



VCU

Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations


Graduate School

2021

EXAMINING THE EFFECT OF ITEM-WRITING FLAWS ON THE PSYCHOMETRIC PARAMETERS OF PHARMACY THERAPEUTICS EXAMINATIONS

Veronica P. Shuford
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), and the [Medical Education Commons](#)

© The Author

Downloaded from

<https://scholarscompass.vcu.edu/etd/6528>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

EXAMINING THE EFFECT OF ITEM-WRITING FLAWS ON THE PSYCHOMETRIC
PARAMETERS OF PHARMACY THERAPEUTICS EXAMINATIONS

A dissertation submitted in partial fulfillment of the requirements for the degree of the
Doctor of Philosophy at Virginia Commonwealth University

by

Veronica Powell Shuford
Bachelor of Science, Virginia Commonwealth University, 1990
Master of Education, Virginia Commonwealth University, 1996

Director: Lisa M. Abrams, Ph.D.
Associate Professor, Foundations of Education
School of Education

Virginia Commonwealth University
Richmond, Virginia
January 2021

ACKNOWLEDGEMENT

It has taken me nearly five years to complete this journey and it would not have been possible without my Lord and Savior, Jesus Christ. There have been many people who have been involved with this dissertation process, people who have supported me; prayed for me and with me; and provided encouraging and inspirational words to make this work possible. I would like to thank each of you for your love, patience, support, guidance, encouragement, and constructive feedback.

I want to thank my husband, Fred, for his unconditional love, attentive ear, and never-ending support throughout this process. I am forever grateful for the many hours you spent reading my work, providing constructive feedback, and helping me with creating balance in my life. To my wonderful sons, Trey and Christian, always remember this quote from American author Brian Herbert, *“the capacity to learn is a gift; the ability to learn is a skill; the willingness to learn is a choice.”* Please continue to be willing to learn, be curious, ask questions, and feed your mind and soul.

I want to express my sincere gratitude to Dr. Lisa M. Abrams, my faculty advisor and the chair of my dissertation committee. I am appreciative of your patience, your guidance, your sustained support, your tenacity in assuring the quality of the dissertation, and your confidence in my ability to complete the task. Thank you from the depths of my heart for serving as my chair. I would like to extend my sincere appreciation to my dissertation committee for taking the time to serve on my committee and providing constructive feedback to improve the quality of my dissertation. I am truly humbled and grateful that you agreed to be a part of this important journey in my career.

This project would not have been possible without the generous contributions made by Dr. Cynthia K. Kirkwood, Executive Associate Dean at the Virginia Commonwealth University School of Pharmacy. I am forever grateful for your support while completing the doctoral program and my dissertation. Mrs. Katherine Henderson and I have worked together for many years in the School of Pharmacy before she joined the Virginia Commonwealth University School of Medicine. Katherine and I have spent hours upon hours reviewing, discussing, and preparing test items for examinations in pharmacy education. I am grateful for the hours you contributed to reviewing and discussing test items for my study. I can never repay you for the time that you allowed me to borrow from Alan and Emma Rose.

I would like to thank Ms. Carol Hampton for planting the seed about earning a doctorate many, many years ago. So much of what I learned about professionalism came from the time I worked under your leadership in the VCU School of Medicine. Ms. Phylliss Moret always offers words of encouragement, wisdom, guidance, and the best advice. Thank you for always lending me your ear. I hope that both of you are aware of the footprint you have left in my life. I will always remember that I am, because you were.

I want to extend a special thank you to Dr. Kristen Tarantino for the editorial assistance and Mr. Kazi R. Moore for allowing me to talk through the statistical concepts that I learned from the “Stats Guru,” Dr. Michael Broda with you. I could not have survived the program without the “Broda Stats Bible” that I created from EDUS 608: Educational Statistics and EDUS 651: Multivariate Statistics. Kazi and Dr. Broda - I am forever grateful!

I would like to thank my colleagues at Virginia Commonwealth University for your encouraging words during my time in the doctoral program. I am blessed to work with such an

enlightened and talented team of professionals. Thank you for encouraging me to sharpen my saw.

Finally, I would like to thank my former classmates, Dr. Cassandra Boyd-Willis and Dr. Hannah Sions, for their encouragement, support, numerous study sessions, and the times we spent with stress-relieving laughs. Thank you for welcoming me into your lives. I will forever cherish the relationship we developed and the time we spent during the doctoral program. Each of you made the experience in the doctoral program more valuable and rewarding.

ABSTRACT

Colleges and schools of pharmacy (C/SOP) use direct measures of assessment to provide evidence of student learning, with multiple-choice questions (MCQs) being one the most common formats used in health sciences education to assess students' knowledge, skills, and abilities (Pate & Caldwell, 2014). This study examined the occurrence of item-writing flaws (IWFs) in the Clinical Therapeutics Module (CTM) sequence of courses at a college of pharmacy at an academic health center in the southeastern United States. The goals of the study were to: (1) identify the most common item-writing flaws on examinations in the CTM sequence of courses, (2) determine what percentage of item-writing flaws included on the CTM examinations contain one or more IWFs, and (3) to examine the relationship between the most frequently occurring IWFs and test item psychometric parameters including item difficulty, item discrimination, and average item answer time.

A total of 1,373 test items from 34 locally developed summative examinations of the second- and third-year CTM sequence of courses during the 2017-2018 academic year comprised the item pool. A stratified random sample of 313 items was used to assure proportionate representation from each course. Eight criteria from the Item-Writing Flaws Evaluation Instrument (IWF EI) were used to identify any item writing flaws in each of the 313 items.

Spearman's rho correlations were conducted to examine the strength and direction of the relationship between the most common item-writing flaws and the psychometric indices, including item difficulty, item discrimination, and average answer time to determine the influence of writing flaws on student performance.

Findings of the current study suggest that item-writing flaws are common within the clinical therapeutics module examinations, with 37% of items having at least one item-writing flaw. Given the use of exam results for program accreditation, the results point to a clear need to examine and improve locally developed measures in pharmacy education programs to ensure the validity of inferences and decisions made on the basis of test scores. This study provides additional guidance for pharmacy educators to support needed improvements of multiple-choice question writing and test design.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	2
ABSTRACT	5
LIST OF TABLES	9
LIST OF FIGURES	10
CHAPTER 1: INTRODUCTION	11
History of Assessment in Higher Education	11
Assessment as Curricular Quality Improvement	13
Statement of the Problem	15
Rationale and Purpose of the Study	15
Background	17
Multiple-Choice Examinations to Measure Pharmacy Student Learning	19
Research Questions	22
Methodology	22
Summary	24
CHAPTER 2: REVIEW OF THE LITERATURE	26
Search Strategy	26
Theoretical Foundation	27
Test Reliability and Validity	30
Sources of Validity Evidence	31
Sources of Reliability Evidence	34
Best Practices for Multiple-Choice Item Construction	36
Content Guidelines	40
Format Guidelines	41
Style Guidelines	41
Writing the Stem Guidelines	43
Writing the Options Guidelines	44
Prevalence of Item-Writing Flaws	46
The Effect of Item-Writing Flaws on Student Achievement	52

The Use of Item Analysis to Improve Assessment of Student Achievement and Validity	58
Summary of the Review of Literature	60
Definition of Terms.....	61
CHAPTER 3: METHODOLOGY	64
Study Context.....	65
Research Design.....	68
Item Pool and Sampling.....	69
Instrumentation	71
Procedure	73
Interrater Reliability.....	76
Data Analysis	79
CHAPTER 4: FINDINGS.....	81
Descriptive Statistics Summaries of Psychometric Indices of Items.....	82
Analysis for RQ1: Summaries of Survey Responses on Most Common Item-Writing Flaws in the Clinical Therapeutics Module Sequence of Courses	82
Analysis for RQ2: Percentage of Items from Locally Developed Summative Examinations for 12 Clinical Therapeutics Module Courses Containing One or More Item-Writing Flaws.....	85
Analysis for RQ3: Results of the Correlation Analyses Between Most Common Item-Writing Flaws in the Clinical Therapeutics Module Examinations and Psychometric Indices	86
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS	95
Introduction.....	95
Discussion of the Findings.....	95
Limitations	107
Recommendations.....	108
Conclusion	111
REFERENCES	113
APPENDIX A - ITEM WRITING FLAWS EVALUATION INSTRUMENT GUIDE.....	A-1
APPENDIX B - CURRICULUM VITAE	B-1

