findings showed that psychological test scores on the GHQ, the STAI, and the BDI rose significantly in medical students between the first and second years, indicating a decreased sense of well-being and increased anxiety and depression after the students’ first year. Niemi and Vainiomäki (2006) designed their own self-rated questionnaire, and with interviews, assessed perceived health, severity of distress, and somatic and emotional anxiety symptoms in a 6-year longitudinal study of Finnish medical students. These studies seem to demonstrate that the negative association anxiety has in medical education is not limited to North America and, ultimately, may have global consequences.

Other researchers focused their interest on measuring stress and generalized anxiety as it relates to program modification or interventions that specifically addressed medical students’ stress. Shapiro, Shapiro, and Schwartz (2000) identified 24 studies between 1966 and 1999 that implemented stress intervention programs in medical student and resident training curricula. Of these studies, five utilized rigorous scientific methods to identify and measure effects of interventions on stress and anxiety (Kelly, Bradlyn, Dubbert, & St. Lawrence, 1982; Kiecolt-Glaser et al., 1986; Nathan et al., 1987; Shapiro, Schwartz, & Bonner 1998; Whitehouse et al., 1996). Each of these studies utilized a standardized outcome measure that had been considered to have acceptable validity evidence. Shapiro et al.’s (2000) criticism of the other studies was varied and included small sample size with
limited statistical power, inappropriate participant recruitment and decreased randomization, lack of control groups, and finally little use of standardized outcome measures. Most notable, Shapiro et al. critically reviewed methodology, but did not question the validity evidence of the measures used in the studies, particularly, if the measures were considered “standardized.” Understandably, the authors applauded the use of available standardized psychological instruments in these studies, but Shapiro et al. failed to recognize that some of these instruments had not been supported by sufficient validity evidence for measuring stress and anxiety among medical students in specific scenarios. The scales had limited scope and their items did not reflect the specific scenarios with which medical students are faced.

In order to gain the most information from the existing instruments, researchers either modified them to better reflect their needs or supplemented them with additional items or scales. Following this trend, medical educators outside the United States also continued to utilize many of the same measures in a similar fashion. For example, Moffat, McConnachie, Ross, and Morrison (2004) supplemented the revised General Health Questionnaire (GHQ-12; Goldberg, & Williams 1988) and the Brief COPE (Carver, 1997), an instrument that measures coping methods, with a measure that listed 59 medical school-related stressors. These scales were used to assess the University of Glasgow first-year medical students’ (N = 275) sources of stress and coping mechanisms. They found that the prevalence of psychological stress (anxiety) and mean GHQ-12 scores increased significantly between the first three terms for all students, indicating increased stress and decreased well-being over the course of a year. The results also indicated that the primary stressors were related to training methodology, and not personal problems. The information may have been helpful and interesting to medical educators. However, because there was no
existing single gold standard to measure stress and anxiety, well-being, or coping mechanisms among medical students, the validity of the findings may have been compromised.

Nearly 10 years later, one study by Finkelstein, Brownstein, Scott, and Lan (2007) attempted to directly address the limitations of previous studies presented by Shapiro et al. (2000) and assessed the effectiveness of a stress reducing elective course on second-year medical students at the University of Washington School of Medicine. In this study, the authors surveyed control and study groups at the beginning of the quarter, at the end of the quarter, and 3 months later. The authors used the SCL-90 Anxiety Scale (Derogatis, 1983), the Profile of Mood States (POMS; McNair et al., 1992), the 2-item Depression Index (Whooley, Avins, Miranda, & Browner, 1997) and the Perceived Stress of Medical School (PSMS; Vitalicano et al., 1984). Analyses included chi-square tests for differences of proportions, t tests for differences between means, and repeated measures analysis of variance (ANOVA) to identify differences between group means across times. Reliability of the data was evaluated for both the study and comparison groups. Study group internal consistency reliability coefficients were high, with Cronbach alpha ranging from .81 to .92 for the SCL-90, .86 to .89 for the POMS, and from .78 to .85 for the PSMS. The comparison group Cronbach alpha ranged from .64 to .93 for the SCL-90, .73 to .92 for the POMS, and from .72 to .79 for the PSMS. Results indicated that for the study group, the stress-reduction course may have helped students develop more effective coping skills to manage the stresses of medical school. Even though Finkelstein and her colleagues did not evaluate the validity evidence of the scales used, the study supported that robust measurement methodologies were employed by medical educators.
Measurement of Medical Students’ Anxiety Toward Clinical Training

Other researchers assessed medical students’ anxiety during their clinical experience. In many cases, the lack of standardized scales to measure anxiety of medical students as it related to their experiences led many researchers to develop their own measures. In one example, the researchers developed and distributed their own questionnaire that rated students’ anxiety toward their initiation of students’ clinical services at St. George’s Hospital Medical School in London (Hayes et al., 2004). The survey is shown in Appendix F. They collected demographic data and included 13 statements relating to the students’ clinical experiences. Responses were rated using a modified, 7-point Likert-type scale, ranging from 1 (Strongly disagree) to 7 (Strongly agree). The higher scores indicated increased comfort in the clinical setting. Two open-ended questions allowed students to comment on the most positive and negative influences perceived during this time. A small pilot study with three members from a junior group with less experience and an older, more experienced student group, was performed to test face validity. Statistical analysis on the 13 items included an age comparison of the two student groups using the Mann–Whitney U test, while distribution of gender was compared between courses using a chi-square test. All reported $p$-values were 2-sided and the significance level was set at .05. The assumption of normality was verified and an analysis of covariance (ANCOVA) was also performed in which total anxiety was regressed on three covariates: (a) gender, (b) course, and (c) age. Results indicated that the two student groups differed in their mean total anxiety scores, with the older students reporting less anxiety ($M = 70.2, SE = 2.4$) versus the younger, less experienced students ($M = 57.2, SE = 0.93$; mean difference 13.0 [95% CI 7.3 – 18.6]. Not surprising, men reported significantly less anxiety ($M = 68.0, SE = 1.56$) than women ($M = 59.4, SE = 1.34$; mean difference 8.6 [95% CI 5.4 – 11.7]. Qualitative assessment was
performed on the open-ended questions. The authors identified a variety of limitations of the study, including the lack of evaluation of validity evidence of the scale.

This study was not the first that had attempted to measure medical student anxiety during their clinical experiences in medical programs, but acted as an example of the type of research outside of the United States spanning the past 30 years (Coburn & Jovaisas, 1975; Moss & McManus, 1992; Pitkälä & Mäntyranta, 2004; Radcliffe & Lester, 2003; Sarikayaya et al., 2006). Of these studies, two (Moss & McManus, 1992; Sarikayaya et al., 2006) had meaningful information as it relates to this study. Both studies identified that of all the clinical skills, performing the rectal examination—a key component of the male GUR exam—was the most anxiety-evoking task, even over the female pelvic exam. Although validity evidence was not addressed in either study, the studies offered clear evidence that performing the male GUR exam and other intimate examinations has been anxiety-evoking for medical students, and there was a need to examine the relationship between medical student anxiety and intimate examinations.

**Measurement of Medical Students’ Anxiety Toward Pelvic and Breast Exams**

Based on 10 years of simple observation, Buchwald (1979) identified six primary sources of anxiety for medical students while they learned and performed pelvic examinations. These were (a) fear of hurting the patient, (b) fear of being judged inept, (c) fear of inability to recognize pathology, (d) fear of sexual arousal, (e) fear of finding the examination unpleasant, and (f) fear of disturbance of the doctor-patient relationship. Buchwald reasoned that having an understanding of the specific anxiety-evoking aspects of the pelvic exam and discussing them during student seminars would reassure medical students as they learn the pelvic examination. Buchwald also discussed the need for objective evaluation methods in which to assess the potential impact these types of seminars
would have on students, patients, and faculty. Buchwald’s seminal paper launched an explosion of international interest in academic medicine in the identification, measurement, and interventions that address medical students’ anxiety as they learn and perform the pelvic examination.

Interestingly, this era also marked an upsurge in interest to define, characterize, and measure anxiety by the cognitive and behavioral psychology theorists. The most common application of anxiety measures in medical education was to identify intervention effects while students learned and performed the pelvic examination. In particular, many early studies focused on the identification of improvements in student competence with the incorporation of a paid, trained patient, later identified as the standardized patient by many medical educators (Livingstone & Ostrow, 1978; Nelson, 1978).

In a similar focus, Vontver et al. (1980) examined students when standardized patients had been utilized to teach intimate exams. The authors designed a randomized, split-halves study on second-year medical students of the University of Washington School of Medicine (N = 349). Forty of these students volunteered to be included in the experimental group. The experimental group was assigned to instruction by a professional patient (standardized patient by current definition), and the control group was assigned to instruction via a patient in an outpatient setting.

Vontver et al. (1980) began the study by implementing the State–Trait Anxiety Inventory (STAI; Spielberger et al., 1970), but modified the scale for the second year of the study. The authors’ reasoning for this was based on previous work by Sarason (1963), who suggested that anxiety scales be designed for specific situations. In addition to measuring anxiety with the modified STAI scale before, during, and after instruction and evaluation sessions, the researchers also recorded an easy-to-measure physiological response to
anxiety—the student heart rates. Additionally, a faculty member and standardized patient evaluated the students’ performance during these sessions. Statistical analyses included ANOVA for comparing the two groups of students to each other and the entire student group, repeated measures ANOVA to measure differences across instruction and evaluation, and regression analysis for examining correlation of student heart rates and self-reported anxiety.

Results indicated that anxiety was highest for the experimental group than for the rest of the class, particularly just prior to performing the pelvic examination during instruction. Because the only difference between the two groups was the heart-rate monitoring and evaluation, the act of the study itself seemed to have induced anxiety in the students. Most notable was the sharp increase in students’ heart rates while they performed the pelvic examination, potentially indicating that students experienced anxiety, even if they were not cognitively aware of that anxiety. Correlations for instructional phase were negative, indicating that when the reported anxiety decreased, the measured heart rate increased. Although these results could not be explained by the authors, this may be explained by the students’ fear of being evaluated as inept if they reported anxiety, compelling them to report their anxiety lower than it actually was. Another explanation may be offered by Meichenbaum (as cited in Hamilton, 1983) who argued that the students’ denial of anxiety may have been a natural coping strategy associated with the phenomena he called “cognitive-affective disguise” (as cited in Hamilton, 1983, p. 251.). Hamilton further explained that cognitive-affective disguise allows a subject to cope when faced with unavoidable situations that may be particularly aversive, frustrating, or conflicted.

In the last five years, there have been a number of other studies that have mirrored Vontver’s work and examined the effects of curriculum modifications on medical students’
anxiety while they learned the pelvic examination. Carr and Carmondy (2004) evaluated the effects of incorporating training assistants (standardized patients) in their Pelvic Examination Educational Program in an Australian medical school. The modification of the curriculum offered students increased opportunities to perform the pelvic exam outside of the clinic environment when compared to the previous year. The authors evaluated the implementation of the program using a variety of criteria. For both years, standardized patients assessed students’ competence at completing a gynecological history and performing a pelvic examination, using an 11-item, 5-point Likert scale questionnaire, ranging from 1 (Strongly disagree) to 5 (Strongly agree). Additionally, students (first year, \( n = 128 \); second year, \( n = 120 \)) rated the overall program and their anxiety toward performing the pelvic exam in a clinical setting, using a 12-item, 5-point Likert scale questionnaire. The second year, the standardized patients evaluated the program as well as students’ communication skills. Later in the academic year, a medical practitioner evaluated the students’ ability to take a sexual history and perform a pelvic exam on a patient.

Median scores for the students and standardized patients were differentiated into three groups representing positive, neutral, or negative evaluation. A comparison of the students’ competence, and anxiety across the two years was performed using the Mann-Whitney U test. Controlling for number of pelvic examinations, a one-way analysis of variance was also performed. Results indicated that the students and standardized patients positively evaluated the new program. More importantly, the students reported improved competence and reduced anxiety at performing the pelvic examination (\( p < 0.05 \)). A similar study by Wånggren et al. (2005) followed with matching results, although the study still lacked validity evidence. In spite of weak validity evidence across the variety of studies that have attempted to measure medical students’ anxiety while they perform the pelvic exam,
two notable studies have attempted to identify and measure specific sources of medical students’ anxiety while they learn and perform the pelvic examination.

In the first study, Abraham et al. (2003) examined anxiety of medical students while they learned and performed their first pelvic examination. In this study students from two Australian medical schools ($n = 226$) were enrolled in a course that training sessions consisted of a video and three learning stations before examining a standardized patient. Students were asked to complete the state scale of the STAI (Spielberger, 1972) and a “feelings” scale before performing the pelvic exam. After performing the pelvic exam the students completed modified versions of the state and trait scales of the STAI. The items of the scale were modified to reflect the specific aspects of the pelvic examination scenario, including embarrassment, personal experiences with the intimate exams, and sexual experience.

A variety of analyses were performed. Principal components analysis (PCA) established two factors for the feelings scale; a positive factor, that contained items that measured confidence, comfort, and a sense of being well-prepared, and the negative factor that contained items that measured anxiety, nervousness, apprehension, and discomfort. The PCA accounted for 70% of the variance of the items. Confirmatory factory analysis was then performed to support the 2-factor model. ANOVA was used to identify differences between the students at the two universities in preexamination state anxiety and confidence. State anxiety and positive and negative feeling before the pelvic examination were the dependent variables, while trait anxiety, age, gender, and language were the covariates. The same analyses were performed on the resulting data collected during the pelvic examination and then postexamination. Finally, a comparison of preexamination and postexamination data was performed. Results indicated that there were no notable relationships of age, gender and
language with the students’ state anxiety, negative or positive feelings before or during the examination.

The most notable results indicated that students’ preexamination state anxiety was significantly higher than students’ postexamination anxiety, \( t(226) = 4.04, p < .001 \). Prior to the examination 50 (22%) of the students reported feeling “shaky and sweaty,” 34 (15%) sick and nauseated. This included one who had vomited. During the examination 93 (42%) students felt embarrassed. Although the information resulting from this study was not dramatic, the researchers attempted to perform analytical methods that helped give the results more strength, and perhaps offer validity evidence. Although not disclosed in this study, the authors suggested in a previous publication (Abraham, 1998) that the scale items were developed with a logical analysis of the content of the theoretical construct, offering evidence relevant to content aspect of validity. Although the authors failed to perform item analysis, evidence relevant to the structural aspect of validity was offered in the form of principal components analysis and confirmatory factor analysis. Additionally, because there were no statistical differences found between the two universities, this could be considered evidence relevant to the generalizability aspect of validity, which may be enough evidence to justify the use of a modified STAI to measure medical student anxiety.

Pugh and Salud (2007) have offered the most information regarding specific sources of medical students’ anxiety toward breast and pelvic examinations. In the first study, the authors administered three 2-part surveys at three separate intervals during the second-year medical students’ breast examination skills training at a large Midwestern research institution \((N = 175)\). The first form was used to identify students’ baseline anxiety, while the second was used to identify changes in students’ anxiety after practicing on a simulated breast. The first part of the first form (Appendix G) requested the students to identify their
top three causes of anxiety from six items, including (a) causing harm or pain to the patient, (b) general performance anxiety, (c) personal/intimate nature of the exam, (d) missing a lesion, (e) palpating the nipple and areola, and (f) communicating effectively. The two other options available were: “I have no anxieties relating to this exam,” and a fill-in-the blank option, Other.

The second part of first form rated students’ overall comfort and comfort performing specific tasks of the breast exam, including visual inspection of the breast, eliciting nipple discharge, and detecting abnormalities. The items were rated on a modified, 6-point, Likert-type scale, with categories ranging from 1 (Extremely anxious) to 6 (Very comfortable) The second form was administered to the students after they were presented with a variety of information associated with the female breast examination via four stations that included visual aids, quizzes, video, and simulated breast models. The first part of the scale asked the students to rate the usefulness of the training modalities, while the second part was the same items from the first form, measuring students’ anxiety toward a specific task of the breast examination.

The third form of the survey was dispersed after the students’ experience of performing a breast examination on a standardized patient. The first part of the survey was in a dichotomous, true-false format and contained four items. These were (a) The breast exam was easier than I thought, (b) I need focused communication skills training for intimate examinations, (c) I felt rushed and did not have enough time to ask questions, and (d) I still have major concerns and questions about breast exams. Although the study was intended to measure student anxiety effects of the addition of simulators to the curriculum, I will focus on the review of the results as they relate to student anxiety in relationship to performing the breast examination. Analyses included ANOVA and chi-square test for
independence. P < .05 was considered significant. Reviewing the resulting frequency data from Form 1, 50% (84/168) of the students selected fear of missing a lesion as their primary cause of anxiety, while 23.8% (40/168) selected personal/intimate nature of the exam.

The authors found no significant differences between male and female students on chi-square analysis for any of the eight variables in part one of the first form of the survey. Analysis of responses from the second part of Form 1 of the survey identified male students as having significantly less anxiety than females for detecting abnormalities. The mean baseline comfort level for male students while detecting abnormalities was 3.10 (SD = 0.84) while female students’ mean was lower at 2.62 (SD = 0.71), p < .001. This gender difference became insignificant after training on the stations and remained so after training with the standardized patient. Interestingly, the outcomes of the ANOVA analysis indicated statistical differences among the students who selected different top causes of anxiety. When comparing the two groups of students who selected fear of missing a lesion, or intimate/personal nature of the exam as their top cause of anxiety, the authors found a significant difference between these groups when examining the second part of the form that asked the students about their comfort level performing specific tasks of the breast examination. Across all forms, the students who selected fear of missing a lesion (n = 84) had significantly higher comfort ratings for the overall rating, and visual inspection than those that selected intimate/personal nature of the exam. This significant difference continued until after training using the simulators. This indicates that for students who select intimate/personal nature of the exam have the most difficulty performing the visual inspection of the breast during the clinical breast exam.

In a parallel study, Pugh, (2009) measured medical student anxiety as it relates to the female pelvic examination. Using similar teaching methods and research design as the
previous study, the authors measured student anxiety during the academic years 2004 and 2005 for two classes of first-year medical students ($N = 344$) at a large Midwestern research institution. Different from the previous study, the author distributed only two surveys. One form of the survey was distributed before training as the baseline, the second after the students performed a pelvic exam on a simulator pelvic model. The forms were developed based on a combination of focus groups with a general practitioner, teaching faculty, and standardized patients. The first section of the first form (Appendix H) requests that students rank their top three causes of anxiety toward the pelvic examination from a selection of six items, including (a) causing harm or pain to the patient, (b) personal/intimate nature of the exam, (c) use of the speculum, (d) religious beliefs, (e) general performance anxiety, and (f) touching the genitalia. “I have no anxieties relating to this exam” was also supplied, as well as a fill-in-the-blank option, “Other,” which allowed the students to write in their own concerns.

The second part of first form rated students’ overall comfort and comfort performing specific tasks of the pelvic exam, including (a) overall comfort, (b) visual inspection of the female perineum, (c) speculum insertion and visualization of the cervix, and (d) performing a bimanual exam and palpation of the uterus. The items were rated on a modified, 6-point, Likert-type scale, with categories ranging from 1 (Extremely anxious) to 6 (Very comfortable). The second form was dispersed to the students after they were presented with a variety of information associated with the female pelvic examination via three stations that included visual aids, specula and simulated pelvic models.

The first part of the scale asked the students to rate the usefulness of the training materials, while the second part was the same items from the first form, measuring students’ anxiety toward specific tasks of the pelvic examination. Reviewing the frequency data,
49.7% (151/304) of the students selected *causing harm or pain* as their primary cause of anxiety, while 25.7% (78/304) selected *intimate/personal nature of the exam*. This was the top cause of anxiety for both genders. Mean comfort levels were computed for the two forms, and analyses included a repeated measures ANOVA and chi-square test for independence to identify gender differences. There were no significant differences across the two groups of students (2004 vs. 2005), but the authors suggest that chi-square test of independence indicated there were significant gender differences, reporting a p value of, \( p < .05 \). No test statistic was reported.

Combining the years, of the 143 female students, 87 (60.8%) selected *causing harm or pain*, while of the 161 male students, 64 (39.8%) selected this item, indicating a significant difference between the genders for this item, \( p < .001 \). However, more male students 51 (31.7%) selected *intimate/personal nature of the exam*, while 27 (18.9%) selected this item, indicating a significant gender difference for this item, \( p < .05 \). Results from the second part of Form 1 indicated that all students were least comfortable with the speculum examination (Pugh et al., 2009). Reviewing data from Form 2, after students performed the pelvic examination on simulator pelvises, indicated that there were no gender by time interactions but male students (\( M = 3.65, SD = .98 \)) continued to rate themselves significantly more comfortable than female students (\( M = 3.31, SD = 1.08 \)) for the speculum examination. This finding was consistent for the bimanual examinations, and male students (\( M = 3.47, SD = .99 \)) rated themselves significantly more comfortable than female students (\( M = 3.11, SD = .98 \)), regardless of timing and training experience. A chi-square test of independence supported this finding and authors reported a p value of \( p < .01 \). No test statistic was reported.
The authors suggested that influences of gender may have important curricular implications for male and female students while they learn the female pelvic examination. Although Pugh, et al. (2009) presented pertinent and interesting information as it relates to this work, they did not deliberately attempt to offer evidence of validity to use the scale for measuring medical student anxiety during these intimate examinations. In spite of this, there is some validity evidence I will point out, and potentially apply to this study. First, the evidence relevant to content aspect of validity was suggested by Pugh et al. as the items of the two forms of the scale were developed with a logical analysis of the content of the theoretical construct, utilizing experts in the field, teaching faculty and standardized patients. For this study, the development of appropriate items may have been supported by interviewing medical students from a variety of levels. To offer validity evidence, I suggest further analyses to include evaluation of evidence relevant to external validity, item analysis to offer evidence relevant to structural validity, and utilizing the scale in a multiuniversity trial with other medical schools, including osteopathic programs, to evaluate evidence relevant to the generalizability aspect of validity.

Measurement of Medical Students’ Anxiety Toward the Male Genitourinary Rectal Exam

In spite of increased interest toward measuring medical students’ anxiety toward intimate exams, studies that have attempted to measure medical students’ anxiety toward the male GUR exam have been sparse. To date, a review of the literature revealed only four studies that have related to measuring medical students’ attitudes toward the male GUR
examination. Two examined student confidence, while the other two focused directly on measuring student anxiety. A review of these studies may help guide the development of a new scale that measures medical students’ anxiety toward the male GUR exam.