LIST OF TABLES

Table 1. Explanation of Discrimination Values for Exam Items.....	29
Table 2. List of Clinical Therapeutics Module Examinations	70
Table 3. Results of the Interrater Reliability Analysis.....	78
Table 4. Descriptive Statistics Summaries of Psychometric Indices	82
Table 5. Frequency and Percentage Summaries for the IWFEI.....	84
Table 6. Results of the Spearman’s Correlation Analysis	87
Table 7. Results of the Point-Biserial Correlation Analysis.....	92

LIST OF FIGURES

Figure 1. Institute for Healthcare Improvement PDSA model for continuous improvement.....	14
Figure 2. Classical test theory.....	28
Figure 3. Common guidelines for measuring Cronbach’s alpha.	35
Figure 4. Example of a multiple-choice question	37
Figure 5. Haladyna & Rodriguez (2013) Revised Taxonomy of Item-Writing Guidelines	39
Figure 6. Item writing flaws evaluation instrument.....	73
Figure 7. Example of the various options available for an item analysis report from the ExamSoft Administrator Portal.	75
Figure 8. An example of a modified item analysis report from the ExamSoft™ Administrator Portal.	76

CHAPTER 1: INTRODUCTION

Assessment has become an increasingly significant practice in doctor of pharmacy degree programs since the Accreditation Council of Pharmacy Education (ACPE) revised its standards and guidelines in March 2003. In an effort to ensure continuous curricular improvement and transparency, ACPE requires evidence that students are meeting intended educational outcomes and professional competencies as part of its accreditation process. Within Section I of the ACPE Standards (i.e., Educational Outcomes), Standard 24 describes the assessment elements necessary for ensuring that students are prepared to enter pharmacy practice. Specifically, Standard 24 requires colleges and schools of pharmacy (C/SOP) to develop, provide resources for, and implement an assessment plan to measure student achievement of educational outcomes at specific milestones during the doctor of pharmacy program to assure that students are prepared to enter practice (ACPE Standards, 2016). “Assessment activities must employ a variety of valid and reliable measures systematically and sequentially throughout the professional degree program. Colleges and schools of pharmacy must use the analysis of assessment measures to improve student learning and the achievement of professional competencies” (ACPE Standards, 2007, p. 27). Since the adoption of the ACPE Standards, 2007, many accredited C/SOP have adopted computer-based testing systems to aid in collecting assessment data to measure student learning and guide instructional practices and policies.

History of Assessment in Higher Education

“Assessment is one area where notions of truth, accuracy and fairness have a very practical application in everyday life” (Williams, 1998). Most, if not all, assessment practitioners would agree that assessment exists primarily to gauge whether students are meeting the intended

educational outcomes of academic degree programs. Assessment in higher education emerged as a recognizable movement in 1985 at the First National Conference on Assessment in Higher Education held in Columbia, South Carolina (Ewell, 2002). The origin of assessment as a recognizable movement in post-secondary education stemmed from political concerns about the quality and cost of higher education, which led to the publication of several reports including the *Involvement in Learning* (1984) and *Integrity in the College Curriculum Report* (1985). These and other reports written during this period questioned the quality of education and challenged educators to bring about programmatic and curricular improvement by developing and revising educational outcomes. Ewell (2002) states that the central argument in these reports was the need for coherent curricular experiences that could be shaped by the ongoing monitoring of student learning. As Angelo (1985) noted:

Assessment is an ongoing process aimed at understanding and improving student learning. It involves making our expectations explicit and public; setting appropriate criteria and high standards for learning quality; systematically gathering, analyzing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance. Assessment helps us create a shared academic culture dedicated to assuring and improving the quality of higher education (p. 7).

To address the concerns about quality and cost of higher education, in 1988 Secretary of Education William Bennett mandated federally-approved accrediting organizations to provide evidence of institutional outcomes in their criteria for accreditation (Anderson, Anaya, Bird, & Moore, 2005). Anderson et al. (2005) contend the core of this mandate required programs to

document what their graduates would be able to do upon completion of their academic program (outcomes) and to provide tangible evidence that students met those expectations (assessment).

There are two paradigms of assessment that are present today which offer different insights about what assessment is and how the process should be operationalized within institutions of higher education. The first paradigm is that of continuous improvement; to enhance teaching and learning. This paradigm emerged out of concern about weaknesses in the quality of higher education in the mid-1980s. The second paradigm is that of accountability; which is concerned with demonstrating value and worth to decision makers, policymakers, and other stakeholders (Ewell, 2009).

Assessment as Curricular Quality Improvement

Assessment in pharmacy education is seen as an opportunity for continuous quality improvement (CQI) of the curriculum. Dew (2004) defines continuous quality improvement as the body of knowledge that helps us learn how to better facilitate the learning that occurs through teaching and research. In higher education assessment, CQI is essentially the ongoing cycle of collecting data and using results to make decisions to gradually improve student learning. The Deming Cycle, also known as the Plan Do Study Act (PDSA) Cycle, is a continuous quality improvement model that includes a sequence of four repetitive logical steps to facilitate continuous improvement and learning. Figure 1 outlines the Institute for Healthcare Improvement PDSA model for continuous improvement. The PDSA Cycle begins with the Plan step, which includes identifying a goal or establishing a purpose, determining measures of success, and developing a plan of action. The second step is Do, the implementation or execution of the plan of action. The Study step reviews and monitors the results to assess the plan for either indicators of progress and success or problems and areas for improvement. The Act step closes

the PDSA cycle, integrating the learning accumulated from the entire process, which can be used to adjust the goal, change methods, or expand learning.

In health sciences, we expect professionals to process large amounts of information in order to make decisions about patient care (Masters, Hulsmeyer, Pike, Leichty, Miller, & Verst, 2001). The mission of C/SOP is to educate and produce competent student pharmacists who are “practice ready” upon graduation. Therefore, the Doctor of Pharmacy (Pharm.D.) curricula incorporates clinical therapeutics modules (CTMs) that allow student pharmacists to apply integrated content from pharmacotherapy, medicinal chemistry, pharmaceuticals, and pharmacology. The CTMs include rapid content delivery, a significant study load, and frequent high-stakes examinations, which are criterion-referenced assessments in nature. Given the changes in accreditation requirements emphasizing evidenced-based reporting metrics, it is essential for faculty to ensure that the assessment practices used provide valid and reliable sources of evidence that students possess the requisite professional competencies and behaviors to progress through the program and to enter pharmacy practice.

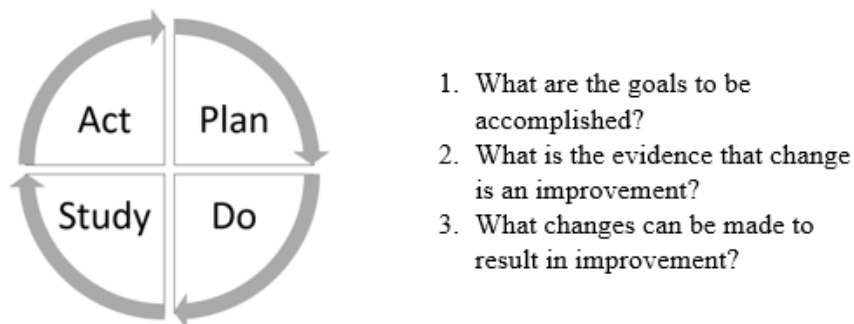


Figure 1. Institute for Healthcare Improvement PDSA model for continuous improvement

Statement of the Problem

Faculty have an ethical obligation to ensure that the scores from examinations are valid and reliable measurements of learning so that students do not fail tests or courses due to poorly written exam items (Hicks, 2014). As such, it is important to examine the extent to which item-writing flaws are present in Clinical Therapeutics Module examinations as these flaws may contribute to the inaccurate measurement of pharmacy students' content knowledge and professional skills. For example, item-writing flaws can result in distorted test results and lowered test reliability (Camili & Shepard, 1994). Downing (2005) states that flawed test items can affect student performance on multiple-choice question (MCQ) examinations by making items easier or more difficult to answer. Conversely, well-constructed test items are essential to accurately assess student learning. If tests are not well constructed, assessments of student performance may be invalid (Tarrant, Knierim, Hayes, & Ware, 2006, Rudolph et al., 2019). The presence of item-writing flaws on pharmacy assessments threatens the validity of the inferences made, or conclusions drawn, on the basis of the examination scores. Thus, any decisions about the academic progression of pharmacy students made by administrators based on these examinations scores are questionable when flaws exist. Research is needed to understand the extent of item-writing flaws on pharmacy assessments and to establish a baseline understanding of the scope of the problem and impact on student achievement.

Rationale and Purpose of the Study

Multiple-choice question (MCQ) examinations are a common method of assessing student learning in pharmacy education. It is reasonable to expect test items to be well-

constructed given that test grades affect students' academic progression. Because of the high-stakes nature of MCQ examinations in the clinical therapeutics module sequence of courses, it is imperative that pharmacy educators have an evidence-based understanding of the best practices for writing MCQ items. It is also important that faculty are aware of common item-writing flaws in pharmacy education and the effect that poorly written items can have on item performance and overall exam performance. There is a considerable amount of literature available on the effect of item-writing flaws and student achievement in nursing education and medical education (Downing, 2005; Masters et al., 2001; Pate & Caldwell, 2014; Tarrant, Knierim, Hayes, & Ware, 2006); however, there are fewer than four published studies related to the item-writing flaws in pharmacy education.

This study is designed to address this gap in the pharmacy education literature by examining the quality of MCQ examinations for Clinical Therapeutics Modules as part of pharmacy education. Identifying the prevalence of common item-writing flaws (IWFs) and the effect of item-writing flaws on item performance will help to better prepare pharmacy faculty to write well-constructed MCQs for assessing student achievement of educational outcomes. If it is found that IWFs adversely affect item performance, this study will provide evidence demonstrating the need to incorporate item-writing training into a new faculty orientation or into a faculty development program at C/SOP. More importantly, the results of this study could be instrumental to the American Association of Colleges of Pharmacy (AACP) in enhancing item construction resources for faculty at C/SOP who are using MCQ examinations to assess student achievement of educational outcomes. Over the years, AACP has responded to the 2016 ACPE Standards and Guidelines by providing professional development in the form of institutes, assessment management systems, and pre-conference workshops to enhance the knowledge and

skills of faculty who are responsible for educating future pharmacists. This study will contribute to the body of evidence about the effect of item-writing flaws on examinations in pharmacy education. Previous studies have focused on item-writing flaws and item construction in nursing education or medical education; however, no research is available on these issues in pharmacy education. In addition, this study has an opportunity to improve the validity and reliability of high-stakes examinations in pharmacy education, providing a more accurate assessment of student achievement of professional competencies.

Background

Colleges and schools of pharmacy (C/SOP) use direct measures of assessment to provide tangible and measurable evidence of student learning, with MCQs being one of the most common formats used in medical and pharmacy education to assess students' knowledge, skills, and abilities (Pate & Caldwell, 2014). Multiple-choice questions are frequently used in pharmacy education because they can be graded quickly and efficiently using software programs. Additionally, multiple-choice exams are designed to assess student knowledge in a content area and provide objective score data for a large number of items and a large number of test takers (Epstein 2007; Kane, 2006; McBrien, 2018). However, the topic of item writing may not be included in the typical faculty orientation or onboarding at C/SOP in the United States, therefore, many pharmacy faculty may not have received adequate professional development on how to construct test items in accordance with assessment design best practices.

Given that many pharmacy faculty have not been trained to write psychometrically sound assessment items, there is greater likelihood of introducing item-writing flaws on exams that may impact item performance, mastery of learning objectives, student exam performance, and limit the reliability of scores and validity of score interpretation. Many C/SOP in the United States

primarily rely on faculty-written, multiple-choice items or items from publisher-provided test banks to assess student achievement of educational outcomes. Faculty could assume that publisher-provided test bank items have been vetted for quality. Richman and Hrezo (2017) questioned the quality of publisher-provided test bank items, finding them to be poorly constructed. Without formal faculty development programs in assessment, many pharmacy faculty may not be familiar with the literature on best practices for the assessment of student learning, especially for test development and item construction, due to the limited literature in the mainstream medical and pharmacy academic journals (Pate & Caldwell, 2014). Moreover, the limited exposure of pharmacy faculty to professional development and peer review resources for item construction threatens the validity of assessments by introducing construct-irrelevant variance, which introduces extraneous, uncontrolled variables that affect assessment outcomes. The lack of training for faculty members who facilitate learning sessions and construct multiple-choice items contributes to disparities in how well or how poorly items are written (American Board of Emergency Medicine [ABEM], 2018; American Board of Internal Medicine [ABIM], 2018; American Board of Physician Specialties [ABPS], 2018). According to Tarrant (2008), multiple-choice questions (MCQs) on many internally developed, cross-discipline examinations are poorly constructed because very few teaching faculty have adequate training in writing high-quality test items. The increased accreditation demands on C/SOP to measure and report student outcomes has required faculty in pharmacy education programs to provide reliable and valid measures of student learning, thus heightening the need for expertise in item-writing among pharmacy faculty.

Multiple-Choice Examinations to Measure Pharmacy Student Learning

Pharmacotherapy course coordinators spend a considerable amount of time preparing and developing course embedded assessments and classroom assessments to measure student achievement against a set of predetermined criteria, outcomes or standards. The assessments are typically high-stakes, summative multiple-choice, case-based exams. Minimal research has been conducted to examine the quality of internally-developed items for pharmacy assessments. Although many of the textbooks used are accompanied by supplemental resources, such as test banks, to help instructors with assessment activities, these resources may lack evidence of best practices in multiple-choice item construction and psychometric test theory and practice (Masters et al., 2001). There can be significant deficiencies in examinations prepared by classroom instructors. Such deficiencies may include item-writing flaws which include excessive verbiage, longest answer choice is correct, responses that include all-of-the-above, responses that include none-of-the above, implausible distractors, or a stem that lacks direction. Research indicates that item-writing flaws can adversely affect student performance on examinations (Downing, 2002; Harasym, Leong, Violator, Brant, and Lorscheider, 2018).

The National Board of Medical Examiners (NBME) offers a comprehensive handbook on best practices for item writing for health sciences instructors and test administrators, which was updated in 2016 (Case & Swanson, 2001; Paniagua & Swygert, 2016). In addition, the *Standards for Educational and Psychological Testing* includes guidance on test construction, evaluation, and documentation (2014). The recurring theme related to best practices for item writing in the numerous guides for writing psychometrically-sound multiple-choice items, suggests that multiple-choice questions should have one best answer. McBrien (2018) stated that each multiple-choice item should stand on its own, each item should measure knowledge acquisition

related to one topic or idea, and distractors should be plausible, clearly incorrect, and avoid cues to the correct answer choice.

Haladyna, Downing, and Rodriguez developed rules of item writing in a taxonomy of 31 item-writing guidelines categorized by content concerns, formatting concerns, style concerns, writing the stem concerns, and writing the options concerns (Pais et al., 2016). There are well-defined, evidence-based principles and standards used to distinguish an effective item from an ineffective or poorly constructed item (Haladyna, Downing, & Rodriguez, 2002). An item-writing flaw is any item that violates one or more of these standard item-writing principles (Downing, 2002; Downing, 2005). There are several empirical studies on item-writing flaws and their effect on psychometric properties of an exam. The findings of a study of item response options conducted by Harasym, Leong, Violato, Brant, and Lorscheider (1998) suggest that using “all of the above” (AOTA) and “none of the above” (NOTA) response choices greatly alter the mean performance scores by students on an MCQ examination.

Other research demonstrates the general effect of item-writing issues on test performance. A study conducted at the University of Illinois at Chicago, Department of Medical Education examined the effects of violations of standard multiple-choice writing principles on test characteristics, student scores, and pass-fail outcomes (Downing, 2005). The study examined the effect of item-writing flaws on the psychometric characteristics of exams that were developed internally at the University of Illinois at Chicago. Using the item-writing taxonomy developed by Haladyna, Downing, & Rodriguez (2002), items were classified as standard or flawed, if they violated at least one of the guidelines established. Downing (2005) found that item-writing flaws were associated with more student failures than comparable items without flaws. The results of the study showed that the increased number of flawed items contributed to the likelihood of poor

performance on the exam, (Downing, 2005). A pass-fail agreement analysis revealed that 102 of the 749 students (14%) in the study passed the standard items and failed the flawed items. Inclusion of flawed exam items reduced the proportion of students meeting or exceeding the passing score. Therefore, item-writing flaws adversely impacted the scores of some of the medical students by incorrectly classifying students as having failed when they should have been classified as having passed (Downing, 2005). These findings suggest that faculty development programs focusing on principles of effective objective test writing should be increased and are necessary to increase the validity of the exam.

Furthermore, Pais et al. (2016) examined the impact of anatomical sites and the presence of item-writing flaws (IWFs) on the psychometric indices of MCQs. Anatomical sites are defined as structures of the human body. Similar to other research cited, the study used item-writing guidelines developed by Haladyna et al. divided into five categories: content guidelines, formatting guidelines, style guidelines, writing the stem guidelines, and writing the options guidelines. The results of the study suggest the categories related to content, writing the stem and writing the options had an adverse effect on the psychometric indices of the test. Specifically, results showed a higher difficulty index and lower discrimination index (Pais et al., 2016). This study suggested the presence of IWFs on an exam could make items more difficult for some students, thereby impacting students' performance on an exam. Moreover, the results of a study conducted in medical education in which researchers evaluated the quality of internally-developed examinations at three U.S. medical schools revealed that the overall quality of questions used on the examinations was low (Pais et al., 2016). Hence, the literature points to the need to improve assessment quality through enhanced item-writing practices and faculty development. This study and others reviewed in this section provide evidence that it is important

to eliminate or minimize item-writing flaws to ensure the reliability and validity of the interpretation of scores from the exams.