Hennigan et al. (1991) were the first to examine the attitudes of medical students toward the male GUR exam, particularly the male rectal exam, a key task of the male GUR exam. In this study, the authors administered a questionnaire to 119 medical students in London. The students were asked if they learned to perform the rectal examination in a variety of conditions, including consultant teaching, formal teaching, and teaching on anesthetized patient. Additionally, the students were asked the number of gastrointestinal or urological specialties in which they received training. The students then rated their confidence at performing the rectal exam in these conditions. Students rated their confidence on a 6-point Likert-type scale, ranging from 0 to 5. Scores were dichotomized around the median confidence score, and scores between 0 and 3 were labeled “no,” with a value of 0, and scores between 4 and 5 were labeled “yes,” with a score of 1. Total scores were summed, and analyzed using Kendall’s τ-c test. Kendall’s τ is a measure of correlation between ordinal variables that have been ranked. The value of Kendall’s τ ranges between -1.0 and +1.0, and represents a correlation that identifies the difference between the probability that observed data are in the same rank order versus the probability the observed data are not in the same rank order (Kendall, 1948). Presented in detail by Everitt (1977) and Siegel and Castellan (1988), there are three types of Kendall tau tests, τ-a, τ-b, and τ-c. For the purpose of this study, Kendall τ-c was likely used to accommodate for any tied pairings. In addition to the Kendall’s τ test, factors that may have deterred students from performing a rectal exam were also examined.
Reviewing frequency data, the median category for number of rectal exams performed was 11 to 30, while 23 (19.3%) students had performed fewer than 10 examinations and 19 (16.0%) had never felt a rectal cancer. Additionally, only 32 (26.9%) students routinely performed rectal examinations. Most prevalent deterrent factors were (a) told not to do so by medical staff, (b) embarrassment, (c) refusal by the patient, and (d) lack of a chaperone. More importantly, the data suggested that the students who had performed more than the median number of rectal exams were more confident in their ability to diagnose rectal cancers ($\tau_C = 0.17, p < 0.05$), benign prostatic hyperplasia ($\tau_C = 0.15, p < 0.01$), prostate cancer ($\tau_C = 0.14, p < 0.05$), and anal fistula ($\tau_C = 0.16, p < 0.05$) when compared to the students who had performed fewer than the median number.

Although the study lacked validity evidence, it was notable as the first study that was concerned about measuring student attitudes toward the rectal exam, and paved the way for more research in the area.

Thirteen years later, Lawrentschuk and Bolton (2004) recreated Hennigan et al.’s (1991) study that examined the experience and attitudes of final-year medical students to the DRE. The study included 222 final-year medical students at the three schools associated with the University of Melbourne in Australia. The authors assessed attitudes toward the digital rectal exam (DRE) via a 27-item questionnaire. The questionnaire requested basic demographic data that included the number of DREs performed, and the number of different abnormalities palpated. Additionally, items relating to teaching conditions and perceived barriers to performing a digital rectal exam, as well as items measuring confidence and self-perceived competence, were included. Analyses included the chi-square test to identify differences between this study and Hennigan’s earlier study, while $t$ tests were performed in order to compare the three groups of students within the contemporary study.
Reviewing the frequency data, only 63% of the students had actually palpated a prostate, 24% felt a prostate malignancy, and 19% a rectal tumor. The median number of DREs performed by the three groups was 2, while 37 (17%) of the students had performed no DREs, and 17 (8%) had performed 10 or more. Although test statistics from the chi-square test were unpublished in this study, the authors reported nonsignificant differences in number of exams performed by school, gender, or age, p > .05.

Most importantly, results indicated that in spite of most (94%) students feeling that it was important to have the skills at performing the exam before practicing medicine, only 30% of the students felt that the school had not been supportive of teaching them how to perform the DRE. Additionally, nearly half (48%) of the students were not confident at diagnosing based on the DRE findings, while 50% of the students were reasonably confident, and only 2% were very confident. The primary student-reported barriers of performing the DRE were (a) not having a doctor chaperone (63%), (b) lack of perceived competence (41%), (c) patient refusal (39%), (d) too bothersome (35%), (e) perceived obstruction by doctors (30%) and nurses (22%).

The importance of this study is not in the reproducibility of Hennigan et al.’s (1991) study, but in that the results suggested that in spite of students’ agreement of the importance of performing the DRE in order to diagnose potential prostate and rectal cancers, a relatively low number actually performed them, and cited a variety of barriers. Although this study directly related to medical education in Australia, it lacked reliability and validity evidence, and failed to recognize perceived barriers to performing the male GUR exam for medical students in the United States.

Robins et al. (1997) examined the effects of implementing standardized patients in the male GUR exam skills curriculum on 190 second-year medical students’ anxiety and
confidence at the University of Michigan Medical School. A 19-item questionnaire was administered to the students immediately following a male GUR examination training session. The questionnaire included demographic questions, such as gender and ethnic identification, questions regarding their training experiences with the standardized patients, and students’ self-reported anxiety and confidence before and after the training sessions. Scoring was rated on a 5-point Likert scale. Analyses included unpaired $t$ test to compare male and female students and paired $t$ tests to compare pretest and posttest anxiety scores by gender and ethnicity. Consideration of effect sizes characterized practical significance of the differences. Results indicated that the use of standardized patients to practice the GUR exam techniques was well-received by the students, and increased their sense of confidence and competence. Most important for the purpose of my research, the study reported that students’ preexamination anxiety was considerable, as 83% of the students reported themselves as “somewhat,” or “very” anxious. Female students (4.13 ± 0.83) rated their anxiety substantially higher than males (3.25 ± 1.20). Female students rated their preexamination anxiety higher regardless of ethnic group, although the magnitude of difference was greater for European American, African American, and Hispanic American groups. Men who identified themselves as Asian and East Asian, Indian, or Pakistani reported significantly higher levels of anxiety than males of other ethnic groups. Postexamination anxiety decreased for all students, with the greatest improvement seen for the groups that had the highest preexamination anxiety, the males of the Asian and East Asian, Indian, or Pakistani groups. Although the study does not offer validity evidence, nor identify specific sources of potential anxiety for students while they learn and perform the male GUR exam, it does effectively present insight on the complexity of the factors that may affect medical students’ anxiety toward the male GUR exam.
More recently, a study by Howley and Dickerson (2003) examined the potential effects of the use of standardized instruction on medical students’ anxiety at the University of Virginia. The authors also investigated the relationship of gender and students’ anxiety. In order to measure the students’ anxiety, the authors utilized a 31-item scale that included the revised form of the state scale of the STAI (Spielberger, 1970), and an open-ended question that prompted the students to describe if and how the training had reduced their anxiety. The scale was administered to 102 students immediately preceding and following the students’ training session with standardized patients. Total scores were calculated for both preanxiety and postanxiety and ranged from *Little or no anxiety* (15) to *High anxiety* (60). Internal consistency was adequate and was identified using Cronbach alpha for both the prescales and postscales (alpha = .86 and .85, respectively).

Analysis included split plot repeated measures ANOVA to identify differences in means between preanxiety and postanxiety for male and female students. Significance was considered $p < .05$, while effect sizes ($\eta^2$) greater than .40 were considered large and less than .10 was small. Reviewing the within-subjects effects (over time) indicated that all students’ mean levels of anxiety surrounding their first male GUR examinations were significantly reduced following participation in the small group instructional session, $F(1,76) = 102.35, p = .001, \eta^2 = .57)$. Interaction effects were not significant. Examination of between-subjects (gender) effects indicated that female students reported their state anxiety prior to the training session significantly higher than their male counterparts, $F(1,91), = 6.96, p = .01, \eta^2 = .08)$. The students attributed this reduction to the format of the session and to the personal affects and demeanor of the teaching associates. These results are consistent with the research performed by Robins et al. (1997).
Howley and Dickerson’s (2003) results have more credibility as they attempted to utilize a scale (albeit revised) that is supported by sufficient validity evidence to measure medical student anxiety, and utilized robust analyses methods and reporting. The work by Howley and Dickerson is evidence of the continued interest in measuring medical students’ anxiety as they learn to perform the male GUR exam. Furthermore, it offered evidence that the negative influence of gender may continue to be a barrier to effective learning of the male GUR exam, and ultimately affect physicians’ decision to perform a GUR exam on a patient later in practice.

Summary of Instruments Used to Measure Medical Student Anxiety Toward Intimate Exams

Since the 1970s there has been a great deal of interest in the identification and measurement of medical students’ anxiety throughout their medical training. Many studies have focused on medical students’ generalized stress and anxiety in hopes of developing curricular improvements to alleviate or minimize the potential effects stress and anxiety may have on students’ learning. Other researchers have directed their attention to measuring student anxieties related to specific clinical-relevant tasks or training methodologies. The research most relevant to this study has focused on measuring students’ anxiety as they learned and performed intimate exams, and not generalized stress and anxiety. Because there are no existing studies that have examined the particular sources of medical student anxiety while they learn and perform the male GUR exam, I must rely on studies (Abrahams et al., 2003; Pugh & Salud, 2007) that have attempted to identify and measure the specific sources of student anxiety while learning and performing the female pelvic examination. In spite of the information presented by the authors, both of these studies failed to formally evaluate validity evidence. Although Howley and Dickerson’s (2003) study did not attempt to
identify specific sources of students’ anxiety, they did present effective research methodology that may be applied to this study. A review of the existing studies that have attempted to identify and measure specific sources of medical student anxiety revealed no studies that have done so for the male GUR exam. Building from the existing research relating to the female pelvic examination, we may gain ideas of how the researchers best identified the specific sources of student anxiety but, at the same time, we need to remember that each of the intimate examinations are associated with different sources of anxieties.

Theoretical Construct of the New Scale

The theoretical construct of the new scale has been directed by the literature review of existing scales used to measure anxiety in clinical studies and medical-education settings. The literature review helped direct the development of a new scale. I have defined anxiety as it related to my purpose, examined instruments commonly utilized to measure anxiety, and reviewed existing measures used to measure medical student anxiety toward intimate examinations. Finally, I presented the construct model for measuring medical student anxiety, a proposition-based approach to presenting reasonable validity evidence for the inferences from the new instrument, and identified the five relevant sources of this validity evidence.

The operational definition of anxiety used for this study is Eysenck et al.’s (2007), “an aversive emotional and motivational state occurring in threatening circumstances” (p. 336). This definition is founded on the concept of state anxiety as defined by Spielberger’s state–trait anxiety theory. Due to the complexity of the influencing factors on medical students’ anxiety, we cannot measure every factor. However, we can focus on the factors that may be easier to identify with a well-designed instrument, based on a theoretical foundation. Construct modeling by Wilson (2004, 2005) allows us to clearly define the
construct, the theoretical object of interest to be measured, and translate it into discrete items that will ultimately be used to measure that construct. The visual representation of this relationship is presented using a construct map. A construct map allows us to understand the relationship of subjects and the construct intended to be measured. The construct also reflects the relationship of subjects to each other along a continuum, as well as the subjects’ relationship with the items.

The construct map for my application (Figure 5) identifies the relationship of medical students and their anxiety level performing specific aspects of the GUR exam. Wilson’s framework of construct modeling contains three other building blocks that support the construct model. These are item design, outcome space, and the measurement model. Item design refers to the quality of the item to measure the intended construct. Well-designed items are appropriate to the targeted audience, and measure the intended construct. Although there are a variety of methods to develop items; for the purpose of this study item quality was ensured via an expert panel comprised of faculty, medical students, and associated health staff using the item writing guidelines presented by Wolfe and Smith (2007a). After multiple items had been developed, a pilot study with faculty experts and medical students helped identify problematic items. The final scale was administered to health professionals who teach and perform the male GUR exam as part of their clinical practice, as well as medical students with variable experience levels. Doing so supplied a heterogeneous sampling of anxiety levels that gave more information about the appropriateness of the items, given the response patterns of these individuals, and lent evidence relevant to the substantive aspect of validity.
Figure 5. Construct map of measuring medical student anxiety toward the male GUR examination.

The outcome space refers to the appropriateness of the categories associated with the items of a rating scale. According to Wilson (2005), a sound and useful outcome space should have rating categories that are “well-defined, finite and exhaustive, ordered, context-specific, and research-based” (p. 64). This concept is the basis for Research Question 2 and may lend evidence relevant to substantive aspect of validity.

Wilson (2004, 2005) best describes the measurement model as a means to relate the examinees’ responses from the outcome space to the construct of measurement. Earlier in this chapter, I discussed classical test theory and Rasch measurement models. For the purpose of this study, a Rasch model is best suited to produce measures of medical student anxiety from the ratings obtained from the new scale and locating the examinees’ responses on the construct map. The strengths of the Rasch model used in this study lie in its interpretability. Wilson points out the two primary requirements of a measurement model,
both offered by the Rasch model. The first is that the model requires invariant ordering of item responses for all respondents, enabling clear interpretation of distance between respondents and their responses on the construct map. Secondly, the Rasch model allows interpretation of the distance between respondents. Both conditions are essential for consistent, meaningful interpretation of the construct map, and fulfill Wilson’s requirements of a measurement model.

All four of the components reviewed here are crucial to this study where I am developing a new instrument intended to measure the factors that influence anxiety during particular tasks. A well-designed construct model that includes solid item development, well-defined outcome space, and a robust measurement model is crucial to support the validity of the inferences from the measures.

The literature review revealed a number of existing instruments used for measuring clinical anxiety. Although the authors of the instruments presented evidence of adequate psychometric properties, in many cases evidence was challenged, suggesting that the scales lacked sufficient validity evidence to support inferences from their scores. Additionally, none of the scales’ items address specific situations relating to the learning scenario in which medical students are presented. For these reasons, I suggest that a new scale should be developed and hopefully supported by sufficient validity evidence in order to measure medical students’ anxiety toward a male GUR examination. Validity propositions, the research questions from which they were derived, and the validity evidence used to test the proposition are listed in Tables 2–6.
### Table 2

*Validity Propositions, Research Questions, and Validity Evidence to Address Research Question 1: Does Evidence Based on the Content Aspect of Validity Support the Use of Student-Anxiety Scores to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?*

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Research Question</th>
<th>Validity Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. The items on the new instrument adequately represent the intended tasks performed during the male GUR exam.</td>
<td>1a. Do the items on the new instrument adequately reflect the tasks performed during the male GUR exam?</td>
<td>1a. Evidence from expert reviews from experienced clinicians, and medical faculty (expert panel) regarding the adequacy of the items to cover the range of tasks (Appendix A).</td>
</tr>
<tr>
<td>1b. The rating scale provides response options that adequately reflect the levels of students’ anxiety they may experience toward the male GUR exam.</td>
<td>1b. Does the rating scale provide response options that adequately reflect the levels of students’ anxiety they may experience toward the male GUR exam?</td>
<td>1b. (1) Evidence from expert panel regarding whether the response-options reflect the level of student anxiety may influence students’ anxiety toward these tasks. (2) Evidence regarding the appropriateness and clarity of items and response-option categories from expert review of experienced clinicians and medical faculty who were administered the new instrument. (3) Evidence from expert review from experienced faculty regarding the categories to ensure that they are ordered, well-defined, mutually exclusive, and exhaustive.</td>
</tr>
</tbody>
</table>
Table 2 (continued)

**Validity Propositions, Research Questions, and Validity Evidence to Address Research Question 1: Does Evidence Based on the Content Aspect of Validity Support the Use of Student-Anxiety Scores to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?**

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Research Question</th>
<th>Validity Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1c. Items on the new instrument form a cohesive scale that measures the theoretical construct.</td>
<td>1c. Do items on the new instrument fit together to provide evidence that only the intended construct is being measured?</td>
<td>1c. Evidence from Rasch analyses to support items contribute to a cohesive scale that measures theoretical construct.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) Evidence from Rasch unweighted mean-square item fit indices within the range of 0.6 to 1.5 supports items are measuring a common underlying construct.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) A positive point-measure correlation indicates that items contribute information to construct measured by the test as a whole.</td>
</tr>
</tbody>
</table>
Table 3

List of Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 2: Does the Evidence Based on the Substantive Aspect of Validity Support the Use of Student Anxiety Scores to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Research Question</th>
<th>Validity Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2b. Students use the rating scale appropriately to self-report their anxiety toward the male GUR exam (i.e., the items have categories that are ordered, well-defined, mutually exclusive, and exhaustive).</td>
<td>2b. Do students use the rating scale appropriately to evaluate and self-report their anxiety toward the male GUR exam, producing ordered, well-defined, mutually exclusive, and exhaustive categories on the rating scale?</td>
<td>2b. Evidence from evaluation of a number of observations, ordered average measures, category fit, and threshold ordering evaluation ensure categories are ordered, exclusive, and reflect intended use.</td>
</tr>
<tr>
<td>2c. Empirical item difficulties on the new instrument are consistent with those predicted from expert opinion.</td>
<td>2c. Does the empirical item hierarchy of the new instrument support an a priori item hierarchy?</td>
<td>2c. Evidence from theoretical item hierarchy compared to those predicted by experts.</td>
</tr>
<tr>
<td>Proposition</td>
<td>Research Questions</td>
<td>Validity Evidence</td>
</tr>
<tr>
<td>-------------</td>
<td>--------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>3. Items on the new instrument form a cohesive scale that measures the theoretical construct.</td>
<td>3. Do items on the new instrument fit together to provide evidence that only the intended construct is being measured?</td>
<td>3. Evidence from Rasch analyses to support items contributes to a cohesive scale that measures theoretical construct.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1) Evidence from Rasch unweighted mean-square item fit indices within the range of 0.6 to 1.5 supports items are measuring a common underlying construct.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) A positive point-measure correlation indicates that items contribute information to construct measured by the test as a whole.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) Principal component analysis (PCA) of the residuals of real and simulated data support unidimensionality.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(4) Evaluation of local independence using the Fisher’s Z statistic of item dependency.</td>
</tr>
</tbody>
</table>
Table 5

Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 4: Does the Validity Evidence Based on the Generalizability Aspect of Validity Support the Use of Student Anxiety Scores of the New Scale to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Research Question</th>
<th>Validity Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>4a. A number of facets contribute to explain variations in the data.</td>
<td>4a. Do gender and ethnicity facets contribute to explain variations in the data collected?</td>
<td>4a. Evaluation of AIC, BIC, and chi-square statistics.</td>
</tr>
<tr>
<td>4b. The empirical ratings of students are reproducible over different items.</td>
<td>4b. Are the empirical ratings of students reproducible over different items?</td>
<td>4b. Evaluation of the internal consistency reliability of students using separation reliabilities, and separation indices (or strata).</td>
</tr>
<tr>
<td>4c. The empirical ratings of items are reproducible over different persons.</td>
<td>4c. Are the empirical ratings of items reproducible over different students?</td>
<td>4c. Evaluation of the internal consistency reliability of items using separation reliabilities, and separation indices (or strata).</td>
</tr>
</tbody>
</table>
Table 6

*Validity Propositions, Specific Research Questions, and Validity Evidence to Address Research Question 5: Does the Validity Evidence Based on the External Aspect of Validity Support the Use of Student Anxiety Scores of the New Scale to Make Inferences About Medical Students’ Level of Anxiety Toward the Male GUR Exam?*

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Research Question</th>
<th>Validity Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. The empirical ratings of students correlate with empirical ratings of another instrument used to measure anxiety.</td>
<td>5. Do the empirical ratings of students correlate with empirical ratings of another instrument used to measure anxiety?</td>
<td>5. Comparison of student group ratings across the A-GUR and the 6-item STAI.</td>
</tr>
</tbody>
</table>
III. METHOD

Research Design

I conducted this study in two phases. The results of a 2005 pilot study involving the administration of a self-report attitude survey toward performing the male genitourinary rectal (GUR) exam was used to develop a new rating scale (A-GUR). The second phase of the study was a prospective study involving the gathering of validity evidence to support the use of scores from the A-GUR to infer the source and level of medical students’ anxiety toward performing the male genitourinary rectal exam.

In the first phase of the study I addressed research question 1: Does evidence based on the content aspect of validity support the use of student-anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam? During the development of the scale I utilized medical students’ responses from the initial scale developed for a pilot study, and incorporated evidence acquired from two expert panels comprised of the clinical skills coordinator, experienced clinicians, medical faculty, and medical students and residents to ensure that the items adequately covered the range of tasks associated with performing the male genitourinary rectal exam discussed in Appendix A. This offered validity evidence that supports research propositions 1a: The items on the new instrument adequately represent the intended tasks performed during the male GUR exam, and 1b: The rating scale provides response options that adequately reflect the levels of anxiety a student may perceive toward performing the male GUR exam as shown in Table 2 in the previous chapter. In doing so, the panel offered evidence regarding the adequacy of the items to cover the range of tasks associated with the male GUR exam, and adequacy of
the items to cover the levels of medical students’ anxiety associated with those tasks, as well as the appropriateness and clarity of items.

The second phase of the study involved data collection and two analyses. After collecting data, the first analysis addressed research question 4a: Do gender and ethnicity facets contribute to explain variations in the data collected? In order to do this, I utilized the Facets model to evaluate the contributions of the facets (gender and ethnicity) to explain any variations in the data collected. This model-building approach, using AIC, BIC, and chi-square statistics, was done to identify the form of the Facets model that works best with the data. The subsequent analyses utilized the most appropriate model identified by the AIC, BIC, and chi-square statistics.

The second analysis included the prospective gathering of validity evidence to support (or refute) the use of the A-GUR scores after the implementation of the rating instrument. I evaluated the validity evidence using the framework offered by Wolfe and Smith (2007a, 2007b) as a guideline.

**Study Setting**

I conducted the study at the medical schools of two major universities located in the Midwest; throughout this document these medical centers will be referred to as University A and University B. University A is an educational center that provides services involving simulated clinical encounters for medical training, assessment, and research for all medical students and residents of the university’s medical school as well as some other nearby medical schools. University B offers similar services.

For the development of the new scale, I scheduled two focus-group sessions. The first included members from both institutions. The first focus group consisted of clinical
skills coordinators, faculty who teach the male GUR exam to medical students, and patient
instructors. The second focus group included a sampling of medical students and junior
residents. For both sessions I used the A-GUR Content Validity Rating Form (Appendix I)
to ensure items are appropriate and clearly-written, and response options adequately reflect
the levels of anxiety a student may perceive toward performing the male GUR exam.

The objective of the A-GUR was to evaluate students’ level of anxiety toward
performing the male GUR exam. For the students of both universities, the A-GUR was
administered immediately preceding the male GUR clinical skills training sessions. Both
institutions employ paid patient instructors (PI) that review the steps in performing the
exam, as well as act as a “patient” for the students to practice their newly learned skills.

**Participants**

Based on two separate focus-group sessions, the first phase of the study was the
development of the A-GUR. One session with the expert panel consisted of the clinical skills
coordinator, experienced clinicians, medical faculty, and patient instructors; the other with
learners consisting of medical students and residents from both universities. The final item
development of the A-GUR scale was directed by the feedback from the expert and learner
panels and results from a 2005 pilot study that measured anxiety toward the male GUR
exam of medical students from a large Midwestern research institution administered during
a clinical skills course in 2005.