The presence of item-writing flaws on high-stakes summative assessments may contribute to increased test anxiety, confusion, and frustration by exam takers (Haladyna, 2004; Sansgiry, Bhosle, & Sail, 2006). “Decisions about the academic progress of a pharmacy student should be based on the interpretation of scores that are reliable and valid,” (McBrien, 2018, p. 3). Therefore, it is essential that the foundation of the clinical therapeutics module exams, and the items included on those exams, be constructed based on best practices for item writing.

Research Questions

The primary research questions (RQ) are:

RQ1. What are the most common item-writing flaws in the clinical therapeutics module sequence of courses at a school of pharmacy at a research-intensive academic health center in the southeastern United States?

RQ2. What percentage of items from locally-developed summative examinations for twelve (12) clinical therapeutics module courses contain one or more item-writing flaws?

RQ3. What is the relationship between the most common item writing flaws in the clinical therapeutics module examinations and the psychometric indices of items, including item difficulty, item discrimination, and average answer time?

Methodology

This study involved the implementation of a descriptive, correlational nonexperimental research design using existing data from examinations completed by second- and third-year students in a doctor of pharmacy program during the 2017-2018 academic year at a large, public research-intensive academic medical center in the southeastern United States. Multiple-choice

examinations from the clinical therapeutics module sequence of courses in the Pharm.D. curriculum were chosen because the content from the sequence of courses aligns with the Center for the Advancement of Pharmacy Education (CAPE) 2013 Educational Outcomes. The CAPE 2013 Educational Outcomes were created by focusing on the end of the Doctor of Pharmacy program and the knowledge, skills, and attitudes recent graduates should possess (Medina et al., 2013). The CAPE 2013 Educational Outcomes provide a structured framework for promoting and guiding curricular change, inspiring innovation, meeting challenges facing pharmacy education, and mapping and measuring programmatic outcomes. The 2017-18 CTM sequence of courses have defined learning outcomes that guide the development of each of the exams. The examinations in the CTM sequence of courses are an example of criterion-referenced assessments, which are designed to measure a student's academic performance against some predetermined standard, learning goal, performance level, or other criterion (Haladyna and Roid, 1983). Criterion-referenced tests assess how well a student masters a specific standard without consideration for how other students perform on the test. There was an average of three exams per module across the 12 CTMs, with a minimum of two exams per module and a maximum of five exams per module. There was an average of 29 questions per exam or a grand total in excess of 1300 questions across the 12 CTMs.

In this descriptive, correlational nonexperimental study, two raters evaluated MCQs from second and third year CTMs in the 2017-2018 academic year, which were designed to measure body systems and disease states, including cardiovascular, endocrinology, respiratory and immunology, psychiatry, neurology, oncology, infectious diseases, nephrology, dermatology, ear-nose-throat, gastrointestinal and nutrition, women's health, and critical care/toxicology. The primary rater evaluated 313 MCQs and the secondary rater evaluated 92 MCQs. The raters used

the Item Writing Flaws Evaluation Instrument (IWFEI) to evaluate each item (Breakall et al., 2019). The IWFEI was developed based on published best practices for item-writing and Haladyna and Rodriquez's guidelines for writing selected response items (2013). Haladyna and Rodriquez's revised taxonomy outlines 14 guidelines for item writing, which are categorized by style concerns, writing the stem, and writing the options.

A series of descriptive statistics were conducted to examine the initial research question. Descriptive statistics were appropriate for measuring the most frequently occurring item-writing flaws and combinations of flaws. The frequencies and percentages for the most frequently occurring item-writing flaws are displayed. Specific descriptive statistics were examined for each item-writing flaw identified including the mean, minimum, maximum, and the standard deviation. In the study, IBM SPSS Statistics 27® was utilized for the data analysis.

A series of Spearman's rho correlations were conducted to examine the strength and direction of the relationship between the most common item-writing flaws and the psychometric properties of CTM MCQs. A Spearman's rho correlation analysis was conducted to examine the relationship between the presence of item-writing flaws and the difficulty index. A Spearman's rho correlation analysis was conducted to examine the relationship between item-writing flaws and the discrimination index. A Spearman's rho correlation analysis was conducted to examine the association between the presence of item-writing flaws and average answer time.

Summary

Multiple-choice exams are common summative assessments used to measure student achievement in pharmacy education. Furthermore, decisions regarding student academic progression based on MCQ examinations have high stakes consequences (Hicks, 2014). This study can have a significant impact for positive social change by providing pharmacy faculty

insight about the effect of item-writing flaws on the psychometric indices of CTM exam items. The knowledge gained from this study can provide guidance about how test development can impact student achievement in C/SOP. Moreover, this study is significant because it helps faculty understand the impact of the most common IWFs on item performance and student achievement. The findings of this study can benefit faculty who are responsible for item writing and test construction at C/SOP in the United States. The social implications are beneficial to academic health centers and clinical educators responsible for test development.

Given the high-stakes nature of MCQ examinations, it is essential that faculty are well-equipped to write well-constructed test items to assure that examinations accurately estimate student achievement. This chapter presents a discussion of the background of assessment in higher education, considers the background of assessment as continuous quality improvement, and provides an overview of the purpose of assessment in pharmacy education and item-writing guidelines. The following chapter provides a thorough discussion of the empirical literature related to item-writing flaws in health sciences education and use of multiple-choice examinations as a means of addressing continuous improvement and accountability in health sciences education.

CHAPTER 2: REVIEW OF THE LITERATURE

This chapter includes an exploration of the topic by examining IWFs from historical and theoretical perspectives. Further, I examined the current empirical literature to provide context for potential relationships between IWFs and student achievement on multiple-choice examinations. Specifically, this chapter explores relevant literature to document a research base that following item-writing guidelines can improve the psychometric indices of examinations in the health sciences and how multiple-choice examinations can be used for continuous curricular improvement and accountability in health sciences education.

Search Strategy

The literature search included three stages: (a) an electronic search of the University Library's databases, (b) a hand search of reference lists from primary sources, and (c) an electronic search of the *American Journal of Pharmacy Education (AJPE)* and *Currents in Pharmacy Teaching and Learning*. I used these stages to identify literature on the effect of IWFs on high-stakes examinations in health sciences education, best practices in item writing, and psychometric properties of high-stakes exams. First, I conducted electronic searches of the university library databases using the following search terms in various combinations: continuous curricular improvement, criterion-referenced tests, health professions education, high-stakes examinations, item analysis, item writing, item-writing flaws, multiple-choice questions, multiple-choice examinations, norm-referenced tests, nursing education, medical education, pharmacy examinations, exam psychometrics, test construction, test items, and validity evidence. I did not place restrictions on publication dates. The ProQuest Dissertation Database was searched using similar keywords in various combinations.

In addition, I hand searched the reference list from each of the primary sources obtained. Lastly, I searched *AJPE* and *Currents in Pharmacy Teaching and Learning* using the same keywords in various combinations. Duplicate citations that appeared in multiple searches were removed, yielding more than 80 unique sources, separated into the following categories: assessment as continuous curricular improvement, classical test theory, item-response theory, item-writing guidelines, multiple-choice examinations in nursing education, multiple-choice examinations in medical education, item-writing flaws and item performance, and item-writing flaws and student achievement.

Theoretical Foundation

In 1904, researchers started examining the reliability of test scores such as the internal consistency of tests as part of the theoretical basis for test measurement (Kean & Reilly, 2014; Spearman, 1907; Traub 2005). Psychometricians created different coefficients to measure a test's internal consistency (e.g., Cronbach's α , Pearson's r). The correlation between test scores is negatively associated with the amount of measurement error that exists in the observed scores. For example, as the relationship between test scores increases, the measurement error in the observed scores decreases. Such diverse correlation coefficients resulted in the beginning of classical test theory (Novick, 1966). Psychometricians use two approaches to analyze items included on examinations: (a) item response theory (IRT) and (b) classical test theory (CTT). Item response theory emphasizes a student's performance on an exam and its relationship between individual items on the examination. Item response theory considers the number of questions answered correctly as well as the difficulty of the item. Essentially, the observed score is the score a test taker achieves on an exam, the true score is the accurate score that shows his or

her theoretical capability, and measurement error is the unexpected outcome of a weak measure of his or her true ability. The basis for classical test theory is presented in Figure 2.

$$\mathbf{O} = \mathbf{T} + \mathbf{E}$$

Observed assessment score	True score (can't be measured)	Measurement error (can be estimated)
---------------------------------	-----------------------------------	--

Figure 2. Classical test theory.