I anticipated that participation rate for the first- and second-year medical students
would be similar to participation rate during the aforementioned 2005 pilot study. Although
the Feinberg School of Medicine requires every medical student to take the clinical skills
course that includes the session on the male GUR exam, student participation in the pilot
study was elective. In spite of this, more than 95% of the students completed the scale used for the pilot study. Of the 331 students that completed the scale, 170 were in the first year, while 161 were in the second year. From the 169 records with gender identification, there were 85 (50.3%) male students and 84 (49.7%) female students. To keep the identity of each participating student confidential, I did not collect any information about age and/or race/ethnicity. Due to interest in gender and ethnicity effects on students’ anxiety, in the new study, unlike the pilot study, these data were collected. In order to maintain confidentiality of participating students during the second phase of this study, no other identifying information was collected. This phase of the study involved the administration of a new scale, the A-GUR, to approximately 521 medical students. Because student anxiety toward the male GUR exam is highest during the junior years of medical training, I collected data from first- and second year medical students. In order to maximize the sample, I collected data from 169 second-year medical students at University A and 299 first- and second-year medical students from University B. Because the first-year students of University A do not participate in clinical skills examination training, they were not requested to participate in this study. Among the 169 second-year University A medical students, 88 (48.3%) were male, and 96 (51.7%) were female. Although all of the students of University A participated in the study, only 32 (18.9%) of these students completed the STAI-related items due to unrecognized misprinted form that did not contain these items. The mean age of the students of University A was 22.8, with a standard deviation (S.D.) of 2.9. Among the estimated 337 University B medical students, 168 were first-year, and 169 were second-year medical students. This group had approximately 182 (54.0%) male and 155 (46.0%) female students. The mean ages of the students of University B were 24.2 (S.D. of 2.8) and 25.0 (S.D. of 2.4) for the first-year and second-year students, respectively. The combined mean age for
University B was 24.6. All students of University B completed all items of the A-GUR, including the STAI-related items.

**Sample Sizes**

Before collecting data, assessment of the sample size was performed in order to ensure a balance between feasibility of the study and stable results. For this study, I evaluated whether the sample size that I had from a 2005 pilot study and the sample size that I had in the 2010-2011 trial were large enough to provide stable item difficulties. In this study, I assessed the adequacy of the sample sizes by considering two aspects of the analysis: the stability of item difficulty (Linacre, 1994), and the functioning of the rating scale (Linacre, 2004b).

According to Linacre (1994), the determination of a sample size that can provide stable item difficulty involves making two key decisions. First, we must determine the largest variability in the item difficulty that is tolerable for the measurement purpose. Secondly, the desired confidence interval should be determined such that the item difficulty will be within the range of tolerable variability. According to Linacre, even a dichotomously-scored test with fewer than 30 items can have random deviations up to 0.5 logits and still be practically bias free (Linacre, 1994). Using desired confidence intervals and level of stability of item difficulty, the required sample size for this study was calculated. Consistent with tradition found in the literature, I selected a 95% confidence interval in the calculation of required sample size (Neyman, 1937; Zar, 1984). In practical terms, I intended to have a sample size large enough to have 95% confidence that no item difficulties are more than 0.3 logits away from their actual value. Given this, an adequate
sample size was large enough to produce item difficulties with standard errors (SE) shown below,

\[ 0.3 \text{ logits} = 1.96 \text{ SE} \]  

\[ \text{SE} = \frac{0.3}{1.96} = 0.153 \text{ logits}. \]  

Typically, a larger sample size will result in a smaller standard error. The standard error is also dependent on how well the items are targeted to the students’ anxiety. For a test that has items reasonably targeted to the students’ anxiety, its standard error is shown by the formula,

\[ \frac{2}{\sqrt{N}} < SE < \frac{3}{\sqrt{N}}. \]  

Here \( N \) represents sample size (Linacre, 1994) with the terms to the left and right identifying the lower and upper boundaries of the standard errors. The lower boundary suggests items are well-targeted to the students’ level of anxiety, while the upper boundary suggests items may be too difficult (or too easy, depending on coding) for learners to endorse. Using this formula, an adequate range for sample size can be estimated. According to Linacre, for a 95% confidence interval and item difficulties stable within 0.5 logits, a sample size ranging between 62–139 is required for dichotomously scored response-options scales (Linacre, 1994). This calculation is shown below as,

\[ \frac{2}{\sqrt{N}} < 0.255 < \frac{3}{\sqrt{N}} \]  

(13)
\[
\frac{4}{0.065} < N < \frac{9}{0.065}
\]
\[
62 < N < 139.
\]

Modifying this calculation to use .3 logits, the calculation is shown as,

\[
\frac{2}{\sqrt{N}} < 0.153 < \frac{3}{\sqrt{N}} \quad (14)
\]

\[
\frac{4}{0.023} < N < \frac{9}{0.023}
\]
\[
174 < N < 391.
\]

Given this calculation, a minimum of 35% of the estimated 520 medical students planned for in the 2010-2011 would be considered a sufficient sample size to provide reliable item difficulties as planned, assuming the data fit the model. Given that it was unlikely that the data would perfectly fit to the model, all attempts were made to maximize the sample size to accommodate any removal of misfitting data. Returning to the computation above, the calculation suggests that I should have adequate sample sizes for dichotomously-scored items, but we must also consider that my study utilized polytomously-scored items. Linacre (2004b) suggested that to estimate stable thresholds, there should be a minimum of 10 observations for each rating-scale category. Further, the minimal number of students required to produce a properly functioning rating scale depends on the complexity of the Rasch model employed in the analysis and targeting of the items to the level of anxiety of the residents. This study employed the rating scale model (equation 4). Given that most medical educational researchers utilize scales with 4 or more rating categories (Howley & Dickerson, 2003; Pugh & Salud, 2007; Pugh, Obadina, & Aidoo, 2009) the new scale
utilizes items with a 5 rating categories. Assuming this, I needed 5 rating categories x 10 observations per category = 50 observations, assuming ratings were uniformly distributed, for stable estimates of thresholds. Given this calculation, the estimated sample size of 520 medical students planned for in the 2010-2011 trial provided sufficient sample size to satisfy the required numbers of observations to estimate stable thresholds in the Rasch model I utilized.

**Instrument**

The *Anxiety toward the Genitourinary Rectal Examination* (A-GUR) is an instrument administered throughout the 2010-2011 academic year and based on the *Male Urogenital Student Evaluation Form* which was administered during a pilot study Fall of 2005.

The *Male Urogenital Student Evaluation Form* is a self-reported instrument that consists of two sections. The first section, *Section A*, prompts the students to rank their top three anxieties by selecting them from nine fixed options and one fill-in-the-blank option, *Other*. The nine options include (a) causing harm or pain to the patient, (b) personal/intimate nature of the exam, (c) self-embarrassment, (d) embarrassment of patient, (e) general performance anxiety, (f) causing patient erection, (g) touching the genitalia, (h) sanitation issues, and (i) no anxieties relating to this exam. The second section, *Section B* consists of five, 6-point Likert-type items with categories ranging from 1 (*Extremely anxious*) to 6 (*Very comfortable*). Students were prompted to rate their comfort in performing specific procedures of the male urogenital rectal examination via the following questions:

1. Please rate your comfort level in performing a clinical male urogenital exam.
2. Please rate your comfort level in performing visual inspection/palpation of the penis/testes.
3. Please rate your comfort level in performing hernia assessment.

4. Please rate your comfort level in performing visual inspection of the anus.

5. Please rate your comfort level in performing a digital rectal/prostate examination.

The administration of this instrument during a 2005 pilot study at University B was performed immediately prior to the first- and second-year students’ experience performing the male genitourinary rectal examination on a patient instructor, a person paid to act dually as an instructor and patient.

Analyses of the data collected during the pilot study included testing of the monotonicity, an assumption of the Rasch Model that supports unidimensionality, evaluation of item and person fit, differential item functioning (DIF; Wright & Stone, 1979), and evaluation of person and item reliabilities. The assumption of monotonicity was found to be upheld as the expected response functions behaved ideally, indicating no decreases in the probability for success with the increase in score. Analysis of person fit indicated there were nine misfitting persons. These persons were removed from the analyses. Analysis of item fit indicated that there was one misfitting item: 1-Clinical Male Urogenital Exam, with an Outfit Z-Standardized of -4.3. Looking at the relationships between persons and items, it was determined that they ranked in order of least confidence to most confidence as:

(Least Confidence)
   1. Digital Rectal Examination
   2. Overall Exam
   3. Hernia Assessment
   4. Visual Inspection of Anus
   5. Visual Inspection of Penis/Testes

(Most Confidence).
Differential item functioning analysis was performed to identify any differences in item functioning across four groups, defined by their self-reported top cause of anxiety: Group 1 (Harm to patient), Group 2 (Intimate nature), Group 3 (Embarrassment), and Group 4 (Performance anxiety). Analysis identified a statistical difference between Groups 2 and 4 for a single item (Item 3-Hernia Assessment), with a Group 2 item difficulty of -.38 (95%CI: -.77 to 0.01 logits), and Group 4 item difficulty of 0.96 (95%CI: 0.12 to 1.80 logits). Although there was some evidence to suggest that the instrument used in the 2005 pilot study was fairly effective at measuring the intended latent trait with reasonable reliability, improvements of the items and additional analyses may improve the function of the scale and better evaluate the evidence relevant to the content, substantive, structural, and generalizability aspects of validity introduced in Chapter 2.

The Anxiety Toward the Genitourinary Rectal Examination (A-GUR) instrument that was administered during the academic year of 2010-2011 is a modified version of the initial instrument used in the 2005 pilot study. Limitations of the pilot study are addressed in this study in two phases. Phase 1 of this study was used to evaluate the evidence relevant to the content aspect of validity. Although some evidence relevant to content aspect of validity may have been addressed in the pilot study, documentation that may have identified supporting processes was lacking, and therefore impossible to evaluate. Another limitation of the pilot study is addressed in this study. Items of the new instrument were rewritten to reflect the theoretical basis on which the instrument was founded. Unlike the instrument administered in 2005, this instrument specifically measures the students’ state anxiety toward each task associated with the male GUR exam. There were additions to the original instrument to better evaluate validity evidence. For example, the form has a section requesting demographic information such as gender and ethnicity, so that evidence of
generalizability aspect of validity can be evaluated. Additional items were added with increased specificity to better evaluate students’ anxiety toward the specific aspects of the male GUR exam, as well as to improve evaluation of substantive aspect of validity. Further, I incorporated items from the STAI Inventory (STAI; Spielberger, et al., 1970), in order to evaluate evidence relevant to the external aspect of validity.

Phase 2 was used to evaluate other evidence relevant to the content validity evidence, as well as substantive, structural, generalizability, and external aspects of validity. Although some evidence relevant to the substantive aspect of validity in the form of person-fit and category function analysis was evaluated in the 2005 pilot study, additional person fit and category function statistics were evaluated during the current study. Also, little evidence relevant to the structural aspect of validity was evaluated during the 2005 pilot study. Therefore, a number of additional statistics were evaluated to potentially offer evidence relevant to the structural aspect of validity. I evaluated Rasch outfit mean-square fit indices, point-measure correlations, principal component analysis of the residuals, and Fisher’s Z statistic for evaluation of local independence. Although the evaluation of item reliability and differential item functioning (DIF) across subgroups offered some evidence relevant to the generalizability aspect of validity for the 2005 pilot study, added evaluation of the internal consistency reliability of students using separation reliabilities and separation indices (or strata) offered added evidence relevant to the generalizability aspect of validity in this study. These statistics and the methods associated with evaluating these statistics are described in greater detail in the Procedures section of this chapter.

Procedures
In this section, I describe the procedures for administering the self-reported instrument, *The University Male Urogenital Student Evaluation Form* at a large Midwestern research institution, during Fall semester 2005, as well as the procedures used for administering the A-GUR during the 2010-2011 academic year. I explain the procedures used in the 2005 session first, and then explain the procedures that will be used during administration of the A-GUR throughout the 2010-2011 academic year. Additionally, I have highlighted variant procedures across the two administrations.

The self-reported instrument, *Male Urogenital Student Evaluation Form* (Appendix H) was developed and implemented at a large Midwestern research institution during Fall semester 2005. The sample consisted of 170 first-year (M1), undergraduate medical students. Of that sample, 86 students were male and 84 were female. A paper form of the instrument was provided to the students during clinical skills training. Clinical skills training spanned five days that rotated approximately 40 students through three training experiences that included (a) watching a video on how to perform the male genitourinary rectal exam, (b) rotating through a number of training stations that reviewed the aspects of the male GUR exam, and finally, (c) participating in small group training sessions with a paid patient instructor (PI). Students were assigned to groups of three or four students and rotated through each training experiences. The students were requested to complete the corresponding form of Student Evaluation at three points during their training—prior to any of the training experiences, after viewing a video demonstration of the GUR examination, and third, after performing the GUR examination on the patient–instructor. The completed forms were collected as the students exited from their experience with the patient–instructor.
Only the first form was used for the pilot study as I was interested only in the students’ baseline anxiety measure.

There were a number of procedural changes for the 2010-2011 administration of the A-GUR. First, the instrument was administered to a broader sampling of medical students, and expanded to include second-year medical students from University A. As discussed earlier, the first-year medical students at University A do not participate in clinical skills examination training. Because of this, only the second-year students were requested to participate during clinical skills training scheduled October 16, 23, 30 and November 18, 2010. The students were asked to complete a paper form of the A-GUR immediately preceding their training session with the patient instructor.

At University B, the first- and second-year students were requested to participate during clinical skills training scheduled for February 28 through March 3, 2011, and December 13 through December 18, 2010, respectively. Similar to the administration for the first- and second-year students of University B, the University A students were asked to complete a paper form of the A-GUR preceding their training session with the patient instructor. In an attempt to best capture accurate student state anxiety measures, all attempts were made to administer the A-GUR immediately prior to the students’ performing a male genitourinary rectal examination for each of the administrations of the A-GUR.

**Analyses**

In order to address the research questions I posed in Chapter 1, I conducted a series of analyses, performed in two phases. In this section, I describe the steps of the analyses that I conducted to address the specific research questions associated with each of the five areas of inquiry.
Phase 1

Research Question 1: Does evidence based on the content aspect of validity support the use of student-anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

1a. Do the items on the new instrument adequately reflect the tasks performed during the male GUR exam?

1b. Does the rating scale provide response options that adequately reflect the levels of students’ anxiety they may experience toward the male GUR exam?

To address each of these research questions, I utilized validity evidence offered by two focus groups with expert panels. The first focus group, comprised of experienced clinicians and medical faculty, reviewed the adequacy of the items to cover the range of tasks associated with the male GUR exam. The same focus group was asked review the quality of the items of the instrument. In order to address research questions 1a and 1b, the expert panel was asked to judge the adequacy of the items to cover the tasks associated with the male GUR examination, as well as judge the clarity and appropriateness of the items, and the adequacy of the response-options to reflect the level of student anxiety toward the male GUR exam. A second focus group, comprised of medical students, was asked to judge the adequacy of the items to cover the tasks associated with the male GUR examination. I used the A-GUR Content Validity Rating Form, seen in Appendix I, for this process.

1c. Do items on the new instrument fit together to provide evidence that only the intended construct is being measured?

Two statistics from Rasch analysis may be used to offer evidence that the items of the A-GUR instrument combine to form a single cohesive construct. First, Rasch outfit
mean-square item fit that fall within 0.6 and 1.5 support that items are measuring a common underlying construct. Additionally, positive point-measure correlations also support that the items contribute information to the construct measured by the test as a whole.

**Phase 2**

The first part of phase 2 of the analysis was utilized to evaluate the contributions of gender and ethnicity to explain the variations in the data collected using the Facets model, and address research question 4(1).

The Facets model I used for this study is the 4-faceted model which takes the following form:

\[
\sum_{k=0}^{m_i} \sum_{t=0}^{m_j} \sum_{l=0}^{m_l} \left( \theta_n - \delta_i - \gamma_j - \lambda_l - \tau_k \right) 
\]

where \( p_{nijkl} \) is the probability that student \( n \) with ethnicity \( j \) and gender \( l \) will have response rating of \( k \) on item \( i \), and \( k \) is the rating category on a rating scale that has \((m_i + 1)\) categories, \( k \in \{0, ..., m_i\} \), \( \theta_n \) is the anxiety level of student \( n \), \( \delta_i \) is the difficulty of item \( i \), \( \gamma_j \) is the effect of ethnicity \( j \), \( \lambda_l \) is effect of gender \( l \), and \( \tau_k \) is the difficulty of the \( k^{th} \) threshold.

Researchers can transform the equation above into a log-odds form for modeling rating data. The commonly used format for the facets model, the log-odds equation, is expressed as:

\[
P_{nijkl} = \frac{\exp \left( \sum_{k=0}^{m_i} \theta_n - \delta_i - \gamma_j - \lambda_l - \tau_k \right)}{\sum_{t=0}^{m_i} \sum_{k=0}^{m_i} \exp \left( \sum_{k=0}^{m_i} \theta_n - \delta_i - \gamma_j - \lambda_l - \tau_k \right)}, \quad (15)
\]
\[
\ln \left[ \frac{(p_{nijkl})}{(p_{niji(k-1)})} \right] = \theta_n - \delta_i - \gamma_j - \lambda_l - \tau_k. \tag{16}
\]

This equation illustrates the additive linear quality of the Facets model that is based on a logistic transformation of observed ratings to a log-odds scale (Engelhard, 2002). The log-odds scale is a common scale of measurement for the four facets where they interact with one another simply by addition or subtraction of their effects to (or from) the net effect—the log-odds of ethnicity and/or gender effects. A student may be likely to self-report their anxiety toward the male GUR exam as high if he/she is highly anxious, and is of a specific ethnic group and gender. On the other hand, a student may self-report their anxiety toward the male GUR exam as low if he/she is less anxious, and is of another ethnic group and/or gender.

To address research question 4a I utilized the Facets model to evaluate the contributions of the facets (gender and ethnicity) to explain the variations in the data collected. This model-building approach was done to identify the form of the Facets model that worked best with the data. I employed Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Kass & Raftery, 1995; Raftery, 1995; Schwarz, 1978). Both the AIC and the BIC are model selection approaches that combine the evaluation of data fitting with a penalization for model complexity. The AIC is a measure of goodness of fit of an estimated statistical model that combines the evaluation of data fitting with a penalization for model complexity, and is written as:

\[
AIC = -2(\text{loglikelihood}) + 2P_d, \tag{17}
\]
where \( P_d \) is the number of parameters in the model. The first term is the log likelihood of the best fitting curve of the considered model, which can be viewed as a measure of the fit between the model and evidence. Researchers call this term deviance. The latter term is a measure of model complexity. This term may be viewed as a penalty for overparameterization. In the event two models fit the evidence approximately equally well, researchers are advised to select the simpler model, indicated by a lower value (Kieseppa, 2002; Lin & Dayton, 1997).

The BIC is a criterion for model selection among a class of parametric models with different numbers of parameters. It takes the form:

\[
BIC = -2(\text{log likelihood}) + P_d \ln(N),
\]

(18)

where \( P_d \) is the number of parameters in the model, and \( N \) is the number of observations.

Similar to AIC, BIC evaluates the fit between the model and evidence with a deviance term. However, it uses a different penalty term for overparameterization. Because the BIC introduces a penalty term for the number of parameters in the model, and takes the number of observations into account, its penalty term increases with sample size. Thusly, BIC accommodates “overfitting” data while avoiding the selection of more complex models. The BIC approach favors simpler models over those that the AIC approach (Kass & Raftery, 1995; Kuha, 2004).

Additionally, a number of chi-square statistics generated from the FACETS program (Linacre, 1994) will be reviewed to help identify whether the facets (gender and ethnicity) have an effect on the proposed model, and should be included in that model. For example, a
“random” (normal deviate) chi-square is possible for each facet included in a model, and takes the form

\[ X^2 = \sum (Z_i^2), \]

(19)

where \( L \) is the number of elements in the facet, and \( Z \) is the standardized residual values (Shumaker & Lunz, 1997). I will also review the data-to-model global “residual” fit chi-square test to evaluate whether the facet has an effect on the model. If the facet does not have an effect on in the model, there should be little difference in the compared chi-square values.

When the AIC, BIC, and chi-square statistics suggest the same model, researchers have strong evidence to support the use of that particular model in the analysis of the data. However, when the AIC statistic conflicts with the BIC or the chi-square statistic, researchers may be able to narrow the choice of measurement models, but also have to consider other factors in choosing an appropriate model, such as the complexity of the analysis, the soundness of the psychological process involved in ratings, and the fit of individual students, items and facets. I used this process to determine the appropriate measurement model used in the later analyses.

1. **Research Question 2.** Does evidence based on the substantive aspect of validity support the use of student anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

2a. Do students exhibit aberrant response patterns indicating problematic response processes?
In order to evaluate student response processes, I conducted a Rasch analysis of the 2010-2011 A-GUR ratings, employing the rating scale model using the WINSTEPS® computer program (Linacre, 2004a). I evaluated outfit mean-square fit statistics. In order to determine aberrant student responses, I considered students with mean-square outfit statistics greater than 1.5 as underfitting, and those with mean-square outfit statistics less than 0.6 as overfitting. After identifying any misfitting students, I examined the response patterns of these students in order to identify the misfitting student ratings and possible explanation of response misfit. Because under- or overfitting students can potentially degrade the measurement process (Linacre, 2002), I followed the commonly-utilized practice of simply removing the misfitting students from the analyses. Because the removal of misfitting students can decrease the power of the fit statistics, alternative treatments of misfitting students should be also considered. A commonly-utilized treatment of misfitting persons is the removal of the offending persons from the analyses in a process described by Linacre (2010a). The method includes multiple reiterations of a 2-step process that includes the removal of the most extreme misfitting persons and a subsequent comparison of estimated person measures before and after the removal of the offending persons. This process continues until the last comparison of estimated person measures shows no obvious differences from the prior estimates, indicating that the fit criteria are adequately met. In this study, I treated any misfitting persons in this way. Finally, there is another consideration in regards to the removal of misfitting persons. After missing persons are removed as described above, the result may be a significant decrease in power of the fit statistics. This was not the case in this study, so I did not considered expanding the measurement model.

Along with the treatment of misfitting data, we must also consider the treatment of potential missing data during analyses. Just as the removal of misfitting persons can
decrease the power of the fit statistics, missing data can decrease researchers’ ability to estimate reasonably precise inferences and detect and evaluate any discrepancies between observed and expected observations. There are a number of methods to treat missing data generated from a rating scale. The two most common treatments include: (a) assume that data are missing at random (MAR) or missing completely at random (MCAR) and perform no special treatments, or (b) the removal of all incomplete records. In order to maximize my ability to estimate reasonably precise inferences using the Rasch measurement model I considered a third treatment that combines the two opposing treatments described above. I compared the item fit statistics of the entire sample that included the responses with missing data to the smaller sample with the missing data removed. If significant differences were found, I removed the responses with the missing data.

2b. Do students use the rating scale appropriately to self-report their anxiety toward the male GUR exam, producing ordered, well-defined, mutually exclusive, and exhaustive categories on the rating scale?

In order to evaluate students’ use of the rating categories, I conducted a Rasch analysis of the 2010-2011 A-GUR ratings, employing the Rasch rating scale model using the WINSTEPS® computer program (Linacre, 2004a). According to Linacre (2004b), there are a number of guidelines used in order to evaluate the functioning of rating scale and used to offer evidence to support the substantive aspect of validity. The essential guidelines are used to ensure responses to the rating scale are consistent with intention of the item developers and to help ensure stable estimates of thresholds. These are:

1. All items are oriented with the construct of measurement interest
2. Each rating scale category contains at least 10 observations,
3. The rating scale categories are used consistently,
4. Average respondent measures increase monotonically with each category,

5. The unweighted mean-square values are less than 2.0.

Linacre (2004b) further states that there are other guidelines that are helpful, but not required to evaluate rating scale function. Transposing the other guidelines on this study, the step thresholds should advance with the values of the rating scale categories to reflect the appropriate anxiety scale. Also, the step thresholds should be a least 1.1 logits apart for a 4-point scale, and these thresholds should be no more than 5.0 logits apart.