Classical test theory has been used in the assessment of undergraduate and graduate medical education to aid in developing examinations that are designed to measure student abilities in terms of item difficulty (De Champlain, 2010). Classical test theory provides useful information to aid in the analysis of test data. A goal of CTT is to improve the reliability and validity of tests by reducing error. Classical test theory enables researchers to create estimates of an item's difficulty and discrimination. The proportion of correct answers is negatively associated with the difficulty of the item, which represents the difficulty index (p). The difficulty index is a measure of the proportion of test takers who answered the item correctly, which is represented by the p -value. The p -value can range between 0.00 and 1.00, with a higher value indicating that a greater proportion of students responded to the item correctly. The higher the difficulty index value, the lower the item difficulty is, and the lower the difficulty index value, the higher is the difficulty of an item (Shete, Kausar, Lakhar, & Khan, 2015).

An exam item's *discrimination* represents the difference in proportion of correct scores between high and low performing students. Specifically, the discrimination index is a measure of how well an item is able to distinguish between those students who are knowledgeable of the content and those who are not knowledgeable of the content. There are three primary uses of the discrimination index for evaluating the quality of MCQs: (a) identifying items that have been

miskeyed, (b) identifying potentially flawed items, and (c) confirming the correct answer choice. Shete, Kauser, Lakhar, and Khan (2015) indicated the higher the discrimination index, the better an item differentiates between those students with higher test scores and those students with lower test scores. Table 1 outlines the range of values for the discrimination index based on Ebel & Frisbie's (1986) guidelines on classical test theory item analysis.

Table 1

Explanation of Discrimination Values for Exam Items

Discrimination Value	Explanation
Negative values	The item contains a flaw or is miskeyed.
D = 0 - 0.19	The item discriminates poorly, is unacceptable, and should be revised or eliminated.
D = 0.20 - 0.39	The item has acceptable discrimination; however, it could be improved following item writing guidelines.
D ≥ 0.40	The item is excellent and has high discrimination.

Classical test theory is a useful theoretical framework for the present study, as the theory helps to explain the relationship between item performance and test performance. “CTT is useful for assessing item discrimination and difficulty and the precision by which scores are measured by an examination,” (De Champlain, 2010, p. 112). De Champlain (2010) suggested that CTT can be applied to examinations to make judgements about whether the evidence supports retaining or excluding items from the scoring process. For example, CTT relates to the research questions for my study because it is suitable for examining the most common IWF in the clinical therapeutic module sequence of courses and the impact of the most common IWF on item performance (i.e., item difficulty, item discrimination, and average answer time). In addition, CTT is appropriate given that the intent is not to generalize beyond clinical therapeutics module examinations in pharmacy education due to the small class size selected for this study. The

indices of item difficulty and item discrimination associated with CTT can be used to examine the quality of measures, identify problematic items, and guide the refinement of test items that may produce more accurate estimates of achievement.

Test Reliability and Validity

Quality test items are essential for an exam to have reliability and to draw appropriate and accurate conclusions from resulting scores (Downing, 2005; Downing & Haladyna, 1997; Rudolph et al., 2019). Tests must be reliable and valid to support accurate and valid inferences based on the results (Sullivan, 2011). Reliability is one of the most important elements of test quality. Reliability refers to the consistency or reproducibility of a student's performance on a test. Validity refers to the extent to which evidence and theory support the interpretation of scores for a test (McMillan, 2016). The *2014 Standards for Educational and Psychological Testing* state validity is an essential criterion in test development and assessing quality of a test. Furthermore, validity refers to the accurate and meaningful interpretation of test scores and the reasonableness of the inferences drawn from the test scores (American Educational Research Association [AERA], 1989; Downing, 2002; Messick, 1989). Essentially, validity refers, in part, to the accuracy of the measurement (i.e., whether a test measures what it claims to measure). Sources of validity evidence for assessment instruments must demonstrate that the assessment measures what it was intended to measure. Specifically, validity evidence to support the inferences of scores from tests are typically found in the content of the examination and in the consequences. Content includes a description of how the test was created, who created the items, and whether the content included on the exam is appropriate. Consequential validity refers to the social positive and negative impacts resulting from the test. (Messick, 1989).

Sources of Validity Evidence

Validity is recognized to be “the most fundamental consideration” in developing and evaluating tests (AERA, American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11). Validity encompasses everything relating to the testing process that makes score inferences useful and meaningful (Albano, 2018). The goal of validity for examinations is to ensure that a representative sample of the intended learning objectives is measured and that students have satisfied the minimum performance level to be competent with respect to the stated objectives (Albano, 2018; AERA, APA, & NCME, 2014). In establishing validity evidence, one must consider whether there is sufficient proof to justify the inferences made based on the test scores. Downing (2003) described validity as a unitary concept with construct validity representing the entirety of validity. However, there are five sources of validity evidence outlined in the *2014 Standards of Educational and Psychological Testing* (AERA, APA, & NCME, 2014) that can be used to support or refute the interpretations of test scores from multiple-choice tests including: (a) evidence based on test content, (b) evidence based on response processes, (c) evidence based on internal structure, (d) evidence based on relations to other variables, and (e) evidence for validity and consequences of testing.

It is essential to keep the unitary nature of validity in mind when reviewing the different types of validity evidence (McDonald, 2018). Occasionally, it may be necessary to use multiple sources for validity evidence to support the interpretation of a test score. Therefore, it is possible that accumulated evidence from each of these types of validity is needed to make a judgement to support the interpretation of a test score.

Evidence based on content (content-related validity evidence) represents the test content and procedures for developing the test. Content-related validity focuses on elements of the

construct and how well they are represented in the test (Albano, 2018). Content-related validity aims to determine whether the test content covers a representative sample of the knowledge or behavior to be assessed (Paniaqua & Swygert, 2018). Test content addresses the themes, language, format of items, and questions included on the test. In addition, consideration should be given to test administration and scoring. There are three main steps in establishing content-related validity evidence: (a) define the purpose of the test and the construct to be measured, (b) create a test outline or blueprint of the test content, and (c) evaluate the test by pilot-testing with a subject matter expert to assess the extent to which the test captures the content domain and the extent to which the test items will adequately sample from the content domain. Developing a test blueprint is a critical step for establishing evidence for the validity of the inferences made based on test scores (McDonald, 2018). Subject matter experts, who may be faculty colleagues, typically evaluate the appropriateness of a test outline or a test blueprint. Tests that are created based on a test blueprint are more likely to have higher content validity. This means the scores from the test can be used to make a judgment about a student's knowledge in that specific content area. McDonald (2018, p. 25) suggested the following steps to enhance validity evidence based on test content:

1. Include specific, action-oriented, and measurable objectives.
2. Identify learning outcomes for the test.
3. Prepare a test blueprint based on items #1 and #2.
4. Write test items that align with the test blueprint.
5. Have peer reviewers examine the test blueprint and test items.
6. Provide adequate time for test completion.
7. Review the item and test analysis report.

8. Use the test only for its intended purpose.

The evidence based on response processes demonstrates that the assessment requires participants to engage in specific behavior necessary to complete a task (AERA, APA, & NCME, 2014; Downing, 2003). Response process is the evidence of data integrity that assures any source of error associated with the test administration is controlled or eliminated. This type of validity evidence addresses the use of documentation to assure data quality control related to the test.

Evidence based on internal structure demonstrates how the relationships between scores on individual test items align with the construct that is being measured. This type of validity evidence is provided when the relationship between items and parts of the instrument are empirically consistent with the theory or intended use of the scores, (McMillian, 2016). The *2014 Standards for Educational and Psychological Testing* address the intended and unintended consequences of test results used to make decisions about different groups. For example, if a test score is congruent with the proposed uses of the assessment. This relates to fairness in testing, which is not being examined in this study. The fairness argument focuses on whether an interpretation is equally plausible for different groups and whether the decision rules are appropriate for the groups (Kane, 2010).

Evidence based on relationship to other variables examines the relationship of test scores to variables that are external to the test (McDonald, 2018). This type of evidence may be useful when it is important to show a relationship with some other measure of performance. The *2014 Standards for Educational and Psychological Testing* focuses attention on the intended and unintended consequences of using test results to make judgements about different groups of students. Test fairness is a critical consideration in validity. The consequences of testing in

pharmacy education on students can be significant. For example, testing is one method of assessing student learning and competence to assure students are ready for professional practice. Student progression in pharmacy school is primarily determined by scores from multiple-choice tests. Test developers should gather evidence based on the consequences of testing by ensuring that scores on their assessments relate to intended future outcomes (McDonald, 2018).