If the rating scale does not function properly according to some of these criteria, an adjustment can be done to improve the scale’s functioning. Turning to Linacre (2004b), rating scale functioning can be improved by combining adjacent categories. Because rating scale functioning of the A-GUR fit the criteria described above, I did not combine rating categories to seek improvement.

2c. Does the empirical item hierarchy of the new instrument support an a priori item hierarchy?

When Eysenck et al. (2007) articulated the relationship of state anxiety and cognitive processes during stressful learning experiences, they identified two specific factors that influence cognitive processes: situation-specific anxiety and anxiety personality of the learner. Further, Eysenck and his colleagues suggested that the state of anxiety is activated when a current goal (performing the male GUR exam) is threatened. The detection of threat to the current goal increases an individual’s anxiety and causes the individual’s attention to be directed toward the source of anxiety and determine a response to that particular threat. Threat stimulus may be internal (negative or worrisome thoughts), or external (threatening tasks students may be faced with toward the male GUR exam). The relationship between external stimuli (tasks associated with the male GUR exam) and the learner’s anxiety
personality laid a theoretical foundation for this study that attempted to measure anxiety in medical students as they perform the male GUR examination.

I proposed a theoretical construct map that shows the theoretical hierarchical ordering of the items on the A-GUR according to the medical students’ state anxiety levels (Figure 6). An expert faculty reviewed this theoretical construct map, and evaluated the rationale used to classify items on the proposed theoretical construct map of the aspects.

<table>
<thead>
<tr>
<th>Medical Students</th>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Anxiety</td>
<td>(Difficult aspects)</td>
</tr>
<tr>
<td></td>
<td>Digital Rectal Examination</td>
</tr>
<tr>
<td></td>
<td>Detection of Abnormality</td>
</tr>
<tr>
<td></td>
<td>Communicating with Patients</td>
</tr>
<tr>
<td>Moderate Anxiety</td>
<td>Hernia Assessment</td>
</tr>
<tr>
<td></td>
<td>Visual Inspection of Anus</td>
</tr>
<tr>
<td></td>
<td>Visual Inspection of Penis</td>
</tr>
<tr>
<td>Low Anxiety</td>
<td>Visual Inspection of Testes</td>
</tr>
<tr>
<td></td>
<td>(Easy aspects)</td>
</tr>
</tbody>
</table>

*Figure 6. Theoretical construct map of measuring medical student anxiety toward the male GUR examination.*

I conducted the Rasch analysis of the 2010-2011 A-GUR ratings and compared the item hierarchy from an empirical construct map to the above theoretical item hierarchy. To do this, I identified whether the extent the ordering of items by their difficulty reflects a construct consistent with the theoretical foundation, using Spearman’s rho. I conducted the
Rasch analysis of the A-GUR ratings employing the Rasch rating scale measurement model using the WINSTEPS® computer program (Linacre, 2004a). I compared the distribution of medical students’ state anxiety measures to the distribution of thresholds to evaluate evidence relevant to the content aspect of validity. Consistent with the ratings of a number of self-reported anxiety instruments (Howley & Dickerson, 2003; Pugh & Salud, 2007; Pugh, et al., 2009), the students recorded their state anxiety using a rating scale that employs a minimum of five categories. Thus, I considered each item as targeting a minimum of four regions of the continuum (i.e., the transitions from the scale points of 1 to 2, 2 to 3, 3 to 4, and 4 to 5). I then compared the distributions of the threshold measures for all items on the instrument to determine the extent to which the A-GUR covers the full range of medical student state anxiety.

2. **Research Question 3.** Does validity evidence based on the structural aspect of validity support the use of student anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

3a. Do items on the new instrument fit together to provide evidence that only the intended construct is being measured?

Like the statistics used to evaluate evidence based on the content aspect of validity, Rasch mean square outfit and positive point-measure correlation statistics from Rasch analysis may be used to offer evidence that the items of the A-GUR instrument combine to form the intended single cohesive construct. As described earlier, Rasch mean square outfit values that fall within .6 and 1.5 will support items measuring a common underlying construct. Positive point-measure correlation will also support that the items contribute information to the construct measured by the test as a whole. An item with a negative point-measure correlation coefficient indicates that the item may have been miscoded or, more
importantly, may tap into a different construct (Linacre, 2005). If any items were found to have a negative point-measure correlation coefficient, I examined the items to identify the basis of the misalignment with the construct.

There are a number of other statistics that can be evaluated in order to offer evidence based on the structural aspect of validity. I will review these now. Principal component analysis (PCA) of the residuals of the data was also used to offer potential evidence of the scale’s unidimensional construct. Under the Rasch models’ assumptions, if the PCA identifies substantive common factors across the residuals, this may indicate more than one latent trait (Linacre, 1998a). If the principal component analysis of the standardized item residuals identifies common variance among items, the analysis would suggest the presence of a secondary component in the data. In this study, the PCA indicated a secondary component, and I examined the loading patterns of the first and second components to determine any meaningful pattern in items’ content. These results will be described in more detail later.

I also evaluated if the data supported the assumption of local independence using Fisher’s Z index of item dependency. There are three primary steps associated with calculating the Fisher’s Z statistic. First, I determined the standardized residuals for each rating of student $n$ on item $i$, using the formula:

$$d_{ni} = \frac{(\text{observed rating} - \text{expected rating})}{SE_n}. \quad (20)$$

Second, I correlated the standardized residuals, $d_{ni}$, for all pairs of items $i, j$ across all students, thusly, producing a residual correlation coefficient ($r_{ij}$) for each item pair. Finally, using the Microsoft (2003) Excel computer program, I computed Fisher’s $Z$ statistic, using the formula:
\[ z_{ij} = \frac{1}{2} \ln \left( \frac{1 + r_{ij}}{1 - r_{ij}} \right) \]  

(21)

which has an expected value (mean) of

\[ E(z) = \frac{1}{2} \ln \left( \frac{1 + E(Q_3)}{1 - E(Q_3)} \right) = 1/2[\ln(1 + E(Q_3)) - \ln(1 - E(Q_3))], \]  

(22)

with \( E(Q3) \) equal to \(-1 / (I-1)\), and variance equal to \( 1 / (N-3) \), where \( Q_3 \) is Yen’s \( Q_3 \) statistic (Yen, 1984, 1993), a Pearson correlation coefficient between the residuals of a pair of items after partialling out the measured construct, and \( N \) is the sample size.

Standardized Fisher’s Z is then shown as:

\[ STDZ_{ij} = z_{ij} - E(z)/\sqrt{1/(N-3)}, \]  

(23)

which is has an approximate normal distribution. Typically, a standardized Z beyond ± 2 (alpha .05) can be used for identifying pairs of items for possible violations of local independence (Kim, De Ayala, & Hsieh, 2009). Because standardized Z statistics using a conventional alpha of .05 will yield an unacceptable overall probability of a Type I error, and a Bonferroni adjustment would be too conservative, I used a standardized Z value beyond ± 3 to identify possible item pairs that may violate local independence, as a balance between multiple type I errors and using a very conservative criteria. Significant dependency among item pairs from the same student group suggests a violation of the local independence assumption. In my case, I evaluated the extent of the violation of the
assumption of local independence by checking the percentage of item pairs that have significant Fisher’s $Z$ statistics. As described later, there were no significant violations, and therefore, no need to remove any of the problematic items.

**Research Question 4.** Does validity evidence based on the generalizability aspect of validity support the use of student anxiety scores of the new scale to make inferences about medical students’ level of anxiety toward the male GUR exam?

4a. Do gender and ethnicity facets contribute to explain variations in the data collected? Although research question 4a was addressed earlier in phase 1, I evaluated evidence based on the generalizability aspect of validity using the following three research questions.

4b. Are the empirical ratings of students reproducible over different items?

4c. Are the empirical ratings of items reproducible over different students?

To address the research questions 4b and 4c, I evaluated the person- and item-internal consistency reliability indices obtained from the Rasch Rating Scale analysis from the 2010-2011 A-GUR. A number of statistics, such as separation reliabilities and separation indices were used to evaluate internal consistency reliability. These statistics indicated the extent to which the student measures and item-difficulty measures could be reproduced (Smith, 2004). The reliability statistics for students reflected the reproducibility of the measures of student anxiety across different items, while the reliability statistics for the items reflected the reproducibility of the measures of item difficulty over different students.

Fisher (1992) explained separation reliability as a correlation coefficient similar to Cronbach alpha in the classical test theory that describes a ratio of true measure variance to observed measure variance. The formula below describes separation reliability ($R$):
In this equation, $SD^2$ represents observed variance of measures, and $RMSE$ is the root mean-square of the standard error of measures, or the “averaged” measurement error of reported measures. For person reliability, the value of person reliability can range from 0 to 1, with higher values indicating increased ability of the items to differentiate persons (students) reliably. For person reliability, higher values indicate that any differences between measures are less likely due to measurement error, and offers validity evidence based on the generalizability aspect of validity to support the use of student anxiety scores of the scale to make inferences from person measures. Similar to person reliability, for item reliability, the value of item reliability can range from 0 to 1, with higher values indicating that the empirical ratings of the items are reliably reproducible over different persons (students). Higher values indicate that any differences between measures are less likely due to measurement error, and offers validity evidence based on the generalizability aspect of validity to support the use of the scale’s items’ ratings to make inferences from person measures.

3. **Research Question 5.** Does the validity evidence based on the external aspect of validity support the use of student anxiety scores of the new scale to make inferences about medical students’ level of anxiety toward the male GUR exam?

5a. Do the empirical ratings of students correlate with empirical ratings of the another instrument used to measure anxiety?

In order to evaluate evidence relevant to the external aspect of validity, I
compared person measures using Pearson Product Moment Correlation across the A-GUR and the short-form, 6-item STAI Inventory (STAI; Spielberger et al., 1970). If the student ratings and person-item maps and strata are consistent across the A-GUR and the 6-item STAI instruments, this offers evidence of external aspect of validity.

Ethical Considerations

The study population for this research are medical students; there was no involvement of special populations (e.g., children, prisoners). All data that I used in this study have been collected using noninvasive means. I conducted this research study in an established educational setting. The 2005 pilot study used the Male Urogenital Student Evaluation Form, which is a preexisting instrument used to collect medical student anxiety ratings during normal educational practice. The study of the 2010-2011 A-GUR used a modified version of the Male Urogenital Student Evaluation Form to collect data. I modified the instrument to hopefully improve the quality of the assessment. The University B CEC administers the anxiety scales for educational purposes, and will continue to collect student anxiety measures, even if I were not conducting this research study. Data was collected using nonidentifiable personal information. The risk of harming or having negative impact to any of the subjects by this study is minimal. This study is eligible for an expedited review by the Institutional Review Board (IRB), according to the U.S. Department of Health and Human Services Regulations for the Protection of Human Subjects for two reasons:

1. The research involves data that have been collected solely for educational purposes. In this case, the data was collected during normal, scheduled educational training sessions and during clinical rotations.
2. The research includes the collection of individual or group characteristics or behavior, employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies. In this case, the data will be collected via an anonymous, self-reported survey.

The Offices for the Protection of Research Subjects (OPRS) of University A and University B will review the research protocol under expedited review procedure.
IV. RESULTS

In this chapter, I present the findings from the analyses I carried out as described in the previous chapter. My presentation of results reflects the order of the research questions, posed in the first chapter.

Research Questions

The research question that I address in this study is: Does evidence based on the content, substantive, structural, generalizability, and external aspects of validity support the use of student-anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam? Broken down into specific areas, the research question becomes five areas of inquiry.

Research Question 1

Does evidence based on the content aspect of validity support the use of student-anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

1a. Do the items on the new instrument adequately reflect the tasks performed during the male GUR exam?

Two independent focus groups consisting of experts and medical students reviewed the items associated with the male GUR exam. During the focus group sessions, all participants completed a three-part survey, called the 2010 Content Validity Rating Form, A-GUR Instrument (Appendix I). I administered the survey to the focus groups to evaluate appropriateness of item content and response options.
The first focus group was comprised of nine experts from University A and University B. The group consisted of two program coordinators, four faculty, one standardized patient coordinator, and two standardized patients. The group collectively reviewed 42 original tasks associated with the male GUR exam that aligned with the seven aspects identified by the theoretical construct map in Chapter 2 (Figure 6). Experts then rated the relevance of each task to performing the male GUR exam using the 2010 Content Validity Rating Form, A-GUR Instrument. Following the completion of the survey, the responses were reviewed, and a discussion resulted in the penultimate A-GUR rating scale. In summary, all experts agreed that all items of the instrument adequately reflected the tasks associated with the GUR exam. One expert suggested adding an additional item, “Addressing patient who has sensation of urination during prostate exam.” The other experts agreed to this addition. When reviewing items for appropriateness and clarity, all experts agreed that each item was relevant to the male GUR exam, but the consensus agreed that Items 8 (Palpation of epididymis) and 9 (Palpation of spermatic cord) should be combined as one item, resulting in Item 8 (Palpation of epididymis and spermatic cord). Following the item review, the experts agreed that the items be organized differently to improve students’ response compliance. The resulting instrument contained 43 items, broken into six sections. The first section contained 17 items associated with the steps of the male GUR exam. For clarity, the experts agreed on the header-prompt, “Please Rate Your Level of Anxiety Associated With Each Examination Task.” The second section contained five items associated with communication during the male GUR exam. For clarity, the experts agreed on the header-prompt, “Please Rate Your Anxiety Toward Communicating Findings With Your Patient.” The third, fourth, and fifth sections each contained four items associated with the managing of patient embarrassment, causing pain or harm to the patient, and
missing a lesion during the male GUR exam, respectively. The sixth section contained nine items associated with managing potentially awkward moments. For clarity, the experts agreed on the header-prompt, “Please Rate Your Anxiety Toward Managing Patient Embarrassment,” “Please Rate Your Anxiety Toward the Possibility of Causing Pain or Harm to Patient,” and “Please Rate Your Anxiety Toward the Possibility of Missing a Lesion,” for the third, fourth, and fifth sections, respectively.

The second focus group, consisting of nine medical students, reviewed the revised version of the scale and, like the expert focus group, completed the 2010 Content Validity Rating Form, A-GUR Instrument. The group consisted of two second-year medical students, one male and one female, from University A; three female first-year medical students, one female second-year, one male second-year, and two male third-year medical students from University B. The focus group consisted of one African American, one Mexican, two Asian, and four Caucasian students. Although students agreed that each item was relevant to the male GUR exam, five new items were added with consensus agreement. Three new items were placed under the header-prompt “Please Rate Your Anxiety Toward Communicating Findings With Your Patient.” These were, Using incorrect naming of medical terminology, Using inappropriate use of verbs, and Using inappropriate use of humor. Two additional items, placed under the header-prompt “Please Rate Your Anxiety Toward Managing Potentially Awkward Moments” included, Omitting task(s) from exam, and Performing task(s) in improper sequence. The resulting instrument contains 47 items, broken into the same six sections determined by the expert panel. The first section remain unchanged, with 17 items associated with the steps of the male GUR exam. The second section contains seven items associated with communication during the male GUR exam. The third, fourth, and fifth sections remain unchanged, and each contains four items. The sixth section
contains twelve items associated with managing potentially awkward moments. After modification, the final rating scale was reviewed independently by the expert members of Focus group 1, who agreed that the items on the new instrument fit together to measure the intended construct.

1b (1). Does the rating scale provide response options that adequately reflect the levels of students’ anxiety they may experience toward the male GUR exam?

During the aforementioned expert focus group, the experts reviewed the existing response options of the new scale to ensure that they reflected the level of student anxiety that may influence students’ anxiety toward these tasks. These experts rated the appropriateness of the existing response options using the 2010 Content Validity Rating Form, A-GUR Instrument referenced earlier (Appendix I). Two faculty suggested expanding the existing 4-response options, 1 (Low Anxiety), 2 (Moderate Anxiety), 3 (High Anxiety), and 4 (Extreme Anxiety) in order to improve differentiation across student anxiety levels. The resulting response options were; 0 (No Anxiety), 1 (Low Anxiety), 2 (Moderate Anxiety), 3 (High Anxiety), and 4 (Extremely High Anxiety). The student focus group was in agreement with the expert focus group that these response options adequately reflected the different levels of student anxiety.

1b (2). Do the response options of the new rating scale appropriately and clearly reflect students’ anxiety they may experience toward the male GUR exam?

During the aforementioned expert focus group, the experts reviewed the existing response options of the new scale to ensure that they appropriately and clearly reflected the students’ anxiety that they may experience toward the male GUR exam. The members of the expert panel indicated the proposed response options of the new rating scale were
appropriate and clear. The student focus group was in consensus with the expert focus group that these response options were appropriate and clear.

1b (3). Are the response options of the new rating scale ordered, well-defined, mutually exclusive, and exhaustive based on subject matter experts’ opinions?

After discussion, the subject matter experts of Focus group 1 agreed that five response options reflected the different levels of student anxiety, and were ordered, well-defined, mutually exclusive, and exhaustive.

1c. Do items on the new instrument fit together to provide evidence that only the intended construct is being measured?

A number of Rasch analyses were performed to offer evidence that support the items contributing to a cohesive scale that measures a single dominant construct. To evaluate evidence of unidimensionality in this study, I reviewed two statistics: the Rasch unweighted mean-square item fit indices and point-measure correlations.

1c (1). Evidence from Rasch unweighted mean-square item fit indices within the range of .6 to 1.5 supports items are measuring a common underlying construct. The first approach I used to assess whether the data evaluate the assumption of unidimensionality was a review of the Rasch unweighted mean-square item fit indices. According to Smith (2004a) and Smith (1996), in a data set where the majority of items contribute to one dimension, the unweighted mean-square item fit indices can be good indicators of items that do not conform to the construct. When including all 47 items in the analysis, outfit mean-square values indicated that all but one of the items’ values ranged between 0.6 and 1.5, which is considered productive
for measurement by Linacre (2010b). Item 36 (*Possibility of missing a lesion during prostate exam*) had an outfit mean-square statistics of 1.54, indicating that this item may be unproductive for construction of measurement, but not degrading.

1c (2). A positive point-measure correlation indicates that items contribute information to construct measured by the test as a whole.

The second approach I employed to evaluate the unidimensionality assumption was the examination of the item point-measure correlation coefficients. All 47 items on the A-GUR showed positive point–measure correlation coefficients. This finding suggests that all the items of the A-GUR contributed information to the construct measured by the test as a whole.

**Research Question 2**

Does the evidence based on the substantive aspect of validity support the use of student anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

2a. Do students exhibit aberrant response patterns indicating problematic response processes?

In order to evaluate validity evidence that supports the use of student anxiety scores to make inferences about their level of anxiety toward the male GUR exam, I reviewed the person-fit statistics obtained from Rasch analysis. Beginning with the underfitting responses, 66 (14.1%) of the 468 students had outfit mean-square statistics higher than 1.5, ranging from 1.52 to 5.20, indicating that these students’ responses may have been idiosyncratic and students may have been responding carelessly or using patterned responses so that students’ true anxiety levels toward the male GUR exam may not be adequately represented. Review
of the response patterns of the underfitting students revealed that these response patterns seem to be consistent with extreme category (No anxiety or Extremely high anxiety) overuse (Linacre, 2010b). The review of the category use frequencies showed that the 66 (14.1%) underfitting students selected categories 1 (No anxiety) and 5 (Extremely high anxiety) 32% of the time, compared to the students with acceptable fit statistics, who selected these combined categories 5% of the time. Reviewing overfitting responses, 68 (14.5%) students had unweighted mean-square statistics below the acceptable boundary of 0.6, ranging from .49 to .12. A review of the category use frequencies indicated middle category overuse, as the overfitting students selected categories 2 (Low anxiety) and 3 (Moderate anxiety) 83% of the time, compared to the students with acceptable fit statistics, who selected these categories only 38% of the time. Analysis indicated 332 (71.0%) students who had unweighted mean-square values within acceptable limits.

2b. Do students use the rating scale appropriately to self-report their anxiety toward the male GUR exam, producing ordered, well-defined, mutually exclusive, and exhaustive categories on the rating scale?

In order to evaluate if the students used the rating scale appropriately to self-report their anxiety toward the male GUR exam, and producing ordered, well-defined, mutually exclusive, and exhaustive categories on the rating scale, I reviewed evidence from evaluation of the number of observations in each category, the order of the average measures, category fit, and the ordering of the thresholds to investigate if the categories are ordered and reflect intended use.

Beginning with category counts, each of the five categories of the A-GUR included well over the minimum of 10 total observations; Category 0 (No anxiety) had 2,064 (9%) counts used, while categories 1 (Low anxiety), 2 (Moderate anxiety), 3 (High anxiety), and
4 (Extremely high anxiety) had 7,741 (35%), 8,132 (37%), 3,056 (14%), and 902 (4%) counts respectively. According to Wolfe and Smith, a minimum of 10 category counts ensures precision of thresholds (Wolfe & Smith, 2007b). In this study, average measures of the five categories were increasing and ordered, ranging from -2.72 for Category 0 (No anxiety) to 1.98 for category 4 (Extremely high anxiety), offering evidence that each category on the printed rating scale is associated with corresponding higher levels of anxiety. The outfit means-square fit statistics for the categories were 1.1 for Category 0, 1.0 for Category 1, 1.0 for Category 2, 0.9 for Category 3, and 1.0 for Category 4. Because all values fell within the range of 0.6 and 1.5 for all five categories, this support responses of the rating categories are used in accordance to Rasch model expectations (Wolfe & Smith, 2007b). Finally, an evaluation of the threshold ordering reveals the threshold values categories were -3.30 for the threshold between Category 1 and 2, -0.75 for the threshold between Category 2 and 3, 1.24 for the Category 3 to 4 threshold, and 2.80 for Category 4 to 5 threshold. The differences across thresholds are within the range of .81 to 5 logits as suggested by Wolfe and Smith (2007b), offering further evidence that the five categories are ordered and each reflects a distinct aspect of the underlying anxiety continuum.

2c. Does the empirical item hierarchy of the new instrument support an a priori item hierarchy?