Furthermore, Goodwin (2002, p. 104) stated that this type of evidence answers the following questions:

- How are the anticipated benefits of testing being realized?
- How do positive and negative unanticipated benefits occur?
- How are different consequences observed for different identifiable subgroups of examinees?

In certain testing situations, one source of validity evidence may be more relevant than another. However, it may be necessary to use multiple types of validity evidence to argue that the evidence supporting a test is appropriate.

Sources of Reliability Evidence

Reliability is used in two specific ways in measurement. First, reliability is used to determine if there is consistency of scores across replications of a testing procedure (i.e., whether the test gives you the same results each time in the same setting with the same student; AERA, APA, & NCME, 2014). Reliability of a test means that the resulting scores from a test are consistent and dependable. Second, reliability is used in CTT to describe the correlation between scores on two equivalent forms of a test (AERA, APA, & NCME, 2014).

Reliability can be measured in diverse ways. First, the relationship between all the variables can be calculated for internal consistency. Cronbach's alpha values can be used to test

internal consistency by measuring the relationship between all the variables. A value that is closer to 1.0 and greater than .70 is ideal (Nunnally, 1978). Figure 3 outlines the guidelines for measuring Cronbach's alpha. Second, the relationship between the two measurements is calculated by using Pearson's r for test/retest reliability, which is a more conservative estimate of reliability. Cronbach's alpha is commonly used in science education because the value can be calculated after a single administration of a test and the coefficients can determine if the test is accurately measuring the construct of interest (Taber, 2017). Third, interrater reliability can be measured to examine the impact of diverse raters by using Cohen's kappa (κ) coefficient. The kappa (κ) coefficient is a statistical measure of interrater reliability that is used to determine agreement between two raters. Reducing the error estimate in CTT improves reliability. One way to reduce error is to improve poorly constructed test items. There are numerous resources (e.g., books, articles, workbooks) available to provide guidance for constructing high-quality test items.

Cronbach's alpha	Internal Consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Figure 3. Common guidelines for measuring Cronbach's alpha.

Though some research has identified positive associations between MCQs and student achievement, the issues of reliability and validity remain. Analyzing the exams of over 1,000 nursing faculty, De Pew (2001) examined the relationship between the reliability and validity of MCQs and student achievement and found that there was no significant correlation between them. DePew found that there was no significant difference between high and low NCLEX-RN

success rates and the extent to which faculty engaged in validity assessment activities. Furthermore, there was no significant difference between high and low NCLEX-RN success rates and the extent to which faculty engaged in reliability estimation activities. Similarly, Rush et al. (2016) found that IWFs did not affect difficulty index or discrimination index values. Standard item-writing guidelines should be followed to improve clarity and consistency of test items to improve student achievement. This contrasts with the Pate and Caldwell (2014) where the research findings indicated that noncompliance with item-writing guidelines may negatively affect student performance. Therefore, efforts to increase faculty awareness and knowledge of the principles of effective item writing should be supported to enhance student achievement.

Best Practices for Multiple-Choice Item Construction

Downing (2002) stated that multiple-choice examinations can produce test scores that provide meaningful assessment of student achievement in medical education by discriminating those students who understand and can apply principles of the discipline from those students who cannot. However, it may be very challenging to reach such positive outcomes with multiple-choice assessments without a considerable level of practice and skill to develop test items that discriminate between low-achieving students and high-achieving students. Standards for best practices in item-writing have been developed and are supported by research (Case & Swanson, 1998; Haladyna & Downing, 1989). A standard MCQ consists of two parts: (a) a problem statement (stem) and (b) a list of suggested solutions (options). The stem is best written as a complete question or statement. The list of options should include one correct answer and a number of incorrect options known as distractors (see Figure 4). There are no clear theories related to the development of test items. However, a considerable amount of the literature on

item-writing guidelines is based on research conducted by Haladyna and Downing (1989).

Haladyna, Downing, and Rodriguez (2002) stated:

The scientific basis for writing test questions appears to be improving but very slowly.

We still lack widely accepted, question writing theories supported by research with resulting technologies for producing many questions that measure the complex types of student learning that we desire (p. 327).

In the absence of available theories, the Haladyna and Downing (1989) item-writing guidelines serve as a model for best practice. Their guidelines are based on an extensive review of more than 46 textbooks on classroom testing and research on existing item-writing guidelines (Ellsworth, Dunnell, & Duell, 1990; Haladyna & Downing, 1989; Haladyna et al., 2002; Stagnaro & Downing, 2006).

STEM - What disorder has unknown causes and is characterized by deafness, tinnitus, and dizziness?

Sensorineural deafness - **DISTRACTOR**

Meniere's disease * - **KEY**

Conduction deafness - **DISTRACTOR**

Otosclerosis - **DISTRACTOR**

Figure 4. Example of a multiple-choice question (Paniaqua & Swygert, 2018).

Many textbooks provide resources to help course coordinators with assessment activities such as test banks. However, these resources may lack evidence that the test bank authors used best practices for developing effective test items yielding poor quality test items (Masters et al., 2001). Furthermore, given the academic training for a pharmacy faculty member, there are significant inadequacies in exams (Masters et al., 2001). The professional curriculum for a doctor

of pharmacy (Pharm.D.) degree program includes clinical therapeutics, pharmaceuticals, pharmacology, medicinal chemistry, leadership, pharmacoeconomics, pharmacy communications, and experiential education. Courses in teaching, education, or assessment are not required courses in a Pharm.D. curriculum. Students would complete such a course as an elective, if it were available at the C/SOP. Haladyna and Downing (1989) examined 46 measurement textbooks dating back to 1935 that provided guidance and instruction on how to write MCQs. From this examination, Haladyna et al. (2002) developed a taxonomy of 31 item-writing guidelines, representing five categories: (a) content guidelines, (b) formatting guidelines, (c) style guidelines, (d) writing the stem guidelines and (e) writing the options guidelines. Figure 5 illustrates the taxonomy that was revised and republished by Haladyna and Rodriguez (2013). There are well-defined, evidence-based principles and standards used to distinguish an effective item from an ineffective or poorly constructed item, and an IWF is any item that violates those standards (Downing, 2002; Haladyna et al., 2002). When developing MCQ examinations, it is important to use best practices or item-writing guidelines to increase the exams validity. This study will focus on style guidelines, writing the stem guidelines, and guidelines for writing the choices. This study will not address content guidelines, given that the primary researcher and secondary researcher do not possess the background and experience in health sciences and pharmacy practice. Furthermore, this study will not address formatting guidelines because all of the items evaluated in this study were configured using the ExamSoft computer-based testing system defaults, which is for vertical display. However, detailed descriptions of each of the categories are presented in the next section.

Content Guidelines

- Base each item on one type of content and cognitive demand.
- Use new material to elicit higher-level thinking.
- Keep the content of items independent of one another.
- Test important content. Avoid overly specific and overly general content.
- Avoid opinions unless qualified.
- Avoid trick items.

Format Guidelines

- Format each item vertically instead of horizontally.

Style Guidelines

- Edit and proof items.
- Keep linguistic complexity appropriate to the group being tested.
- Minimize the amount of reading in each item. Avoid window dressing.

Writing the Stem Guidelines

- State the central idea clearly and concisely in the stem and not in the options.
- Word the stem positively; avoid negative phrasing.

Writing the Options Guidelines

- Use only options that are plausible and discriminating. Three options are usually sufficient.
- Make sure that only one of these options is the right answer.
- Vary the location of the right answer according to the number of options.
- Place options in logical or numerical order.
- Keep options independent; options should not be overlapping.
- Avoid using the options “None of the above,” “All of the above,” and “I don’t know.”
- Word the items positively; avoid negative words such as NOT.
- Avoid giving clues to the right answer.
- Keep the length of options about equal.
- Avoid specific determiners including always, never, completely, and absolutely.
- Avoid clang associations, options identical to or resembling words in the stem.
- Avoid pairs or triplets of options that clue the test taker to the correct choice.
- Avoid blatantly absurd, ridiculous options.
- Keep options homogeneous in content and grammatical structure.
- Make all distractors plausible. Use typical errors of test takers to write distractors.
- Avoid the use of humor.