I conducted a Facets analysis of the scores of the A-GUR implemented during the 2010-2011 academic year. This analysis revealed the construct map shown as Figure 7. Because the ratings of students were hypothesized to depend on the interaction between students’ anxiety, student gender, student ethnicity, and item difficulty, the empirical construct map obtained from the analysis has four panels. This construct map displays
students’ anxiety level (the second column), student gender (the third column), student ethnicity (the fourth column), and item difficulty (the last column) on to the same logit interval scale (the first column). The map revealed the students’ anxiety toward the male GUR exam measures ranged from -5.78 to 4.56 logits, with a mean of 1.00 logits and a standard deviation of 1.39 logits. The student gender measures were -.67 and -.08 for the males and female students, respectively. These findings are consistent with the aforementioned literature that suggests female medical students self-report higher anxiety than male students (Hayes, et al., 2004; Moffat et al., 2004; Robins, et al., 1997; Toews, et al., 1997). Although these studies suggest female medical students self-report their anxiety higher than males, they offer few explanations for the phenomenon. One common explanation suggests female medical students’ anxiety may be linked to less self-perceived competence, and because females tend to self-rate their competence much lower than male medical students, their self-reported anxiety will also be lower (Blanch, Hall, Roter, & Frankel, 2008; Sheets, Gorenflo, & Gorney 1993). This could be the basis of decreased self-reported anxiety for female medical students toward the male GUR exam, but further research is required before making inferences in this setting. The student ethnicity measures ranged from –0.85 to 0.61 logits, with a mean of 0.0 logits and a standard deviation of 0.42 logits. Looking more closely at the measures for each ethnicity group, measures were found to be 0.61, 0.39, 0.37, 0.31, 0.29, 0.16, -0.21, -0.29, -0.36, -0.41, and -0.85 for the Undisclosed, Japanese, Korean, Asian, Pacific Islander, Hispanic, Chinese, Mexican, African American, Caucasian, and American Indian/Native American groups, respectively. This seems consistent with the work by Robins et al. (1997) that suggested men from Asia, East Asia, Indian, or Pakistan self-reported their anxiety highest prior to performing the male GUR exam of all student groups, although there is some ambiguity in implied findings,
given the differences in categorization of ethnic groups. Although not related to the male GUR exam, another study, Smith and colleagues posited that students from Indian/Pakistani, Asian, Pacific Islander, African-American and Hispanic ethnic groups had increased anxiety founded on academic and time pressures, day-to-day functions, social issues, and health concerns when compared to Caucasian medical students (Smith, Peterson, Degenhardt, & Johnson, 2007). Although the breadth of Smith and colleagues’ work was intended to cover a broad scope, it seems to support the view that certain minority groups may have increased anxiety toward the academic load of associated with medical education, among other aspects of medical school.

Finally, the item difficulty measures ranged from –1.68 to 1.67 logits, with a mean of 0.00 logits and a standard deviation of 0.59 logits. Based on the item difficulty measures, the ordering of items from easiest to most difficult was 1, 24, 10, 20, 13, 29, 19, 40, 12, 18, 22, 23, 27, 37, 3, 21, 31, 42, 26, 16, 33, 41, 25, 38, 30, 46, 39, 2, 11, 7, 17, 9, 28, 35, 5, 45, 6, 8, 32, 43, 4, 36, 47, 44, 15, and 14, respectively (Figure 7). Figure 8 shows the item difficulty measures for each item and their placement within each of the aspects identified within the theoretical framework.

I compared this item order with the a prior theoretical item hierarchy (Figure 9). The theoretical and empirical item orderings showed some similarities, but also some differences. Consistent with the theoretical framework posited earlier, the two most anxiety-evoking aspects associated with the male GUR exam were found to be Digital Rectal Examination and Detection of Abnormality. The mean of the item difficulties of the items
Figure 7. An empirical construct map showing the relationships among medical student anxiety, student gender, student ethnicity, and items’ difficulty to endorse.
<table>
<thead>
<tr>
<th>Measure</th>
<th>DRE $(M = .91)$</th>
<th>Abnormality $(M = .27)$</th>
<th>Penis $(M = .24)$</th>
<th>Testes $(M = .10)$</th>
<th>Performance $(M = -.19)$</th>
<th>Anus $(M = -.28)$</th>
<th>Communication $(M = -.37)$</th>
<th>Hernia $(M = -.41)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.65</td>
<td>14</td>
<td>44</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>47</td>
<td>36</td>
<td>4</td>
<td>36</td>
<td>4</td>
<td>43</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>.77</td>
<td>28</td>
<td></td>
<td>32</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>.01</td>
<td></td>
<td>7</td>
<td></td>
<td>30</td>
<td>7</td>
<td>38</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>-.12</td>
<td>33</td>
<td>25</td>
<td>33</td>
<td>26</td>
<td>26</td>
<td>37</td>
<td>41</td>
<td>16</td>
</tr>
<tr>
<td>-.43</td>
<td>42</td>
<td></td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>-.71</td>
<td>29</td>
<td>3</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>-1.02</td>
<td>24</td>
<td>16</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>-1.66</td>
<td></td>
<td>26</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

**Figure 8.** An empirical construct map showing the relationships among items, mean items difficulty to endorse, and the aspects in which these items fall.
<table>
<thead>
<tr>
<th>Medical Students</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Anxiety</td>
<td>(Difficult aspects)</td>
</tr>
<tr>
<td></td>
<td>Digital Rectal Examination</td>
</tr>
<tr>
<td></td>
<td>Detection of Abnormality</td>
</tr>
<tr>
<td></td>
<td>Communicating with Patients</td>
</tr>
<tr>
<td>Moderate Anxiety</td>
<td>Hernia Assessment</td>
</tr>
<tr>
<td></td>
<td>Visual Inspection of Anus</td>
</tr>
<tr>
<td></td>
<td>Visual Inspection of Penis</td>
</tr>
<tr>
<td></td>
<td>Visual Inspection of Testes</td>
</tr>
<tr>
<td>Low Anxiety</td>
<td>(Easy aspects)</td>
</tr>
</tbody>
</table>

*Figure 9.* Theoretical Construct map of measuring medical student anxiety toward the male GUR examination.

associated with the DRE aspect (Items 14-*Inserting finger into rectum*, 15-*Palpation of prostate*, 28-*Managing patient embarrassment during prostate exam*, 32-*Possibility of causing pain or harm to patient during prostate exam*, and 47-*Management of stool spills*) was .91, indicating that for this group of medical students, the DRE aspect was most anxiety-evoking.

The second-most anxiety-evoking aspect, Detection of Abnormality, consisted of 4 items (Items 35-*Possibility of missing a lesion during hernia exam*, 34-*Possibility of missing a lesion during testes exam*, and 36-*Possibility of missing a lesion during prostate exam*) and had a mean item difficulty of .27.
The third highest anxiety-evoking aspect is associated with the penis. There were a number of additional items the focus groups felt should be included in the new scale that were associated with examination of the penis. Because many of the new items extended beyond the range of the originally posited aspect of the a priori construct, Visual Inspection of Penis, I revised the name of the new aspect to Examination of Penis to reflect the range of all new items associate with the aspect. Examination of the Penis, consisted of 9 items (44-Addressing patient who has erection, 4-Retraction of penis foreskin, 43-Potentially causing patient erection, 6-Palpation of the penis, 5-Inspection of urethral meatus, 25-Managing patient embarrassment during penis exam, 42- Being in close proximity to the patient’s genitalia, 3-Examination of pubic hair, and 29-Possibility of causing pain or harm to patient during penis exam) and had a mean item difficulty of .24.

The fourth highest anxiety-evoking aspect of the male GUR exam is Examination of the Testes. Similar to the previously discussed aspect, there were a number of additional items the focus groups felt should be included in the new scale that were associated with examination of the testes. Because many of the new items extended beyond the range of the originally posited aspect of the a priori construct, Visual Inspection of Testes, I revised the name of the new aspect to Examination of Testes to reflect the range of all new items associate with the aspect. Examination of Testes consisted of 4 items (8-Palpation of epididymis and spermatic cord, 7-Inspection of testes/scrotum, 30-Possibility of causing pain or harm to patient during testes exam, and 26-Managing patient embarrassment during testes exam), and had a mean item difficulty of .10.

The fifth highest anxiety-evoking aspect was Performance, a new aspect added during the A-GUR item development phase. The Performance aspect contained 4 items (39-Omitting task(s) from exam, 38-Performing exam with faculty supervision, respectively, 37-
Performing exam with peer supervision, and 40-Performing task(s) in improper sequence) and had a mean item difficulty of -.19.

The sixth anxiety-evoking aspect, Visual Inspection of Anus, consisted of two items (41-Being in close proximity to the patient’s anus, and 12-Visual inspection of sacrocoxygeal and perianal area) and had a mean item difficulty of -.28.

The seventh highest anxiety-evoking aspect, Communication, consisted of 11 items (11-Positioning patient for rectal/prostate exam, 2-Requesting patient to expose genitals, 46-Inspection of urethral meatus, 16-Communication regarding patient “clean up” following prostate exam, 21-Communicating findings with patient during prostate exam, 23-Using inappropriate use of verbs, 22-Using incorrect naming of medical terminology, 18-Communicating findings with patient during penis exam, 19-Communicating findings with patient during testes exam, 24-Using inappropriate use of humor, and 1-Giving the patient an overview of the examination plan) and had a mean item difficulty of -.37.

The eighth highest anxiety-evoking aspect, Examination of the Hernia, consisted of 6 items (31-Possibility of Causing pain or harm to patient during hernia exam, 27-Managing patient embarrassment during hernia exam, 13-Requesting patient to bear down/cough, 20-Communicating findings with patient during hernia exam, and 10-Visual inspection of inguinal ring area for bulges) and had a mean item difficulty of -.41.

These findings suggest the originally posited theoretical construct map used or my application (Figure 9), requires modification. Figure 10 illustrates the updated construct. Although the top two anxiety-evoking aspects, Digital Rectal Exam and Detection of Abnormalities, are placed in alignment with the original construct map, the other aspects have shifted. For example, Examinations of Penis and Testes are ordered similarly as the Visual Inspection of Penis and Testes of original construct, but they have both moved much
higher than the original construct. Hernia Assessment has shifted lower than the original construct, and now follows Examinations of Penis and Testes. The Communication aspect has shifted lower, and now follows Examinations of Penis, Testes and Hernia, a measurable difference compared to the original construct. The new construct also includes a new aspect, which was added during the A-GUR item development. The new aspect, called Performance, sits below all the aspects described above. Finally, Visual Inspection of Anus is the least anxiety-evoking of all aspects of the male GUR exam. Its placement contradicts the original construct, and sits below Examinations of Penis and Testes.

<table>
<thead>
<tr>
<th>Medical Students</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Anxiety</td>
<td>(Difficult aspects)</td>
</tr>
<tr>
<td></td>
<td>Digital Rectal Examination</td>
</tr>
<tr>
<td></td>
<td>Detection of Abnormality</td>
</tr>
<tr>
<td>Moderate Anxiety</td>
<td>Examination of Penis</td>
</tr>
<tr>
<td></td>
<td>Examination of Testes</td>
</tr>
<tr>
<td></td>
<td>Hernia Assessment</td>
</tr>
<tr>
<td></td>
<td>Performance</td>
</tr>
<tr>
<td>Low Anxiety</td>
<td>Communicating with Patients</td>
</tr>
<tr>
<td></td>
<td>Visual Inspection of Anus</td>
</tr>
<tr>
<td></td>
<td>(Easy aspects)</td>
</tr>
</tbody>
</table>

*Figure 10. Modified construct map of measuring medical student anxiety toward the male GUR examination.*
Research Question 3

Does the validity evidence based on the structural aspect of validity support the use of student anxiety scores to make inferences about medical students’ level of anxiety toward the male GUR exam?

3. Do items on the new instrument fit together to provide evidence that only the intended construct is being measured?

3 (1). Evidence from Rasch unweighted mean-square item fit indices between 0.6 and 1.5 supports items are measuring a common underlying construct.

In order to assess if items are measuring a common underlying construct, I reviewed the Rasch unweighted mean-square item fit indices. According to Smith (2004a) and Smith (1996), in a data set where the majority of items contribute to one dimension, outlying fit statistics can be good indicators of items that do not measure a common underlying construct. As described earlier in question 1c, outfit mean-square values of all but one of the 47 items ranged between 0.60 and 1.54, within the acceptable range between 0.6 and 1.5. One item (36- Possibility of missing a lesion during prostate exam), had an outfit mean-square of 1.55, considered unproductive for construction of measurement, but not degrading by Linacre (2010b).

3 (2). A positive point-measure correlation indicates that items contribute information to construct measured by the test as a whole.

As described earlier in question 1c, all 47 items on the A-GUR showed positive point-measure correlation coefficients. Because there were no items that were found to have a negative point-measure correlation coefficient, there is no need to examine the items further to identify the basis of the misalignment with the construct. This finding suggests that all the items of the A-GUR contributed useful information to the construct measured by
the test as a whole, and offers evidence based on the structural aspect of validity to support
the use of student anxiety scores to make inferences about medical students’ level of anxiety
toward the male GUR exam.

3 (3). Principal component analysis (PCA) of the residuals of real and simulated data
support unidimensionality.

As the item fit statistics described earlier are not useful in all situations for
detecting model violations, a principal component analysis (PCA) of the standardized
residuals was performed. In addition to performing PCA of the standardized residuals, a
comparison of real and simulated data was performed. Eigenvalues of the real data set are
compared to those of a software-generated, simulated, unidimensional, data set using the
parameters from the real data set as the generating values for the simulation. As described
by Smith (2002), when the components have approximately equal number of items
contributing to each component and the components are not highly correlated, PCA will
be better able to detect multidimensionality than the Rasch fit-statistics. Further, if the
PCA revealed the real data explained more of the residual variance than the simulated
data explained, multidimensionality may be indicated. The PCA of the standardized item
residuals revealed a first component that could explain 6.5% of the residual variance,
higher than the 1.7% of variance explained by the data simulated to fit the Rasch model.
Therefore, the amount of common variance among residuals was higher than the amount
of common variance that occurred by chance (when running the analysis of the data that
fit the model). Approximately half (23) of the items had positive loadings on the first
component. These items with the loading values are shown in Table 7.

Table 7 displays the loadings of the items on the first component (the second
dimension of the data), which ranged from .04 to .67. The remaining 24 items have little
association with this dimension. Based on the magnitude of the loadings, item 6 was the
item that was most strongly associated with this second dimension, followed by items 7, 5,
8, 4, 14, 3, 10, 15, 42, 12, 41, and 2, respectively. Considering the content of these items, the
second dimension these items might be measuring could be the “anxiety associated with
performing the tasks of the GUR exam.”

The remaining 24 items had negative loadings on the first component. Considering
the content of these items, the dimension these items might be measuring could be the
“sources of worry toward the GUR exam.” These items with the negative loading values are
shown in Table 8.

Given the complexity of the model that included multiple facets, I performed a PCA
of the standardized person residuals in relation to student year, institution, gender, and
ethnicity. In doing so, I could determine if contrasts appeared to separate across the first-
and second-year students, male and female students, University A and University B, and
finally, the 11 designated ethnicity groups, to potentially explain the violation of
unidimensionality indicated by the PCA of the standardized item residuals. The PCA of the
standardized person residuals indicated the contrast did not separate first- and second-year
students, male and female students, University A and University B, and the 11 designated
ethnicity groups.
Table 7

Loading Pattern of Positively-Loaded Items on the First Extracted Component from the Principal Component Analysis of the Standardized Item Residuals (Items Sorted by Loadings)

<table>
<thead>
<tr>
<th>Item</th>
<th>Item description</th>
<th>Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Palpation of penis</td>
<td>0.67</td>
</tr>
<tr>
<td>7</td>
<td>Inspection of testes/scrotum</td>
<td>0.65</td>
</tr>
<tr>
<td>5</td>
<td>Inspection of urethra meatus</td>
<td>0.61</td>
</tr>
<tr>
<td>8</td>
<td>Palpation of epididymis and spermatic cord</td>
<td>0.53</td>
</tr>
<tr>
<td>4</td>
<td>Retraction of penis foreskin</td>
<td>0.49</td>
</tr>
<tr>
<td>14</td>
<td>Inserting finger into rectum</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>Examination of pubic hair</td>
<td>0.48</td>
</tr>
<tr>
<td>10</td>
<td>Visual inspection of inguinal ring area for bulges</td>
<td>0.43</td>
</tr>
<tr>
<td>15</td>
<td>Palpation of prostate</td>
<td>0.40</td>
</tr>
<tr>
<td>42</td>
<td>Being in close proximity to patient’s genitalia</td>
<td>0.38</td>
</tr>
<tr>
<td>12</td>
<td>Visual inspection of sacrocoxygeal and perianal area</td>
<td>0.37</td>
</tr>
<tr>
<td>41</td>
<td>Being in close proximity to patient’s anus</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>Requesting patient to expose genitals</td>
<td>0.33</td>
</tr>
<tr>
<td>9</td>
<td>Palpation of inguinal ring</td>
<td>0.31</td>
</tr>
<tr>
<td>11</td>
<td>Positioning patient for rectal/prostate exam</td>
<td>0.27</td>
</tr>
<tr>
<td>13</td>
<td>Requesting patient to bear down/cough</td>
<td>0.23</td>
</tr>
<tr>
<td>16</td>
<td>Communication regarding “clean-up” following prostate exam</td>
<td>0.21</td>
</tr>
<tr>
<td>43</td>
<td>Potentially causing patient erection</td>
<td>0.20</td>
</tr>
<tr>
<td>17</td>
<td>Handling stool for occult blood test</td>
<td>0.17</td>
</tr>
<tr>
<td>44</td>
<td>Addressing patient who has an erection</td>
<td>0.16</td>
</tr>
<tr>
<td>45</td>
<td>Addressing patient who has passed gas</td>
<td>0.12</td>
</tr>
<tr>
<td>47</td>
<td>Management of stool spills</td>
<td>0.06</td>
</tr>
<tr>
<td>46</td>
<td>Addressing patient who has sensation of urination during prostate exam</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 8

Loading Pattern of Negatively-Loaded Items on the First Extracted Component from the Principal Component Analysis of the Standardized Item Residuals (Items Sorted by Loadings)

<table>
<thead>
<tr>
<th>Item</th>
<th>Item description</th>
<th>Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>Possibility of missing a lesion during testes exam</td>
<td>-0.64</td>
</tr>
<tr>
<td>35</td>
<td>Possibility of missing a lesion during hernia exam</td>
<td>-0.62</td>
</tr>
<tr>
<td>36</td>
<td>Possibility of missing a lesion during prostate exam</td>
<td>-0.61</td>
</tr>
<tr>
<td>33</td>
<td>Possibility of missing a lesion during penis exam</td>
<td>-0.58</td>
</tr>
<tr>
<td>39</td>
<td>Omitting task(s) from exam</td>
<td>-0.39</td>
</tr>
<tr>
<td>22</td>
<td>Using incorrect naming of medical terminology</td>
<td>-0.37</td>
</tr>
<tr>
<td>23</td>
<td>Using inappropriate use of verbs</td>
<td>-0.35</td>
</tr>
<tr>
<td>40</td>
<td>Performing task(s) in improper sequence</td>
<td>-0.35</td>
</tr>
<tr>
<td>31</td>
<td>Possibility of causing pain or harm during hernia exam</td>
<td>-0.34</td>
</tr>
<tr>
<td>27</td>
<td>Managing patient embarrassment during hernia exam</td>
<td>-0.30</td>
</tr>
<tr>
<td>30</td>
<td>Possibility of causing pain or harm during testes exam</td>
<td>-0.29</td>
</tr>
<tr>
<td>32</td>
<td>Possibility of causing pain or harm during prostate exam</td>
<td>-0.28</td>
</tr>
<tr>
<td>24</td>
<td>Using inappropriate use of humor</td>
<td>-0.26</td>
</tr>
<tr>
<td>26</td>
<td>Managing patient embarrassment during testes exam</td>
<td>-0.26</td>
</tr>
<tr>
<td>25</td>
<td>Managing patient embarrassment during penis exam</td>
<td>-0.25</td>
</tr>
<tr>
<td>29</td>
<td>Possibility of causing pain or harm during penis exam</td>
<td>-0.24</td>
</tr>
<tr>
<td>28</td>
<td>Managing patient embarrassment during prostate exam</td>
<td>-0.24</td>
</tr>
<tr>
<td>38</td>
<td>Performing exam with faculty supervision</td>
<td>-0.21</td>
</tr>
<tr>
<td>20</td>
<td>Communicating findings with patient hernia prostate exam</td>
<td>-0.14</td>
</tr>
<tr>
<td>18</td>
<td>Communicating findings with patient during penis exam</td>
<td>-0.13</td>
</tr>
<tr>
<td>19</td>
<td>Communicating findings with patient during testes exam</td>
<td>-0.10</td>
</tr>
<tr>
<td>21</td>
<td>Communicating findings with patient during prostate exam</td>
<td>-0.09</td>
</tr>
<tr>
<td>37</td>
<td>Performing exam with peer supervision</td>
<td>-0.09</td>
</tr>
<tr>
<td>1</td>
<td>Giving the patient an overview of the examination plan</td>
<td>-0.04</td>
</tr>
</tbody>
</table>
In addition to the PCA of the standardized person residuals, I also examined the relationship of the many facets to the positively- and negatively-loaded items to evaluate their impact item loading by evaluating bias estimates across item designation (positive or negative) and each of the facets in question. The findings indicated that student year, institution, gender had little impact on item loading designation, while ethnicity may have. Beginning with positively-loaded items, there were no ethnic groups that had bias sizes larger than .5 logits, with the exception of one ethnic group. The Undisclosed ethnic group had a bias size of -.97, $t(140) = -7.87, p < .0001$, indicating that for this group of students, the positively-loaded items were less anxiety-evoking than other student groups. For the negatively-loaded items, there were significant biases indicated for three ethnic groups, Pacific Islander, Hispanic, and American Indian/Native American. The Pacific Islander group had a bias size of .73, $t(46) = 3.63, p < .001$, indicating that for this group of students, the negatively-loaded items were more anxiety-evoking than for other student groups. For the Hispanic group (bias = -.59, $t(375) = -7.75, p < .0001$), and American Indian/Native American group (bias = -.87 $t(46) = -3.64, p < .001$), the negatively-loaded items were less anxiety-evoking than for the other ethnic groups.

Given these findings of the PCA, it might be more appropriate to define the model in a multidimensional context. Using the software ConQuest v.2 (Wu, Adams, Wilson, & Haldane; 2007), I evaluated the best model fit to the data from two model options. Model 1 is a composite unidimensional model that contains all items in Tables 8 and 9 treated as a single dimension, while Model 2 is a two-dimensional model that contains the items of Tables 8 and 9 as two distinct dimensions). I compared the data-to-model-fit using the difference in deviance statistics, where the difference in deviance statistics between the two
models is distributed approximately as Chi-square. A summary of these statistics is shown in Table 10.

When comparing the data-to-model fit of the two models, the difference in deviance was 1,184.86, while the difference in parameters was 2. Treating this similar to the chi-square statistic, the results of the comparison of deviances suggest there is a statistical difference of the two models. Given these finding, I calculated the AIC and BIC to further evaluate which of the two models was more appropriate. As described in Chapter 4, the AIC and the BIC are model selection approaches that combine the evaluation of data fitting with a penalization for model complexity. The AIC is a measure of goodness of fit of an estimated statistical model that combines the evaluation of data fitting with a penalization for model complexity, and is written as:

\[ AIC = -2(\text{loglikelihood}) + 2P_d, \]

where \( P_d \) is the number of parameters in the model. The first term is the log likelihood of the best fitting curve of the considered model, which can be viewed as a measure of the fit between the model and evidence. Researchers call this term deviance. The latter term is a measure of model complexity. This term may be viewed as a penalty for overparameterization. In the event two models fit the evidence approximately equally well, researchers are advised to select the simpler model, indicated by a lower value (Kieseppa, 2002; Lin & Dayton, 1997).
Table 10

_A Summary of Deviances for the Two Model Options_

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidimensional</td>
<td>46735.64</td>
<td>51</td>
</tr>
<tr>
<td>Two-dimensional</td>
<td>45550.78</td>
<td>53</td>
</tr>
</tbody>
</table>

The BIC is a criterion for model selection among a class of parametric models with different numbers of parameters. It takes the form:

\[
BIC = -2(\log\text{likelihood}) + P_d \ln N, \tag{26}
\]

where \( P_d \) is the number of parameters in the model, and \( N \) is the number of observations.

Similar to AIC, BIC evaluates the fit between the model and evidence with a deviance term. However, it uses a different penalty term for overparameterization. Because the BIC introduces a penalty term for the number of parameters in the model, and takes the number of observations into account, its penalty term increases with sample size. Thusly, BIC accommodates “overfitting” data while avoiding the selection of more complex models. The BIC approach favors simpler models over those that the AIC approach (Kass & Raftery, 1995; Kuha, 2004). A summary of these findings are shown in Table 11.
Table 11

*The AIC and BIC of the Two Models Under Consideration for the Analysis of the A-GUR*

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45272.3</td>
<td>49113.3</td>
</tr>
<tr>
<td>2</td>
<td>44770.2</td>
<td>48027.9</td>
</tr>
</tbody>
</table>

According to the AIC, the model with the lowest value AIC and was most appropriate for the analysis of the A-GUR data was Model 2, the two-dimensional model. Similarly, the BIC supported these findings, and suggested the multidimensional model better represents the underlying data.

3 (4). Evaluation of local independence using the Fisher’s Z statistic of item dependency.

In order to evaluate the local independence assumption, I used Fisher’s Z modification of Yen’s $Q_3$ statistic. I correlated the standardized residuals obtained from the Facets analysis using the PASW Statistics 19, Release Version 19.0.0 (SPSS, Inc., 2011). Given the difficulty of computing correlations from a very large number of the items from the 47 items of the A-GUR, I randomly selected 50 item pairs—which was considered an adequate number—during consultation with G. Karabatsos (personal communication, May 24, 2011). I then computed Fisher’s Z statistics (see Equation 23), using the Microsoft (2010) Excel computer program. I calculated the mean and standard deviation of Fisher’s Z for the sample of items to be .15 and .46, respectively. As suggested by Kim et al. (2009), I considered a pair of items to be significantly dependent if their Fisher’s Z statistic was more
than two standard deviations above, or less than two standard deviations below the mean of Fisher’s $Z$. Given this standard, I considered item pairs with Fisher’s $Z$ statistic more than 1.08 or less than -0.78 significantly dependent. From the evaluation of the Fisher’s $Z$ statistics of the sampled 50 item pairs, I found three item pairs with significant Fisher’s $Z$ statistics. Looking closer at the items’ relationships, two item pairs [items 9 (Palpation of inguinal ring) and 26 (Managing patient embarrassment during testes exam), and items 6 (Palpation of penis) and 23 (Using inappropriate verbs)] seem to have little in common, suggesting the associated significant Fisher’s $Z$ statistic may be a function of type I error, a relatively small threat to the accuracy of estimates of item difficulties and student anxiety measures. The third item pair consisting of items 31 (Possibility of causing pain or harm during hernia exam) and 35 (Possibility of missing a lesion during hernia exam), do, however, seem to have possible redundancy, as both are associated with the hernia exam. This may suggest these items may pose a possible threat to the accuracy of estimates of item difficulties and student anxiety levels that, ultimately, could result in overestimation of reliability and information functions as suggested by Sireci, Thissen, and Wainer (1991). I will consider either removing of one item, or combining the items into one item to minimize artificial inflation of reliability estimates.

**Research Question 4**

Does the validity evidence based on the generalizability aspect of validity support the use of student anxiety scores of the new scale to make inferences about medical students’ level of anxiety toward the male GUR exam?

4a. Do gender and ethnicity facets contribute to explain variations in the data collected?
In order to evaluate if gender and ethnicity facets contribute to explaining variations in the data, I evaluated the AIC, BIC, and chi-square statistics of the three possible models that reflected the association of gender and ethnicity on medical student anxiety toward the male GUR exam.

Model 1 is a 3-faceted model where \( \theta_n \) (theta) is the anxiety level of student \( n \), \( \delta_i \) (sigma) is the difficulty of item \( i \), \( \gamma_j \) (gamma) is the effect of ethnicity \( j \), and \( \lambda_l \) (lamda) is the difficulty of the \( l^{th} \) threshold. In this model, the effect of gender is not considered, and the rating scale for all the items shared a common set of threshold difficulty parameters in that all students used the rating scale in a similar fashion.

\[
\ln\left[ \frac{p_{nijl}}{p_{nijl(k-1)}} \right] = \theta_n - \delta_i - \gamma_j - \lambda_l \tag{27}
\]

Model 2 is a 3-faceted model where \( \theta_n \) (theta) is the anxiety level of student \( n \), \( \delta_i \) (sigma) is the difficulty of item \( i \), \( \gamma_j \) (gamma) is the effect of gender \( j \), and \( \lambda_l \) (lamda) is the difficulty of the \( l^{th} \) threshold. In this model, the effect of ethnicity is not considered, and the rating scale for all the items shared a common set of threshold difficulty parameters.

\[
\ln\left[ \frac{p_{nijl}}{p_{nijl(k-1)}} \right] = \theta_n - \delta_i - \gamma_j - \lambda_l \tag{28}
\]

Model 3 is a 4-faceted model where \( \theta_n \) (theta) is the anxiety level of student \( n \), \( \delta_i \) (sigma) is the difficulty of item \( i \), \( \gamma_j \) (gamma) is the effect of ethnicity \( j \), \( \lambda_l \) (lamda) is the
effect of gender $l$, and $\tau_k$ (tau) is the difficulty of the $k^{th}$ threshold. In this model the rating scale for all the items shared a common set of threshold difficulty parameters.

$$\ln \left[ \frac{p_{nijkl}}{p_{nij(k-1)}} \right] = \theta_n - \delta_i - \gamma_j - \lambda_l - \tau_k \quad (29)$$

Model 3: $$\ln \left[ \frac{p_{nijkl}}{p_{nij(k-1)}} \right] = \theta_n - \delta_i - \gamma_j - \lambda_l - \tau_k \quad (29)$$

I evaluated the AIC and BIC of each model to determine the most appropriate model for the analysis. Table 11 provides all the AIC and BIC values for the three forms of the Facets models under consideration.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46095.8</td>
<td>49112.5</td>
</tr>
<tr>
<td>2</td>
<td>46497.8</td>
<td>49024.9</td>
</tr>
<tr>
<td>3</td>
<td>45272.3</td>
<td>48706.7</td>
</tr>
</tbody>
</table>

According to the AIC, the model with the lowest AIC value would provide the best balance between fit and parsimony. In this case, Model 3, the 4-faceted model where ethnicity and gender effects were taken into consideration, was the most appropriate model according to AIC. Similarly, the BIC supported these findings, and further suggested that Model 3 was the most appropriate model for the analysis of the A-GUR data.

4b. Are the empirical ratings of students reproducible over different items?
In order to evaluate if the empirical ratings of students are reproducible over different items, I evaluated the internal consistency reliability of students using separation reliabilities and separation indices (or strata). In this study, the reliability of student separation was estimated to be .97. The high degree of student separation reliability suggested that the students participating in this assessment are well-differentiated in terms of their level of anxiety. The student separation ratio was estimated to be 5.95, which indicated that the variability of student anxiety measures is about 6 times larger than the precision of those measures. Finally, the student separation index (strata) was estimated to be 8.27, indicating that this instrument was able to statistically distinguish between approximately eight levels of student anxiety.

4c. Are the empirical ratings of items reproducible over different students?

Continuing my evaluation of validity evidence based on the generalizability aspect of validity, I also evaluated the internal consistency reliability of items using separation reliabilities and separation indices (or strata). The reliability of item separation was estimated to be .99. The high degree of item separation reliability suggested that the items of the A-GUR are well-differentiated in terms of their level of difficulty. The item separation ratio was estimated to be 8.49, which indicated that the variability of item difficulty measures is about 8 times larger than the precision of those measures. Finally, the item separation index (strata) was estimated to be 11.65, suggesting that among the 47 items included in this analysis, there are about 12 statistically distinct strata of item difficulty.
Research Question 5

Does the validity evidence based on the external aspect of validity support the use of student anxiety scores of the new scale to make inferences about medical students’ level of anxiety toward the male GUR exam?

5. Do the empirical ratings of students correlate with empirical ratings of another instrument used to measure anxiety?

In order to evaluate if the empirical student ratings correlate with empirical ratings of another instrument used to measure anxiety, I compared student ratings across the A-GUR and the short form, 6-item Spielberger State Trait Anxiety Inventory (STAI; Spielberger et al., 1970). Before running any correlational analyses, it was essential to recode the ratings of the STAI. Of the six items of the short-form STAI, three were negatively worded in terms of anxiety; Items 1 (I feel calm), 3 (I am content), and 4 (I am relaxed) originally indicated less anxiety as the ratings increased. Because of this, I reverse coded the ratings of these three items to improve alignment with the A-GUR ratings. In doing so, I modified each rating of 1 (Not at all) to reflect a rating of 4 (Very much so), of 2 (Somewhat) to reflect a rating of 3 (Moderately so), of 3 (Moderately so) to reflect a rating of 2 (Somewhat), and 4 (Very much so), to reflect a rating of 1 (Not at all).

In order to evaluate if the empirical student ratings of the A-GUR correlate with empirical ratings of the short form six-item STAI, an established rating scale used to measure anxiety, I used Pearson Product Moment Correlation (PPMC). In doing so, I compared the Rasch person measures of the A-GUR to those of the STAI from the 329 students who completed both ratings scales. Findings suggested that the ratings were moderately correlated, \( r(329) = .48, p < .01 \). Correcting for attenuation, the correlation was estimated to be .53.
V. DISCUSSION

In this chapter, I summarize the key findings obtained from my validation study and point out the practical implications of those findings. I first discuss the validity evidence related to the use of the A-GUR. I then discuss practical implications of the findings and point out the limitations of this study. I also suggest many relevant research questions that other researchers might find interesting to explore in the future. I conclude the chapter with a short summary of this study and its significance to the medical education community.

The Anxiety Toward the Male Genitourinary Rectal Exam (A-GUR) Scale

The examination of validity evidence related to the interpretation of scores obtained from the 2010-2011 administration of the A-GUR revealed a large amount of evidence that supported the use of the A-GUR to measure medical students’ anxiety toward the male GUR exam. However, this study also revealed some aspects of the instrument that might benefit from modification. In this section, I briefly review the validity evidence, and make suggested scale and administration changes that may improve the A-GUR’s measurement capabilities.

Evidence Relevant to the Content Aspect of Validity

The content analysis consisted of item review from two independent focus groups comprised of experts and medical students. The analysis revealed the items associated with the A-GUR adequately reflected the tasks performed during the male GUR exam items, and the response options adequately reflected the different levels of student anxiety. The expert focus group also agreed that the response options were ordered, well-defined, mutually
exclusive, and exhaustive. Further, evidence from Rasch analyses supported the items may be consistent with a cohesive scale that measures the theoretical construct. First, the Rasch unweighted outfit mean-square item fit statistics for all but one of the 47 items of the A-GUR fell within the range of 0.6 to 1.5, supporting the notion that the items likely measures a common underlying construct. Although Item 36 (Possibility of missing a lesion during prostate exam) had an outfit mean-square statistics of 1.55, this value was not high enough to become degrading to the measurement construct. The point–measure correlation coefficients were all positive, indicating that there were no recoding errors and all 47 items of the A-GUR were positively correlated with the total score of the A-GUR. Given this finding, Item 36 was not removed from the scale.

**Evidence Relevant to the Substantive Aspect of Validity**

In order to evaluate validity evidence that supports the use of student anxiety scores to make inferences about their level of anxiety toward the male GUR exam, I reviewed the Rasch mean-square person-fit statistics, category functioning, and compared the resulting item hierarchy with a theoretical item hierarchy predicted by experts. A review of the response patterns of the underfitting students revealed that the 66 (14.1%) underfitting students selected the extreme categories—categories 1 (No anxiety) and 5 (Extremely high anxiety)—at approximately a 6 times higher rate than the students with acceptable fit statistics. Response patterns of the overfitting students suggested that the 68 (14.5%) overfitting students selected the low-middle categories—categories 2 (Low anxiety) and 3 (Moderate anxiety)—83% of the time. This compares to the students with acceptable fit statistics, who selected these categories only 38% of the time. Only the overfitting students’
responses were centralized in the low-middle categories, indicating that the individual
responses of the students who fit the Rasch model expectations were quite varied.

A review of the category functioning indicated that (a) each category on the rating
scale was associated with corresponding higher levels of anxiety, (b) responses of the rating
categories were used in accordance to Rasch model expectations, and (c) the four thresholds
were ordered and, therefore, each category reflects a distinct aspect of the underlying
anxiety continuum. The category counts indicated Category 0 (No anxiety) had 2,064 (9%)
counts used, while categories 1 (Low anxiety), 2 (Moderate anxiety), 3 (High anxiety), and 4
(Extremely high anxiety) had 7,741 (35%), 8,132 (37%), 3,056 (14%), and 902 (4%) counts,
respectively. The diverse use of the categories, in addition to the response patterns
described above, supports previous researchers’ findings that first and second-year medical
students, as a whole, have a broad range of anxiety levels toward the male GUR exam
(Robins et al., 1997; Howley & Dickerson (2003).

Using the mean item difficulty measures for each aspect, I compared the empirical
construct map defined by the mean item difficulty for each aspect with that originally
posited hierarchy. This comparison may indicate a modified construct, with the eight aspects
ranging from most difficult (most anxiety-evoking) to least difficult (least anxiety-evoking)
as Digital Rectal Examination, Detection of Abnormality, Examination of Penis,
Examination of Testes, Hernia Assessment, Performance, Communicating with Patients, and
Visual Inspection of Anus. The two most anxiety-evoking aspects, Digital Rectal Exam and
Detection of Abnormalities, remain the same, while the other aspects have shifted slightly
when compared to the originally posited aspect hierarchy. I may make some inferences from
these findings, such as (a) students’ anxiety toward specific aspects of the male GUR exam
may be influenced by other variables I did not account for (i.e., not measured by the
A-GUR) when forming my original hierarchy, or simply, (b) the original construct was a
good estimate, but the new construct map reflects a more precise estimate of the hierarchy of
students’ anxiety toward the aspects of the male GUR exam.

Evidence Relevant to the Structural Aspect of Validity

In order to evaluate if the items on the new instrument fit together to provide evidence that
only the intended construct is being measured, I evaluated a range of evidence, including the
unweighted mean-square item-fit indices, point-measure correlations, principal component
analysis (PCA) of standardized residuals, and the Fisher’s Z statistic of item dependency.

Beginning with the latter evaluation described in more detail in Chapter 4, from the
evaluation of the Fisher’s Z statistics of the sampled 50 item pairs, I found 3 item pairs with
significant Fisher’s Z statistics. Given the small percentage of items exhibiting dependency
problems (6%; 5% would be expected by chance alone), I determined that there was a trivial
threat of violating the local independence assumption of the measurement model, and I will
consider removal of one item of the violating pair to see if this might improve outcomes.

Continuing on, I reviewed the three other types of evidence relevant to the structural aspect
of validity. As described earlier, outfit mean-square values of all but one of the 47 items
ranged between 0.60 and 1.5, within the acceptable range between 0.6 and 1.5. One item
(36-Possibility of missing a lesion during prostate exam), had an outfit mean-square of 1.55.
Although considered unproductive for construction of measurement, an outfit mean-square
of 1.55 was not degrading (Linacre, 2010b), so I did not remove the item. A review of the
point-measure correlations revealed that all point-measure correlations were positive, and
suggested that all the items of the A-GUR were positively correlated with the total score of
the A-GUR. Interestingly, the PCA of the standardized residuals of real data indicated that
there may be two dimensions measured by the A-GUR. The first component consisted of 23
items with positive loadings, while the second component consisted of 24 items with
negative loadings. The items with positive loadings ranged from 0.04 to 0.67. In Chapter 4, I
described these items as “anxiety associated with performing the tasks of the GUR exam.”
The 24 items with negative loadings ranged from -0.04 to -0.61. I described these items as
“sources of worry toward the GUR exam.” Although the PCA of the standardized person
residuals in relation to student year, institution, gender, and ethnicity indicated no impact of
these facets on the model, a deeper examination of the interactions between all facets
(gender, institution, student-year, and ethnicity) and positively- and negatively-loaded items
suggested ethnic groups may be associated with the items’ loadings.

Further, I calculated the AIC and BIC of the two model options—a composite
unidimensional model that contains all 47 items treated as a single dimension, and a
2-dimensional model that contains the positively and negatively-loaded items as two distinct
dimensions. Both the AIC and BIC suggested that the 2-dimensional model better
represented the underlying data. Although the Rasch outfit mean-square indices suggested
the items were unidimensional, further dimensionality evaluation via PCA of the
standardized residuals suggested the instrument consisted to two dimensions with equal
number of items, giving less meaning to the Rasch outfit mean-square indices (Smith, 2001).
AIC and BIC calculations further supported the A-GUR may be 2-dimensional, having the
“anxiety toward tasks,” and “worry toward tasks” dimensions.

Returning to the attentional control theory presented by Eysenck et al. (2007) on
which this study was based, it becomes clear why worry may be an influential variable and a
potential dimension that ought to be considered and measured in this setting. Although this
theory is strongly founded on anxiety and its negative effect on cognitive processes, the
authors’ second assumption suggested that *worrisome thoughts* may interfere with cognitive function and tax the self-regulating mechanism and auxiliary processing used to inhibit these negative thoughts. To apply this to the setting of this study in a worse-case scenario, a student’s anxiety may be compounded by their worrisome thoughts such that if the student cannot inhibit negative thoughts (worry) toward the male GUR exam, these worrisome thoughts inhibit cognitive function even further. The spiral continues as the student’s anxiety increases, resulting in increased worry, until they are focused solely on the negative thoughts. All cognitive function is diverted from the learning moment, and focused on negative affect.

In summary, evaluation of evidence relevant to the structural aspect of validity seems to suggest two critical points, (a) worry may be as equally important as anxiety on medical students’ negative affect toward the male GUR exam, and (b) researchers may be able to utilize the A-GUR scale to measure the sources of medical students’ worry toward the male GUR exam.

**Evidence Relevant to the Generalizability Aspect of Validity**

I evaluated a variety of evidence to address a number of research questions related to the evidence based on the generalizability aspect of validity. First, I evaluated whether gender and ethnicity facets contributed to explain the variations in the data collected. I evaluated AIC, BIC, and chi-square statistics that reflected the three possible models: Model 1—the 3-faceted model that considered the effects of students, items, and ethnicity; Model 2—the 3-faceted model that considered the effects of students, items, and gender; and Model 3—the 4-faceted model that considered the effects of students, items, ethnicity, and gender. Calculated values of the AIC and BIC suggested that the 4-faceted model, where ethnicity
and gender effects were taken into consideration, was the most appropriate model for the analysis of the A-GUR data. Following this, I evaluated internal consistency reliability using separation reliabilities and strata to evaluate if the ratings of the students were reproducible over different items. The reliability of student separation was estimated to be .97, while the student separation ratio was estimated to be 5.95, and the student separation index (strata) was estimated to be 8.27. These findings suggested the A-GUR was able to statistically distinguish between approximately eight levels of student anxiety. In a similar fashion, I examined parallel indices to evaluate if the ratings of the items were reproducible over different students. The reliability of item separation was estimated to be .99, while the item separation ratio was estimated to be 8.49, and the item separation index (strata) was estimated to be 11.65. These findings suggested the 47 items of the A-GUR were able to distinguish between more than 11 levels of item difficulty.

**Evidence Relevant to the External Aspect of Validity**

In order to evaluate if the empirical student ratings correlate with empirical ratings of another instrument used to measure anxiety and offer evidence relevant to the external aspect of validity, I compared student ratings across the A-GUR and the short form, 6-item Spielberger State Anxiety Inventory (STAI; Spielberger, et al., 1970). Findings suggested that the ratings were moderately correlated, $r (329) = .48$, $p < .01$. Correcting for attenuation, the correlation was estimated to be .53. The modest correlation may indicate a number of things. It could indicate that there are other variables that influence the students’ anxiety toward the male GUR exam that are not measured by the STAI state form. A closer inspection of the two item sets from the anxiety and worry domains offered other evidence. For the comparison of the A-GUR anxiety domain items and the short form, 6-item STAI,
the correlation was estimated to be \( r(320) = .39, p < .01 \). Correcting for attenuation, the correlation was estimated to be .43. Similarly, for the comparison of the A-GUR worry domain and the short form, 6-item STAI, the correlation was estimated to be \( r(320) = .36, p < .01 \). Correcting for attenuation, the correlation was estimated to be .40. It seems to make sense that the isolated worry domain of the A-GUR had a decreased correlation with the STAI, which was intended to only measure anxiety and not worry. A comparison of the A-GUR anxiety and the A-GUR worry domains indicated the estimated correlation was \( r(320) = .74, p < .01 \). Correcting for attenuation, the estimated correlation was .76. This finding may suggest that although there may be two domains of the A-GUR indicated, these domains may continue to have such an influential relationship on each other that separating the associated items may degrade the psychometric properties of the A-GUR. Separate treatment of the items of these domains should be considered only after evaluating the validity evidence.

**Implications**

There are a number of implications of the findings of this study. The primary implication is that the evidence from the content, substantive, generalizability, and external aspects of validity suggested the A-GUR can be used to measure medical students’ anxiety toward the male GUR exam. Evaluation of the evidence of the structural aspect of validity was less decisive. The point-measure correlations were all positive, indicating all the items of the A-GUR were positively correlated with the total score of the A-GUR, and evaluation of Fisher’s Z statistic upheld the assumption of local independence. Although Rasch unweighted mean-square item-fit indices suggested the A-GUR was unidimensional, we should remember that if the A-GUR is truly composed of two dimensions (worry and
anxiety), and that each dimension is represented by approximately equal number of items, then the mean-square item-fit indices are not effective at detecting departures from unidimensionality (Smith, 2001). Under these conditions, the PCA would be better able to detect multidimensionality. The PCA of the standardized residuals suggested the A-GUR consisted of two domains, “anxiety” and “worry”. As discussed earlier, evidence from this study suggests a high correlation between the anxiety and worry domains, with estimated $r_{(320)} = .74, p < .01$. For the purpose of this work, the intimate relationship of anxiety and worry is not troublesome, as Spielberger described worry as a component of anxiety in Chapter 2. Further, Eysenck and his cohorts whose work provided the theoretical foundation of this study, considered worry equally powerful as anxiety at negatively impacting cognitive function (pp. 21-26). Although the multidimensionality does not contradict the theoretical foundation of the study, it may pose a problem from an assessment perspective. First and foremost, these findings suggest that the assumption of unidimensionality, a requirement of the many-facet Rasch model may be violated. This creates a problem when scores are interpreted from a unidimensional perspective when the scale is truly multidimensional. Because it is fairly common for the requirement of unidimensionality to be unmet in practical settings, there are acceptable methods in which to treat multidimensional instruments so that inferences about examinees’ scores can be made with confidence. For example, future research that utilizes the multidimensional Random Coefficient Multinomial Logit (MRCML) model could be used for fitting data to multidimensional extensions of the many-facet Rasch model used in this study. Given these findings of this study, I would recommend that future analyses be performed and data interpreted in a 2-dimensional context. This may allow more accurate estimation of person
measures in regard to the two specific sources of negative affect (anxiety and worry) toward the male GUR exam.