Figure 5. Haladyna & Rodriguez (2013) Revised Taxonomy of Item-Writing Guidelines

Content Guidelines

While this study does not address the appropriateness of content concerns in the development of multiple-choice items for clinical therapeutics module examinations in pharmacy education, it is important to consider its relationship to IWFs and its impact on student achievement. Haladyna and Rodriguez (2013) addressed six specific concerns related to content in item writing. These include (a) basing each item on one type of content and cognitive demand, (b) using new material to elicit higher-level thinking, (c) keeping content items independent of one another, (d) testing important content and avoiding overly specific and general content, (e) avoiding opinions unless qualified, and (f) avoiding trick items (Haladyna & Rodriguez, 2013).

Research suggests that items included on a multiple-choice assessment or an item bank should be organized by topic or by competencies. A test blueprint would be beneficial to guide the development of the examination (Ray et al., 2018). Some faculty have a tendency to provide information in one item that helps students answer another item correctly. Guidelines suggest it is best to avoid creating dependency among items to prevent providing clues to another item's answers (Haladyna & Rodriguez, 2013). Test-wise students use such strategies to select answers they may not know. When considering content on an examination, faculty must make good judgements about specific content that is important to assess student achievement of learning outcomes. Faculty must create the perfect balance between how specific and how general each item must be to adequately reflect the desired content. To address the issue of content that is too general or too specific, it is best to have a committee of subject matter experts determine what is most important. The committee of subject matter experts could be used to conduct a collaborative review or peer review of test items prior to test administration. Test content should be determined by lesson objectives which are created by a consensus of subject matter experts.

Testing an opinion about something is unfair (Haladyna & Rodriguez, 2013). Haladyna and Rodriguez (2013) suggested it is acceptable to test an opinion only when it comes from a documented source, from evidence, or from a presentation cited in a curriculum. Haladyna and Rodriguez also suggested avoiding the use of trick questions on multiple-choice examinations. Item writers should not appear to deceive, confuse, or mislead students. Failing to adhere to item-writing guidelines can introduce variance in student performance stemming from construct irrelevant variance, such as a student's test-taking strategies and testwiseness. Trickiness in item writing development is a source of construct-irrelevant variance, which erroneously inflates or deflates test scores due to certain types of uncontrolled or systematic measurement error (Downing, 2002; Haladyna & Downing, 2004; Ray et al., 2018).

Format Guidelines

The format guidelines include only one recommendation for best practice, which is to format test items vertically versus horizontally (Haladyna & Rodriguez, 2013). Formatting items vertically minimizes the cramped spacing of test items that affect the look of the tests, which supports university initiatives in conservation (i.e., Go Green and sustainability initiatives). In addition, horizontal formatting may make the test more difficult to read, possibly lowering student achievement on the test. Most computer-based testing systems have default settings that present items in vertical format.

Style Guidelines

Consistency in style is essential in test development and administration. Style concerns can affect the validity of test score interpretations; therefore, it is very important to address style concerns in the test development phase by assuring a consistent and professional appearance is created throughout the test. McDonald (2018) states that consistency improves test validity and

reliability by reducing ambiguity, increasing item quality, and increasing student respect for the test. Furthermore, style guidelines include editing and proofreading items; using correct grammar, punctuation, and spelling; and minimizing the amount of reading required in each item (Haladyna & Rodriguez, 2013). Editing and proofreading test items for grammatical errors is a consistent theme throughout the literature on best practices for item writing. The purpose of proofing test items is to ensure that the test and all test items are accurately and completely presented (Haladyna & Rodriguez, 2013). Developing a style guide for testing that includes information about how each item is to be formatted, the use of acronyms and how acronyms should be presented, and a list of do's and don'ts would be helpful in adhering to best practices for item writing.

In team-taught courses, a course coordinator who has been trained and has considerable practice in item writing should be accountable for enforcing item-writing guidelines, proofreading items, and editing items before the items are added to an exam bank and included on an exam. The validity of test score interpretations can be improved by editorial work through proofing of the test and its items (Haladyna & Rodriguez, 2013; Panigua & Swanson, 2016). It is important that item writers use correct grammar, punctuation, capitalization, and spelling when developing exams as items with errors may confuse students. Following these guidelines can improve sentence structure, which improves clarity of test items for test takers.

Because establishing reliability and validity based on the inferences made from test scores is important, researchers suggest reducing the reading time for an item (Board & Whitney, 1972; Haladyna & Downing, 1989; Haladyna & Rodriguez, 2013). Test items with unnecessary content in the stem extend the reading time and lengthen the time it takes to complete an exam. McDonald (2018) suggests avoiding the use of excessively lengthy items if the information is not

necessary to answer the question correctly. This is particularly important in situations where there is a set amount of time to administer an exam, which is common in C/SOP. Extraneous information included in the stem of an item increases the ambiguity of the item and the time it takes students to process the statement to understand what is being asked.

Writing the Stem Guidelines

Writing the stem is as essential to test item development as editing for grammatical errors and punctuation errors. An essential goal in writing the stem is to state the central idea clearly and concisely in the stem and not in the options (Haladyna & Rodriguez, 2013). The stem should be written as short as possible, with completeness, clarity, and conciseness. Ideally, students should be able to read the statement and arrive at the answer without reading the options. It is common to have an unfocused stem, where there is minimal information provided in the stem and a substantial amount of content provided in the options. Students should not have to read through the options to answer the test question. Another common IWF is the inclusion of negative phrasing in the problem statement or case statement. An example of a negatively phrased stem would be: which is NOT the primary cause of cardiovascular disease? Haladyna et al. (2002) suggested that students have difficulty understanding the meaning of MCQs that include negatively phrased words.

There is a considerable amount of literature available that suggests using negative words in multiple-choice items can contribute to students having difficulty understanding the meaning of the negatively phrased items (Chiavaroli and Familiari, 2017; Haladyna et al., 2002). In testing environments where students may have a limited amount of time to complete an examination (e.g., 50-minute exam block), a negatively phrased stem may increase the time for students to read the question and correctly respond to the item. In addition, students may not process the

meaning of “NOT” and may forget to reverse the logic of the relation being tested (Haladyna & Rodriquez, 2013). Negative phrasing can cause unnecessary confusion and make items unnecessarily difficult. In the event that the item writer deems that it is necessary to include the negative phrasing, Haladyna and Rodriquez (2013) suggested that the phrase be emphasized by placing it in bold type, capitalizing the phrase, underlining the phrase, or using all of these.

Writing the Options Guidelines

There is a considerable body of literature that suggests that writing plausible distractors is very challenging (DiBattista & Kurwaza, 2011; Paniagua & Swygert, 2018; Haladyna & Rodriquez, 2013). Keeping distractors relevant is essential as they can affect item difficulty in the same way that the content of the stem does. Haladyna and Rodriguez (2013) stated that an effective or optimal distractor would be selected by lower achieving students and disregarded by higher achieving students. An additional guideline for writing plausible options is to provide only one correct answer. Researchers suggest having items peer-reviewed by faculty subject matter experts to assure there is one best answer per item (Haladyna & Rodriguez, 2013; Ray et al., 2018).

Faculty should consider varying the location of the correct answer according to the number of options to prevent test-wise students from correctly answering the question based on clues (Attali & Bar-Hillel, 2003). Options should always be presented in alphabetical, logical, or numerical order to minimize confusion and to facilitate ease of reading (Considine et al., 2005; Haladyna & Rodriguez, 2013; Nadeau-Cayo, 2013, Town, 2014). Furthermore, options or answer choices should always be presented in ascending or descending order. To expand on presenting item options in a logical or numerical order, consideration must be given to assuring that options do not overlap and are distinctly different. Most item-writing guidelines suggest item

writers avoid using options that include “All of the above” (AOTA), “None of the above” (NOTA), or “I don’t know.” The rationale behind avoiding the use of AOTA, NOTA, or “I don’t know” is that the options should include one correct answer. Haladyna & Downing (2002) suggest that there is no obvious benefit for omitting the correct answer from the list of options. However, there is conflicting information about the use of NOTA as an option (Pate & Caldwell, 2013; Nadeau-Cayo et al., 2013). Haladyna and Rodriguez (2013) suggest that the use of NOTA requires students to solve a problem rather than select the correct answer. Paniagua and Swygert (2018) suggested NOTA is problematic when students need to identify the correct answer and the options are not clearly true or false. “None of the above” options should be replaced with more plausible and specific items that minimize ambiguity by the student.

Another consideration when writing the options is that faculty should assure that items are written without giving clues to the correct answer. Haladyna and Rodriguez (2013) suggest making all options around the same length and avoiding the use of absolute terms (i.e., always, never, totally, absolutely, and completely). Distractors should be plausible, discriminating and have a similar length as the answer. Plausibility, another guideline for writing the options, refers to the notion that high-achieving students will answer the question correctly, while low-achieving students are less likely to answer the item correctly. Another commonality that emerged in item-writing guidelines is the use of humor. Introducing humor into a testing environment may encourage students to take the test less seriously. Therefore, researchers suggest it is best to avoid the use of humor when writing MCQs (Case & Swanson, 2002; Haladyna & Rodriguez, 2013).