Because the goal of the medical educators is to minimize and/or alleviate any negative affect toward the male GUR exam, I would not suggest, at this point, treating this instrument as two distinct subscales— the “A-GUR” that measures anxiety toward the male GUR exam, and the “W-GUR,” that measures worry toward the male GUR exam until further research can support or refute the current findings regarding the dimensionality of the data. Until further validity evidence is evaluated, I suggest that the best practice seems to be to administer the entire A-GUR to capture both domains.

In addition to the validity evidence offered by this study, findings from this work may also lead us to improved teaching modalities to minimize or alleviate medical students’ anxiety toward the male GUR exam. The modified construct map based on this study (Figure 10), suggests a new hierarchy of medical students’ anxiety toward the aspects of the male GUR examination. Consistent with the work of Lawrentschuk and Bolton (2004), the findings of my study suggest that the Digital Rectal Exam continues to be the most anxiety-evoking aspects of the male GUR exam for medical students learning the exam. This work also suggests Detection of Abnormality continues to be more anxiety-evoking than the other aspects of the GUR exam. Understanding the hierarchy of the aspects of the GUR exam and the medical students’ anxiety toward these aspects can translate directly into teaching methodologies that may improve medical students’ affect toward these particularly problematic aspects of the male GUR exam. For example, simply allowing students more opportunities to practice the DRE on a manikin in a simulated clinical environment prior to their experience with the patient instructor may be enough to minimize their anxiety toward the DRE aspect. Offering increased resources on pathologies associated with the DRE,
penis/testes, and hernia exams, along with realistic examples on the manikins, may help alleviate medical students’ anxiety toward detection of abnormalities during the GUR exam. Further, novel supplemental teaching programs that utilize some of the commonly utilized behavior intervention or cognitive therapy approaches discussed in Chapter 2 could be implemented for these aspects associated with the highest levels of student anxiety. With the understanding that specific aspects may require more attention or simply more opportunities to practice the steps of the GUR exam, more resources can be allocated to help improve student affect toward and increase learning during the male GUR exam. Furthermore, the study suggests that the less anxiety-evoking aspects, Communicating with Patients and Visual Inspection of Anus, may require less teaching time and sensitivity toward the students’ affect, so that more effort can be directed to the more anxiety-evoking aspects of the male GUR exam.

**Limitations**

There are a number of limitations related to the interpretation and applications of the findings from this study. The first limitation is associated with the context in which the A-GUR scale was administered at both institutions. At University A there was an administrative limitation to consider. The male GUR exam skills were taught in a combined session with the breast exam skills at this institution. Because the students completed the A-GUR scale prior to the hands-on experience with breast and male GUR exams, it may have been difficult for students to differentiate anxiety toward the two exams, and self-report their anxiety toward the male GUR exam with a high degree of accuracy. There was a similar trend for the administration of the A-GUR at University B for the second-year (M2) students. During this administration, the “mock oral exam” was scheduled to immediately
follow the A-GUR session. For some students, the mock oral exam was highly anxiety-evoking. This may have diverted students’ attention and erroneously decreased self-reported anxiety toward the male GUR exam.

Another limitation of this study is the homogeneity of the medical student sample with which I worked. Although I selected the two institutions to increase ethnic diversity of the sample, both institutions were set in Midwestern urban environments. Further, given the focus of the study, the sample of students was limited to first and second-year medical students. This sample of medical students had a relatively narrow range of clinical experience, which may have decreased the variability of the students’ anxiety measures. Expansion of the administration of the A-GUR to include medical students with more clinical experience may result in greater variability in the anxiety levels. A study of the A-GUR function in a more heterogeneous group of medical students would help address this limitation.

Finally, this study is limited in terms of the generalizability of the findings. First, there may have been some sampling bias introduced, particularly with the sample from University B. Although there was a very high participation rate (91.8%) from University A and the first-year students of University B (97.6%), there was a smaller sample (79.9%) from the second-year students of University B. The circumstances of the combined “mock-oral” exams and the male GUR training for the second-year students of University B may have influenced the focus of the students. Given the relative high-stakes of the “mock-oral” exam, fewer students may have selected to participate in the study, possibly introducing some sampling bias, a systematic error due to a non-random sample of a population. This may have ultimately compromised the external aspect of validity. Also, generalizability may
have been compromised due to the decreased sample of students that completed the six items relevant to the STAI. Reviewing the student group that completed the STAI-related items, there were 332 out of the 468 students that completed the STAI-relevant items. Of these students, 33 were from University A, while the remaining 299 were from University B. Comparing gender percentages across the students who completed the STAI items and those who did not, of the sample that completed the items 52.1% were male and 47.9% were female, while for the student group that did not complete the items 48.5% were male and 51.5% female. The ethnic composition of the two groups varied. Although the top three highest percentage ethnic groups for both groups were Caucasian, Asian, and African American, the respective percentage of these three ethnicities combined within both groups was 80.1% and 63.2% for the students that completed the STAI items and those who had not. Regarding the ethnicities with smaller percentages, there were even greater variations across the student group who had completed the STAI items and those who had not. Given the differences in ethnic compositions of the groups may limit the generalizability of the correlational findings of the A-GUR and STAI. Finally, I carried out this study across two undergraduate medical programs that have different ethnic populations, but at the same time teach the male genitourinary rectal exam via a similar method. The various psychometric properties of the scale I reported in this study might not be replicable for other programs that use different teaching methods for the male GUR exam. Finally, the findings may also not generalize to examinees having different characteristics than the students in this study.
**Future Directions**

This study not only provided answers to several questions related to the validity of the interpretations of the scores obtained from the A-GUR, but also generated a number of interesting research questions for me and other researchers to address in future studies. First, expansion of this study could be performed in order to bolster the validity evidence evaluated in this study. For example, this study could be expanded across institutions outside of the Midwest. This would offer a larger sample of medical students and increase geographic variability, and ultimately, may increase cultural diversity. Although my study had an adequate sample size for the analyses conducted, it lacked the power to examine differential item functioning across the 11 designated ethnic groups. Findings from the PCA of standardized person residuals suggested ethnicity may impact dimensionality of the A-GUR. Given this, expanding the current research to a multi-institutional study would increase the overall sample size and the sample sizes for the 11 ethnic groups under consideration. This larger sample size may permit differential item functioning (DIF) evaluation across ethnic groups and help evaluate the generalizability of the current findings.

If DIF is found, findings from this work might provide a new perspective on how to modify the instrument to better serve medical students and medical educators. For example, if there are trends indicated, new subscales could be developed to further isolate the source of negative affect toward a specific aspect of the male GUR exam. A specific hypothetical example might be an administration of the A-GUR indicates that a group of medical students rates their anxiety higher toward all of the items associated with the penis exam. Potentially, after consulting with focus groups consisting of subject matter experts, specific items could be developed to further isolate the root cause of the negative affect toward the penis exam for that group of medical students. Once identified, best practices used to
alleviate and/or minimize students’ anxiety may be implemented during the training sessions and throughout their clinical experiences.

Other changes may benefit the administration of the A-GUR rating scale. First, the development of cut-scores implementing one of a number of methods described by Downing and Yudkowsky (Downing & Yudkowsky, 2009), may help educators differentiate students whose anxiety toward the male GUR exam is high enough to affect their learning of the exam. Educators could then implement additional or specialized intervention for these students. Additionally, medical educators should also standardize future administrations of the A-GUR to a single time-point in the medical students’ learning continuum. Doing so will allow educators to compare groups while reducing the potential impact from maturation and/or practice effect differences. Another area of research stemming from my research revolves around the introduction of the negative affect, worry, and how worry should be treated in the context of this work. The first steps stemming from this area includes further evaluation of the validity evidence for the A-GUR to be split into two subscales that are intended to measure anxiety and worry. The evaluation of further validity evidence might bolster evidence relevant to structural aspect of validity. Further, this research might particularly benefit from additional evidence relevant for the external aspect of validity, where I would compare the empirical student ratings of the worry subscale with empirical ratings of an established instrument used to measure worry, such as the Penn State Worry Questionnaire (PSWQ; Meyer et al., 1990), introduced in Chapter 2. Findings from this proposed study might suggest the anxiety and worry subscales could be administered and analyzed independently so that medical educators may improve estimations of the various contributing factors to overall anxiety toward the male GUR exam and address them more effectively in the medical curriculum.
Finally, an area of interest that seems to be relevant to this work includes the addition of measuring trait-anxiety of medical students while also measuring their specific sources of anxiety toward the male GUR exam. Although this study focused on measuring state-anxiety via the short-form STAI state scale, it might be informative if trait anxiety, the disposition to respond to stressful situations, is measured in future studies using the short-form STAI trait scale. First, the modest correlation of A-GUR scores with the STAI state scale seems to indicate the presence of another, unmeasured, variable that may influence medical students’ anxiety toward the male GUR exam. Inclusion of the STAI trait scale will allow researchers to evaluate correlations of students’ scores of the anxiety and worry subscales and those of the STAI trait form, and potentially offer additional evidence relevant to the external aspect of validity. More importantly, it may help medical educators more accurately pinpoint best practices to address students’ anxiety toward the male GUR exam. Given that there are differing techniques described in Chapter 2 used to alleviate and/or minimize state and trait anxieties, it seems important to identify the typical anxiety personality of medical students so that the most effective practices are utilized to address students’ anxiety toward the male GUR exam.

Conclusion

I examined validity evidence related to the use of scores obtained from the 2010-2011 administration of the A-GUR at University A and University B. I focused my analyses on the evidence relevant to the content, substantive, structural, generalizability, and external aspects of validity. I stated several validity propositions based on a construct modeling approach and evaluated each proposition using a many-facet Rasch model. In phase one I developed and refined the items of a new instrument, called the A-GUR, using input from
two focus groups of subject matter experts to support evidence of content validity. In phase two I used the A-GUR to measure medical student anxiety toward the male GUR exam and gather validity evidence for score interpretation. Although my findings offered evidence to support the content, substantive, generalizability, and external aspects of validity, the evidence relevant to the structural aspect of validity requires further examination. As suggested earlier, the potential for multidimensionality introduced by the worry dimension should be evaluated in future studies. Additional studies may help to identify best analyses practices that might increase accuracy of student measures and ultimately, better identify the sources of medical students’ negative affect toward the male GUR exam.

The findings from this study also make a scientific contribution to the medical education literature. The various types of analyses I carried out in this study might provide researchers with practical examples of how to use the recommendations of the 1999 Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999) as a framework for gathering validity evidence related to content, substantive, structural, generalizability, and external aspects of validity, the evidence relevant to measuring anxiety of learners in medicine. The results of this study have also led to a modified construct map that suggests a new hierarchy of medical students’ anxiety toward the aspects of the male GUR examination. Understanding the hierarchy of the aspects of the GUR exam, and the medical students’ anxiety toward these aspects, may translate directly into novel teaching methodologies that may improve medical students’ affect toward the particularly problematic aspects of the male GUR exam, the digital rectal exam and detection of abnormality. Furthermore, the study suggests that the less anxiety-evoking aspects—communicating with patients and visual inspection of anus—may require less teaching time and sensitivity toward the students’ affect, so that more effort can be
directed at the more anxiety-evoking aspects of the male GUR exam. Additionally, this study offers some new insights regarding the medical students’ negative affect toward the specific aspects of the male GUR exam such that my research may lead to a refining of existing theories or the development of new theories associated with negative affect toward intimate exams. Finally, this study also generated many interesting research questions for researchers to explore in the future.
REFERENCES


Predictive Analytics SoftWare (2011). SPSS (Version 19) [Computer program]. Chicago, IL: SPSS.


APPENDIX A: MALE GENITOURINARY RECTAL (GUR) EXAMINATION

The techniques are demonstrated and the students are asked to perform the examination as follows: The patient is requested to either stand or lay supine, and expose the genitals. Beginning with the *genital exam*, the physician sits and with gloved hands examines the distribution of pubic hair, and the size and shape of the penis and testes, as well as the color of the skin. The physician then examines the skin at the base of the penis for inflammation, excoriations, nits or lice. The physician retracts the foreskin of the penis, if necessary, and inspects the glans of the penis, looking for lesions, or scars. The urethral meatus is then inspected for discharge by pressing it between thumb and forefinger. If discharge is present, the physician notes the color, and a culture is taken. The shaft of the penis is then examined for indurations and/or tenderness. Next, the scrotum is examined. The contours of the scrotum are examined for swelling, and the skin of the scrotum is examined for rashes or nodules. The testes are palpated gently and size and consistency are noted. Existing nodules are also noted. The physician then palpates the right epididymis, noting any nodules and tenderness. The right spermatic cord, including the vas deferens, is then examined for nodules or swelling. This examination is repeated on the left epididymis and spermatic cord. With the patient still standing, the *hernia assessment* begins with the physician visually inspecting the right inguinal ring and femoral canal for bulges. The physician requests the patient to cough or bear down, as in a bowel movement, and palpates the inguinal ring to assess any potential bulging mass. The physician also palpates for a femoral hernia, again asking the patient to cough or bear down, noting any swelling or tenderness. This examination is repeated on the left inguinal ring and femoral canal.

Examination of the anus, rectum and prostate begins with the patient lying on his left side,
APPENDIX A (continued)

close to the edge of the table, and draped. The patient is asked to lift his right leg, so to flex it at the hip and knee. The light is adjusted to maximize exposure, and the physician begins the examination by inspecting the sacrococcygeal and perianal areas, noting any lesions, redness, rashes, and excoriations. Abnormal areas are palpated to note tenderness. Asking the patient to bear down again, the physician inspects the anus for hemorrhoids, fissures or lesions. The physician advises the patient that they will be examining the patient’s rectum, and requests they bear down again. The physician inserts a lubricated forefinger, pointing toward the umbilicus, checking for sphincter tone, tenderness, irregularities, and nodules. The rectal wall is palpated by inserting the finger fully, rotating clockwise, and then counterclockwise, noting any nodules, irregularities or tenderness. Asking the patient to bear down at this point allows the physician to reach any inaccessible irregularities. With finger in the full-counterclockwise position, the posterior surface of the prostate gland can be then felt on the anterior wall of the rectum. The size, shape, and consistency are noted. The lateral lobes and median sulcus are palpated, and any hardening, nodules, and tenderness are noted. The physician returns their hand to the neutral position, and removes their finger. The physician removes their gloves, noting the presence of fecal matter. If fecal matter is present, it is sampled and tested for the occult blood. The examination concludes with the physician either wiping the lubricant from the patient’s anus or supplying the patient with tissues to do it himself. During the entire examination and process of learning, the physician and patient educator direct the students to the proper techniques of examination, as well as appropriate language to use during the examination. After viewing a demonstration of the
examination, the students perform the same sequence of examination; starting with the penis and testes, hernia assessment, anus and rectum, and finally, the prostate.
APPENDIX B: SAMPLE FROM STAI FORM Y-1

STAI-AD
Self- Evaluation Questionnaire
STAI Form Y-1

Name__________________________________ Date___         ___ S _____
Age_________   Gender (Circle)   M   F    T _____

DIRECTIONS

A number of statements which people have used to describe themselves are
given below. Read each statement and then blacken in the appropriate circle to
the right of the statement to indicate how you feel right now, that is, at this
moment. There are no right or wrong answers. Do not spend too much time on
any one statement but give the answer which seems to describe your present
feelings best.

1. I feel calm 1 2 3 4
12. I feel nervous 1 2 3 4

Reproduced from State Trait Anxiety Inventory for Adults, Copyright © 1983 Spielberger,
C. D., with the permission of publisher, Mind Garden, Inc.
APPENDIX C: SAMPLE FROM STAI FORM Y-2

STAI-AD
Self- Evaluation Questionnaire
STAI Form Y-2

Name__________________________________________ Date_____________

DIRECTIONS

A number of statements which people have used to describe themselves are given below. Read each statement and then blacken in the appropriate circle to the right of the statement to indicate how you generally feel.

25. I feel like failure
27. I am "calm, cool, and collected"
32. I lack self confidence

1  2  3  4

ALMOST ALWAYS
ALMOST NEVER
SOMETIMES
OFTEN

Reproduced from State Trait Anxiety Inventory for Adults, Copyright © 1983 Spielberger, C. D., with the permission of publisher, Mind Garden, Inc.
# APPENDIX D: HAMILTON ANXIETY RATING SCALE (HAM-A)

## Hamilton Anxiety Rating Scale (HAM-A)

### Patient Information

<table>
<thead>
<tr>
<th>Patient</th>
<th>Date</th>
<th>Day</th>
<th>Mth</th>
<th>Year</th>
<th>Time</th>
<th>Hr</th>
<th>Min</th>
</tr>
</thead>
</table>

**Personal notes**

### 1. Anxious mood

This item covers the emotional condition of uncertainty about the future, ranging from worry, insecurity, irritability and apprehension to overpowering dread.

| 0- The patient is neither more or less insecure or irritable than usual. | □ |
| 1- Doubtful whether the patient is more insecure or irritable than usual. | □ |
| 2- The patient expresses more clearly to be in a state of anxiety, apprehension or irritability which he may find difficult to control. However, the worrying still is about minor matters and thus without influence on the patient’s daily life. | □ |
| 3- At times the anxiety or insecurity is more difficult to control because the worrying is about major injuries or harms which may occur in the future. Has occasionally interfered with the patient’s daily life. | □ |
| 4- The feeling of dread is present so often that it markedly interferes with the patient’s daily life. | □ |

### 2. Tension

This item includes inability to relax, nervousness, bodily tensions, trembling and restless fatigue.

| 0- The patient is neither more or less tense than usual. | □ |
| 1- The patient seems somewhat more nervous and tense than usual. | □ |
| 2- The patient expresses clearly unable to relax and full of inner unrest, which he finds difficult to control, but is still without influence on the patient’s daily life. | □ |
| 3- The inner unrest and nervousness is so intense or frequent that it occasionally interferes with the patient’s daily work. | □ |
| 4- Tensions and unrest interfere with the patient’s life and work at all times. | □ |

### 3. Fears

This item includes fear of being in a crowd, of animals, of being in public places, of being alone, of traffic, of strangers, of dark, etc. It is important to note whether there has been more phobic anxiety during the present episode than usual.

| 0- Not present. | □ |
| 1- Doubtful whether present. | □ |
| 2- The patient experiences phobic anxiety but is able to fight it. | □ |
| 3- It is difficult to fight or overcome the phobic anxiety, which thus to some extent interferes with the patient’s daily life and work. | □ |
| 4- The phobic anxiety clearly interferes with the patient’s daily life and work. | □ |
APPENDIX D (continued)

4. Insomnia
This item covers the patient’s subjective experience of sleep duration and sleep depth during the three preceding nights. Note: Administration of hypnotics or sedatives is disregarded.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-</td>
<td>Usual sleep duration and sleep depth.</td>
</tr>
<tr>
<td>1-</td>
<td>Sleep duration is doubtfully or slightly reduced (e.g. due to difficulties falling asleep), but no change in sleep depth.</td>
</tr>
<tr>
<td>2-</td>
<td>Sleep depth is also reduced, sleep being more superficial. Sleep as a whole is somewhat disturbed.</td>
</tr>
<tr>
<td>3-</td>
<td>Sleep duration and sleep depth is markedly changed. Sleep periods total only a few hours per 24 hours.</td>
</tr>
<tr>
<td>4-</td>
<td>Sleep depth is so shallow that the patient speaks of short periods of slumber or dozing, but no real sleep.</td>
</tr>
</tbody>
</table>

5. Difficulties in concentration and memory
This item covers difficulties in concentration, making decision about everyday matters, and memory.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-</td>
<td>The patient has neither more or less difficulty in concentration and/or memory than usual.</td>
</tr>
<tr>
<td>1-</td>
<td>Doubtful whether the patient has difficulty in concentration or memory.</td>
</tr>
<tr>
<td>2-</td>
<td>Even with major effort it is difficult for the patient to concentrate on his daily routine work.</td>
</tr>
<tr>
<td>3-</td>
<td>The patient has pronounced difficulties with concentration, memory, or decision making, e.g. in reading newspaper article or watching a television programme to the end.</td>
</tr>
<tr>
<td>4-</td>
<td>During the interview the patient shows difficulty in concentration, memory, or decision making.</td>
</tr>
</tbody>
</table>

6. Depressed mood
This item covers difficulties both the verbal and non-verbal communication of sadness, depression, despondency, helplessness and hopelessness

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-</td>
<td>Not present.</td>
</tr>
<tr>
<td>1-</td>
<td>Doubtful whether the patient is more despondent, or sad than usual, or is only vaguely so.</td>
</tr>
<tr>
<td>2-</td>
<td>The patient is more clearly concerned with unpleasant experiences, although he still lacks helplessness, or hopelessness.</td>
</tr>
<tr>
<td>3-</td>
<td>The patient shows clear non-verbal signs of depression and/or hopelessness.</td>
</tr>
<tr>
<td>4-</td>
<td>The patient remarks on despondency and helplessness or the non-verbal signs dominate the interview and the patient cannot be distracted.</td>
</tr>
</tbody>
</table>

7. General somatic symptoms: Muscular
Weakness, stiffness, soreness, or real pain, more or less diffusely localized in the muscles, such as the jaw ache or neck ache.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0-</td>
<td>The patient is neither more or less sore or stiff in the muscles than usual.</td>
</tr>
<tr>
<td>1-</td>
<td>The patient seems somewhat more stiff or sore in the muscles than usual.</td>
</tr>
<tr>
<td>2-</td>
<td>The symptoms have a character of pain.</td>
</tr>
<tr>
<td>3-</td>
<td>Muscle pain interferes to some extent with the patient’s daily work and life.</td>
</tr>
<tr>
<td>4-</td>
<td>Muscle pain is present most of the time and clearly interferes with the patient’s daily work and life.</td>
</tr>
</tbody>
</table>
### APPENDIX D (continued)

#### 8. General somatic symptoms: Sensory
This item includes increased fatigability and weakness or real functional disturbances of the senses, including tinnitus, blurring of vision, hot and cold flashes and prickling sensations.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not present</td>
</tr>
<tr>
<td>1</td>
<td>Doubtful whether the patient’s indications of symptoms are more pronounced than usual.</td>
</tr>
<tr>
<td>2</td>
<td>The sensations of pressure reach the character of buzzing in the ears, visual disturbances, and prickling or itching sensations in the skin.</td>
</tr>
<tr>
<td>3</td>
<td>The generalized sensory symptoms interfere to some extent with the patient’s daily life and work.</td>
</tr>
<tr>
<td>4</td>
<td>The generalized sensory symptoms are present most of the time and clearly interfere with the patient’s daily work and life.</td>
</tr>
</tbody>
</table>

#### 9. Cardiovascular symptoms
This item includes tachycardia, palpitations, oppression, chest pain, throbbing in the blood vessels, and feelings of faintness.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not present.</td>
</tr>
<tr>
<td>1</td>
<td>Doubtful whether present.</td>
</tr>
<tr>
<td>2</td>
<td>Cardiovascular symptoms are present, but the patient can still control them.</td>
</tr>
<tr>
<td>3</td>
<td>The patient has occasional difficulty controlling the cardiovascular symptoms, which thus to some extent interferes with his daily life and work.</td>
</tr>
<tr>
<td>4</td>
<td>Cardiovascular symptoms are present most of the time and clearly interfere with the patient’s daily life and work.</td>
</tr>
</tbody>
</table>

#### 10. Respiratory symptoms
Feelings of constriction or contraction in throat or chest, dyspnea or choking sensations and sighing respiration

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not present.</td>
</tr>
<tr>
<td>1</td>
<td>Doubtful whether present.</td>
</tr>
<tr>
<td>2</td>
<td>Respiratory symptoms are present, but the patient can still control them.</td>
</tr>
<tr>
<td>3</td>
<td>The patient has occasional difficulty controlling the respiratory symptoms, which thus to some extent interferes with his daily life and work.</td>
</tr>
<tr>
<td>4</td>
<td>Respiratory symptoms are present most of the time and clearly interfere with the patient’s daily life and work.</td>
</tr>
</tbody>
</table>

#### 11. Gastro-intestinal symptoms
The items covers difficulties in swallowing, “sinking” sensation in stomach, dyspepsia (heartburn or burning sensation in the stomach, abdominal pains related to meals, fullness, nausea and vomiting), abdominal rumbling, and diarrhea.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not present.</td>
</tr>
<tr>
<td>1</td>
<td>Doubtful whether present (or doubtful whether different from usual).</td>
</tr>
<tr>
<td>2</td>
<td>One or more gastro-intestinal symptoms are present, but the patient can still control them.</td>
</tr>
<tr>
<td>3</td>
<td>The patient has occasional difficulty controlling the gastro-intestinal symptoms, which thus to some extent interferes with his daily life and work.</td>
</tr>
<tr>
<td>4</td>
<td>The gastro-intestinal symptoms are present most of the time and clearly interfere with the patient’s daily life and work.</td>
</tr>
</tbody>
</table>
**APPENDIX D (continued)**

12. **Genito-urinary symptoms**  
This item includes non-organic or psychic symptoms such as frequent or more pressing passing of urine, menstrual irregularities, anorgasmia, dyspareunia, premature ejaculation, loss of erection.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not present.</td>
</tr>
<tr>
<td>1</td>
<td>Doubtful whether present.</td>
</tr>
<tr>
<td>2</td>
<td>One or more genito-urinary symptoms are present, but the patient can still control them.</td>
</tr>
<tr>
<td>3</td>
<td>Occasional, one or more genito-urinary symptoms are present to such a degree that they interfere to some extent with the patient’s daily life and work.</td>
</tr>
<tr>
<td>4</td>
<td>The genito-urinary symptoms are present most of the time and clearly interfere with the patient’s daily life and work.</td>
</tr>
</tbody>
</table>

13. **Other autonomic symptoms**  
This item includes dryness of the mouth, blushing or pallor, sweating and dizziness.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not present.</td>
</tr>
<tr>
<td>1</td>
<td>Doubtful whether present.</td>
</tr>
<tr>
<td>2</td>
<td>One or more autonomic symptoms are present, but they do not interfere with the patient’s daily life or work.</td>
</tr>
<tr>
<td>3</td>
<td>Occasional, one or more autonomic symptoms are present to such a degree that they interfere to some extent with the patient’s daily life and work.</td>
</tr>
<tr>
<td>4</td>
<td>Autonomic symptoms are present most of the time and clearly interfere with the patient’s daily life and work.</td>
</tr>
</tbody>
</table>

14. **Behavior during interview**  
The patient may appear tense, nervous, agitated, restless, tremulous, pale, hyperventilating or sweating during the interview. Based on such observations a global estimate is made.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The patient does not appear anxious.</td>
</tr>
<tr>
<td>1</td>
<td>It is doubtful whether the patient is anxious.</td>
</tr>
<tr>
<td>2</td>
<td>The patient is moderately anxious.</td>
</tr>
<tr>
<td>3</td>
<td>The patient is markedly anxious.</td>
</tr>
<tr>
<td>4</td>
<td>The patient is overwhelmed by anxiety, for example with shaking and trembling all over.</td>
</tr>
</tbody>
</table>

Total score _______

**Ham-A Score level of anxiety**

- <17: mild
- 18 – 24: mild to moderate
- 25 -30: moderate to severe
APPENDIX E: ANXIETY IN MEDICAL STUDENTS QUESTIONNAIRE

<table>
<thead>
<tr>
<th>Course: (circle one)</th>
<th>GEP</th>
<th>Undergraduate Course</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Age........... |       |                      |
|               |       |                      |

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
</table>

1) I feel confident about starting work in the clinical years of my course  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

2) I have been really looking forward to this stage of my medical training  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

3) I do not think I will be able to cope with the new demands of the clinical years  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

4) My previous clinical skills teaching has prepared me adequately to start on the wards  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

5) I feel confident about taking medical histories from real patients  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

6) I feel confident in discussing personal and/or sensitive issues with patients  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

7) I feel confident about performing physical examinations on real patients  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

8) I feel confident in performing practical procedures on real patients  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

9) I am prepared for the uncertainty I may feel during this period  
   Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

10) I am concerned that I will not be appreciated as a member of the clinical team  
    Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

11) I feel frightened about the prospect of working in clinical areas  
    Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

12) I feel I may be intimidated by certain members of the clinical team e.g. consultants  
    Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

13) I feel other life experiences will help me cope with clinical practice  
    Strongly disagree (1) (2) (3) (4) (5) (6) (7) Strongly agree

Please enter any thoughts you may have to the following two questions:

What concerns you most about starting clinical work on the wards?

What do you feel has been the most useful/helpful thing in helping you to cope with the challenge of clinical work?
APPENDIX F: M2 STUDENT EVALUATION FORM 1

Breast Examination Training Session
M2 Student Evaluation Form 1

Sex

Number

Please rank, the top three (1st, 2nd & 3rd) causes of your anxiety, if any, in performing clinical female breast exams: 1 = Highest cause, 2 = 2nd cause, 3 = 3rd cause.

Causing harm or pain to the patient
The personal/intimate nature of the exam
Palpating the nipple and areola
I have no anxieties relating to this exam
Other:

Please rate your overall comfort level in performing a clinical female breast exam.

Extremely Anxious
1 2 3 4 5 6

Very Comfortable

Please rate your comfort level in performing visual inspection of the female breast?

Extremely Anxious
1 2 3 4 5 6

Very Comfortable

Please rate your comfort level in eliciting and interpreting nipple discharge.

Extremely Anxious
1 2 3 4 5 6

Very Comfortable

Please rate your comfort level in detecting abnormalities during the clinical female breast exam.

Extremely Anxious
1 2 3 4 5 6

Very Comfortable

Reproduced from Proceedings for Surgical Education Week. Tucson, AZ, March, 2007
Copyright © 2007 with the permission of authors, C.M. Pugh & L. Salud.
APPENDIX G: STUDENT EVALUATION FORM 1

M1 Pelvic Teaching Session
Student Evaluation Form 1

Insert 1 for your highest cause of anxiety in performing the female pelvic examination.
Insert 2 for your second highest cause of anxiety.
Insert 3 for your third highest cause of anxiety.

___ Causing harm or pain to the patient
___ The personal/intimate nature of the exam
___ General performance anxiety
___ I have no anxieties relating to this exam
___ Use of the Speculum
___ Religious beliefs
___ Touching the genitalia
___ Other: __________________

Please rate your comfort level in performing a clinical female pelvic exam.

<table>
<thead>
<tr>
<th>Extremely</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxious</td>
<td>Comfortable</td>
</tr>
<tr>
<td>1  2  3  4  5  6</td>
<td></td>
</tr>
</tbody>
</table>

Please rate your comfort level in performing visual inspection of the female perineum.

<table>
<thead>
<tr>
<th>Extremely</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxious</td>
<td>Comfortable</td>
</tr>
<tr>
<td>1  2  3  4  5  6</td>
<td></td>
</tr>
</tbody>
</table>

Please indicate your comfort level in performing speculum insertion and visualization of the cervix.

<table>
<thead>
<tr>
<th>Extremely</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxious</td>
<td>Comfortable</td>
</tr>
<tr>
<td>1  2  3  4  5  6</td>
<td></td>
</tr>
</tbody>
</table>

Please indicate your comfort level in performing a bimanual exam and palpation of the uterus.

<table>
<thead>
<tr>
<th>Extremely</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxious</td>
<td>Comfortable</td>
</tr>
<tr>
<td>1  2  3  4  5  6</td>
<td></td>
</tr>
</tbody>
</table>

Reproduced from Teaching and Learning in Medicine, Copyright © 2008 with the permission of authors, C.M. Pugh, E.T. Obadina, E. T. & K. A. Aidoo.
Please rank, in order, in the top three causes of your anxiety, if any, in performing clinical male urogenital exam: 1= primary cause of anxiety, 2 = second highest cause of anxiety, 3- third highest cause of anxiety.

___Causi3ng harm or pain to the patient   ___General performance anxiety
___The personal/intimate nature of the exam  ___Causing patient erection
___Self Embarrassment                     ___Touching the genitalia
___I have no anxieties relating to this exam ___Sanitation Issues
___Other: ____________________________

Please rate your comfort level in performing a clinical male urogenital exam.

<table>
<thead>
<tr>
<th>Extremely Anxious</th>
<th>Very Comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Please rate your comfort level in performing visual inspection/palpation of the penis/testes.

<table>
<thead>
<tr>
<th>Extremely Anxious</th>
<th>Very Comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Please indicate your comfort level in performing hernia assessment.

<table>
<thead>
<tr>
<th>Extremely Anxious</th>
<th>Very Comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Please indicate your comfort level in performing visual inspection of the anus.

<table>
<thead>
<tr>
<th>Extremely Anxious</th>
<th>Very Comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Please indicate your comfort level in performing a digital rectal/prostate examination.

<table>
<thead>
<tr>
<th>Extremely Anxious</th>
<th>Very Comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX I: 2010 CONTENT VALIDITY RATING FORM A-GUR INSTRUMENT

**Part 1 Instructions:** The following items are being considered for inclusion in the A-GUR, a survey that measures medical student anxiety toward tasks involved during the male Genitourinary Rectal (GUR) examination.

**(A)** Please assist us in reviewing the *item content* by indicating how strongly you feel that the particular task be included in the scale used to measure medical student anxiety toward the male GUR examination.

Please rate how sure you are that the task is relevant to performing the male GUR exam. Please use the following scale:

1 = Definitely do not include this task,  2 = Not sure if this task should be included   3= Pretty sure this task should be included,    and 4= Definitely include this task

<table>
<thead>
<tr>
<th>Task</th>
<th>Definitely do not include this task (1)</th>
<th>Not sure if this task should be included (2)</th>
<th>Pretty sure this task should be included (3)</th>
<th>Definitely include this task (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requesting patient to expose the penis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examination of pubic hair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retraction of penis foreskin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inspection of urethral meatus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Touching the penis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inspection of testes/scrotum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of epididymis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of spermatic cord</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicating my finding with the patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient embarrassment during penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient embarrassment during testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient erection during the penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient erection during the testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facing the patient while your perform the exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being in close proximity to the patient’s genitalia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing pain to the patient during penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing pain to the patient during testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially missing a lesion during the penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially missing a lesion during the testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of the inguinal ring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual inspection of inguinal ring area for bulges</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicating my finding with the patient during penis/testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Continue scale on next page]
### APPENDIX I (continued)

<table>
<thead>
<tr>
<th>Task</th>
<th>Definitely do not include this task (1)</th>
<th>Not sure if this task should be included (2)</th>
<th>Pretty sure this task should be included (3)</th>
<th>Definitely include this task (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potentially causing pain to the patient during hernia exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially missing a lesion during the exam for hernia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requesting patient to lie in left decubitus position</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Inspection of sacrocoxygeal and perianal area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requesting patient to bear down</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inserting finger into rectum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of prostate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicating findings with patient during prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient embarrassment during prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient erection during the prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicating during potentially awkward moment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fear of patient passing gas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being in close proximity to the patient’s anus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing pain to the patient during prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially missing a lesion during the exam of prostate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handling stool for occult blood test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Part 2 Instructions.**

Please assist us in developing *item content* by indicating if there are other relevant tasks associated with the male GUR examination that were not included in the section above. Please use the space provided below to make suggestions.
APPENDIX I (continued)

**Part 3 Instructions:** Please assist us in reviewing the appropriateness of the *proposed response options*. Below are the 5 response options I feel cover the full range of medical student anxiety toward the tasks associated with the male GUR exam.

<table>
<thead>
<tr>
<th>No Anxiety (0)</th>
<th>Little Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extreme Anxiety (4)</th>
</tr>
</thead>
</table>

Please indicate how strongly you feel that the response options above adequately reflect the possible range of medical student anxiety toward the tasks associated with the male GUR examination. Please check the option below that represents your feelings.

- □ 1 = The response options do *not* adequately reflect the possible range of anxiety experienced by medical students
- □ 2 = The response options do adequately reflect the possible range of anxiety experienced by medical students

If you feel alternative options for response options (ratings) would be more appropriate for measuring medical students' levels of anxiety while performing the male GUR examination, please provide them in the space below:

Thank you for your participation.

Any questions you may have about this study may be directed to my faculty sponsor, Everett Smith, PhD, Associate Professor, Educational Psychology, University of Illinois at Chicago at phone 312-996-5630 and email, evsmith@uic.edu.

Deb Rooney, PhDc
derooney@nmh.org
APPENDIX J: 2010 ANXIETY TOWARD MALE GUR EXAM INSTRUMENT

2010
A-GUR Instrument

Part 1 Instructions: Please select your gender and ethnic group.

Gender: I am

- Male
- Female

Ethnic Group: The ethnic group I most identify with is

- Caucasian
- African American
- Pacific Islander
- Mexican
- Hispanic
- Prefer not to disclose
- Asian
- Chinese
- Korean
- Japanese
- American Indian/Alaskan Native

Part 2 Instructions:
This survey will be used to identify specific sources of medical student anxiety toward the genitourinary rectal (GUR) examination. Before starting, please try to picture yourself in a clinical exam room, where a male patient is waiting for you to perform the examination. Now, please ask rate your anxiety level toward each aspect of the male GUR exam listed below.

<table>
<thead>
<tr>
<th>Please Rate Your Level of Anxiety Associated with Each Examination Task</th>
<th>No Anxiety (0)</th>
<th>Low Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extremely High Anxiety (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giving the patient an overview of the examination plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requesting patient to expose genitals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Examination of pubic hair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retraction of penis foreskin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inspection of urethral meatus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of the penis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inspection of testes/scrotum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of epididymis and spermatic cord</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of the inguinal ring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual inspection of inguinal ring area for bulges</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positioning patient for rectal/prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual inspection of sacroccygeal and perianal area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requesting patient to bear down/cough</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inserting finger into rectum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palpation of prostate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication regarding patient “clean up” following prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Handling stool for occult blood test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please continue on opposite side of page

A-GUR Instrument, v.2 (11/15/10)
<table>
<thead>
<tr>
<th>Please Rate your Anxiety Toward Communicating Findings with your Patient</th>
<th>No Anxiety (0)</th>
<th>Low Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extremely High Anxiety (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>During penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During hernia exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using incorrect naming of medical terminology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using inappropriate use of verbs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using inappropriate use of humor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please Rate your Anxiety Toward Managing Patient Embarrassment</th>
<th>No Anxiety (0)</th>
<th>Low Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extremely High Anxiety (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>During penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During hernia exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please Rate Your Anxiety Toward the Possibility of Causing Pain or Harm to Patient</th>
<th>No Anxiety (0)</th>
<th>Low Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extremely High Anxiety (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>During penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During hernia exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please Rate Your Anxiety Toward the Possibility of Missing a Lesion</th>
<th>No Anxiety (0)</th>
<th>Low Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extremely High Anxiety (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>During penis exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During testes exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During hernia exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please Rate Your Anxiety Toward Managing Potentially Awkward Moments</th>
<th>No Anxiety (0)</th>
<th>Low Anxiety (1)</th>
<th>Moderate Anxiety (2)</th>
<th>High Anxiety (3)</th>
<th>Extremely High Anxiety (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performing exam with peer supervision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performing exam with faculty supervision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omitting task(s) from exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performing task(s) in improper sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being in close proximity to the patient’s anus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being in close proximity to the patient’s genitalia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially causing patient erection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addressing patient who has erection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addressing patient who has passed gas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addressing patient who has sensation of urination during prostate exam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Management of sexual spills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please Describe How You Are Feeling Right Now</th>
<th>Not at all (1)</th>
<th>Somewhat (2)</th>
<th>Moderately so (3)</th>
<th>Very Much so (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel calm</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel upset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am content</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am relaxed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel tense</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am worried</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thank you for your participation. Any questions you may have about this study may be directed to my faculty advisors, Everett Smith, PhD, Associate Professor, Educational Psychology, University of Illinois at Chicago at phone 312-996-5630 and email, evsmith@uic.edu, or Carla M. Pugh, MD, PhD, Associate Professor of Surgery, Northwestern University at phone (312) 695-1419 and e-mail drpugh@northwestern.edu.

Deborah Rooney, PhDC, d-rooney@northwestern.edu

A-GUR Instrument, v.2 (11/15/10)
VITA

NAME: Deborah Moulton Rooney

EDUCATION:

University of Illinois at Chicago
August, 2011: Doctor of Philosophy, Education Psychology (PhD)
Focus: Measurement, Evaluation, Statistics, and Assessment (MESA)
Thesis Topic: Validation of Instrument that Measures Medical Students Anxiety While Learning and Performing Intimate Examinations.

University of Illinois at Chicago
December 1997: Master of Associated Medical Sciences (MAMS)
Biomedical Visualization, University of Illinois at Chicago, Chicago, IL
Medical software/design/animation. Portfolio consists of traditional/digital illustration, 2D/3D modeling and animation, web publishing, graphic design, with anatomy, physiology, histology and surgical content.

Michigan State University
December 1991: Bachelor of Art (BA), Studio Art
June 1990: Bachelor of Science (BS), Biology with emphasis on anatomy and physiology

TEACHING:

Course developer, Surgical skills curricula for General Surgery residents (PGY1-5), since 2004.

Instructor, Surgical skills curricula for General Surgery residents (PGY1-5), since 2006.


Instructor, University of Illinois at Chicago, Biomedical Visualization. Multimedia applications for the medical illustrator (1998-2000). 4 credit Master-level course. Topic included development of interactive educational tools for medical education. Softwares covered include Macromedia Director and Flash, as well as HTML programming for integration of video into web-based applications.
DEBORAH ROONEY VITA, continued.  

TEACHING, continued.  

Academic Mentor, NCASE Summer Surgical Internship for High School Students  
2011- Leigh Miron  
2011- Sean Watts  
2010- Sarah Polito  
2009- Ashley Pollack  
2009- Isabella Mizzi-Moulton

PROFESSIONAL MEMBERSHIP:  
American College of Surgeons, Committee on Administration and Management of the Consortium of ACS-accredited Education Institutes, member since 2010.  
Association of Surgical Education, Subcommittee, Coordination and Administration, member since 2009.

ACADEMIC COMMITTEES:  
Feinberg School of Medicine Educational Support Committee, member since 2010  
Feinberg School of Medicine Assessment Committee, member since 2011  
STIL Research Planning Committee, foundation member since 2011

PROFESSIONAL MEMBERSHIPS:  
Society for Simulation in Healthcare (SSH), member since 2011  
Association of Surgical Education, associate member since 2007  
American Educational Research association (AERA), student member since 2006  
National Council on Measurement in Education (NCME), student member since 2006

PROFESSIONAL AND SCIENTIFIC SERVICES:  
Reviewer, Medical Education, 2007-present  
Reviewer, Canadian Medical Education Journal, 2011-present  
Proctor, SAGES Fundamentals of Laparoscopic Surgery (FLS), Central region, 2006-present

ABSTRACTS:  
DEBORAH ROONEY VITA, continued.

PUBLICATIONS:


INVITED CHAPTERS:


PRESENTATIONS:

International Objective Measurement Workshop, Vancouver, British Columbia
April, 2012. “Using a Many-Facet Rasch Measurement Approach to Identify
Specific Sources of Medical Student Anxiety Toward the Male Genitourinary
Rectal Examination.”

American College of Surgeons, Division of Education – 4th Annual Meeting of the
Consortium of ACS-Accredited Education Institutes, Chicago, IL  April 2011.
Rooney, D., Hungness, DaRosa, D., & Pugh, C. “ACS Verifications of Proficiency:
Faculty versus skills-coach.”

American College of Surgeons, Division of Education – 4th Annual Meeting of the
Consortium of ACS-Accredited Education Institutes, Chicago, IL  April 2011.
(FLS) Manual Skills Assessment: Surgeon Versus Non-surgeon Raters.”

American College of Surgeons, Division of Education – 2nd Annual Meeting of the
Consortium of ACS-Accredited Education Institutes, Chicago, IL  March 2009.
DaRosa, D., Hungness, E., Rooney, D. & Pugh, C. “Organization of a Curriculum
Designed to Transfer Learning from Skills Lab to Bedside.”

Midwest Objective Measurement Seminar, Chicago, IL April, 2008. “Using HLM to
Analyze a Self-report Survey that Measures Medical Student Anxiety While
Learning and Performing the Male Genitourinary Rectal Examination.”

Simulation Technology in Medical Education, Northwestern University, Chicago, IL
to ProMIS Simulator.”

Midwest Objective Measurement Seminar, Chicago, IL May 20, 2005.
“Comparison of Resident and Faculty Perceived Preparedness of Residents for
Surgical Cases.”

Slice of Life: Computers in Healthcare Education Symposium, 2002. Toronto,
Quebec “The UltraSlide Studio, a virtual microscopic environment.”

Slice of Life: Computers in Healthcare Education Symposium, 2001. Munich,
Germany “Real Learning in a Virtual World: A Prototype for The UltraSlide Studio,
a virtual microscopic environment.”
http://slice.gsm.com/2001/Fri_poster/Moulton.html

PA “Motivating Faculty to Use Technology in Education: The Monkey See,
Monkey Do Approach.”

Association of American Medical Colleges Central Group on Student Affairs,
Internet in Academia.”
DEBORAH ROONEY VITA, continued.

POSTERS:

7th Annual Lewis Lansberg Research Day, Northwestern University, Chicago, IL April 7, 2011. D.M. Rooney. “Using Many-facet Rasch model to Measure Medical Student Anxiety Toward the Male Genitourinary Rectal Examination at Northwestern University.”


American College of Surgeons, Division of Education – 2nd Annual Meeting of the Consortium of ACS-Accredited Education Institutes, Chicago, IL March 2009. E. Hungness, D. Rooney, D. DaRosa & C. Pugh. “Psychometric Analysis of the Phase 1 ACS/APDS Skills Curriculum Verification of Proficiency Examination.”


SOFTWARES DEVELOPED:


GUTS (Growth and Understanding through Technology): Copyright Deborah Moulton, 1997. Prototype of self-paced teaching tool for veterinary students. CD-ROM. Module reviewing a variety of concepts as related to canine gastro-intestine, such as gross anatomy, radiology, pathology. Developed using Macromedia Director™.