Prevalence of Item-Writing Flaws

Writing psychometrically sound MCQs can be challenging, especially for novice item writers. Furthermore, it can be difficult to remember the numerous rules related to item writing at the time of test construction. Breakall et al. (2019) developed the Item Writing Flaws Evaluation Instrument (IWFEI) to assist faculty with assuring that MCQs adhere to the standards of accepted item-writing guidelines thereby improving the quality of MCQ examinations. The study conducted by Breakall et al. (2019), included two phases: (a) instrument development and (b) item analysis. The IWFEI was developed based on recommendations outlined in the literature as item-writing guidelines (Downing, 2002; Frey et al., 2005; Haladyna et al., 2010). The IWFEI instrument was pilot tested with four chemistry education graduate students who were given a 20-minute orientation on how to use the IWFEI and the instruction manual. The graduate students rated 10 general chemistry multiple-choice items individually and provided feedback as part of a focus group in two iterations. After testing interrater reliability and conducting item analysis, Breakall et al. (2019) found that the instrument was reliable. The instrument was found to have a high degree of interrater reliability with a 91.8% agreement and a Krippendorff's alpha of 0.836. Krippendorff's alpha is a reliability coefficient used to measure the interrater reliability among observers, judges, coders, and raters by calculating the disagreement rather than agreement. It is frequently used to quantify the extent of agreement between raters.

The second phase of the Breakall et al. (2018) study involved using the IWFEI to evaluate 1,019 multiple-choice items on 43 chemistry examinations. The items were developed by a committee of instructors who taught the different sections of the same course. A total of 33-unit exams and 10 final exams were analyzed. Findings from the study indicated that 83% of the items contained at least one IWF. The most common IWF was the "inclusion of implausible

distractors” where a student could “answer without looking at the answer choices” as the second most common (Breakall et al., 2019). However, because the instrument was validated using chemistry examinations, it is not clear how the instrument would perform in other examinations. Breakall et al. (2019) concluded that the use of the IWFEI by chemistry faculty could improve their assessment practices.

Rush, Rankin, and White (2016) evaluated the effect of IWFs and item complexity on item difficulty and discrimination using 1,925 examination questions administered to 112 veterinary students at Kansas State University. In 33.9% of the questions, Rush et al. identified one IWF. In 37.3% of questions, two or more IWFs were identified. Twenty-nine (28.8%) of items were free of IWFs. Rush et al. also found that item complexity was positively and significantly correlated with higher cognitive skills. As item complexity increased, item difficulty values decreased and item discrimination values increased. The most common IWFs identified in this study were awkward stem structure (29.4%), implausible distractors (22.9%), longest response correct (20.6%), true/false (17.1%), grammatical clues (15.4%), negative stem (11.8%), and vague language (11.2%). Item-writing flaws that included absolute terms, AOTA, and NOTA consisted of less than 9% of the IWFs. Although IWFs did not appear to adversely impact the item difficulty or item discrimination indices in this study, Rush et al., suggest faculty adhere to standard item-writing guidelines to improve clarity and consistency of examination items.

A study conducted by Masters et al. (2001) assessed MCQ test banks included in nursing education textbooks. Masters et al. (2001) found 2,233 violations of item-writing guidelines by performing a chi-square test to examine 2,913 MCQs used in nursing education that were selected from a convenience sample of 17 textbook test banks. The researchers selected a

random sample of 30% of the chapters from each of 17 test banks. Items were evaluated based on whether or not they violated generally accepted item-writing guidelines and the cognitive level of the question. Most of the common IWFs noted in this study included inadequate space ($n = 960$), uneven length options ($n = 239$), negative questions ($n = 166$), more than one correct answer ($n = 120$), nonplausible options ($n = 98$), and grammatical errors ($n = 97$). This finding contrasts with the findings from the study conducted by Breakall et al. (2019) in which implausible distractors was the most common violation of item-writing guidelines. Even though the findings of this study are not generalizable, they do support the idea that IWFs are prevalent in health sciences education. In addition, the results of the study were similar to the findings from the study conducted by Breakall et al. (2019), where the majority of items evaluated during the study violated at least one item-writing guideline and the inclusion of implausible distractors as the most common.

Similarly, Tarrant, Knierim, Hayes, and Ware (2006) examined 2,770 MCQs collected from tests and examinations from 2001 to 2005 using a chi-square analysis. The tests were administered in two baccalaureate nursing programs over a 5-year period. All clinical nursing and health assessment courses were included in this study. Tarrant et al. evaluated MCQs from eligible assessments, determined the source of the questions, identified duplicate questions, and examined all questions for IWFs, cognitive level assessed, and the distribution of correct answers. The items included in this study were evaluated based on 19 common IWFs. Four reviewers evaluated each MCQ for IWFs and cognitive level. Reviewers discussed items where there was no agreement during a consensus panel, in which discussion and agreement was made about categorization of the item and cognitive level. Tarrant et al. found that guidelines were violated in almost half (46.2%) of the questions. Of the 2,770 items evaluated, 1,490 had no

IWFs, 939 MCQs had one flaw, and 290 MCQ had two IWFs. The most common IWFs were “ambiguous or unclear information in the stem” ($n = 208$), negatively worded stems ($n = 192$), and implausible distractors ($n = 184$) (Tarrant et al., 2006).

Nedeau-Cayo, Laughlin, Rus, and Hall (2013) conducted a systematic replication study based on work conducted by Tarrant et al. (2006). The purpose of the study was to examine the frequency of IWFs in organizationally developed multiple-choice test questions in computer-based learning modules in a hospital setting. The sample for this study was composed of 405 computer-based learning (CBL) modules written by content experts from multiple disciplines. The tool used to evaluate the MCQ items was created by researchers based on the work of Tarrant et al. (2006). Tarrant et al. identified 19 IWFs that were consistent with the item-writing guidelines developed by Haladyna et al. (2002). A pilot study of the evaluation tool was conducted using 200 MCQs (Nedeau-Cayo et al., 2013). Each of the reviewers reviewed 50 questions each. The researchers individually and collectively identified the IWFs for four questions that were randomly selected. Researchers arrived at a consensus after the results were discussed. Nedeau-Cayo et al. (2013) found that most items included flaws by performing a frequency distribution and chi-square test. Interrater reliability was established by four researchers, and consensus of at least 90% was attained. The frequency of IWFs was reported using descriptive statistics. A chi-square test was conducted to determine the association between IWFs and the Bloom’s taxonomy level of the question. Of 2,491 MCQs, 386 (15.5%) items were not flawed, and 862 (34.6%) had more than one IWF (Nedeau-Cayo et al., 2013). Similar to the study conducted by Masters et al. (2001), the most frequent IWFs were all of the above ($n = 713$), more than one correct answer ($n = 387$), implausible distractors ($n = 380$), repeating word ($n = 314$) dissimilar length options ($n = 268$) and none of the above ($n = 215$). The findings from

this study support research conducted in similar studies in healthcare education, which found that most MCQ contain at least one IWF (Tarrant et al., 2006).

DiBattista and Kurzawa (2011) randomly selected 12 courses from 240 undergraduate courses in applied health sciences, business, humanities and sciences, math sciences, and social sciences from which to examine 1,198 multiple-choice items. The study included 16 submitted tests with a range of items from 24-211 and administered to a range of 109-547 test takers from freshmen to seniors. The results of the study indicated that more than 40% of the test items were flawed. A chi-square test, an ANOVA, a correlation analysis, and Mann-Whitney test, were performed. The results of the study indicated that discrimination coefficient was consistent with the findings of other studies (e.g., Tarrant and Ware, 2008). The findings of the study are consistent with other studies on IWFs and suggest that there is an opportunity to improve the quality of multiple-choice tests by using item analysis and by modifying distractors that adversely affect the discriminatory power of items (DiBattista & Kurzawa, 2011).

Pate and Caldwell (2014) measured the differences in student performance on multiple-choice items based on multiple-choice, item-writing guideline adherence and nonadherence in a cardiovascular module in the fall 2011 semester. The curriculum consisted of team-taught integrated modules that cover clinical therapeutics and basic sciences content of body systems such as endocrinology, dermatology, renal, and respiratory. The researchers chose the cardiovascular module because (a) it provided the best sampling of questions from faculty members in clinical therapeutics content and basic sciences content, (b) the module included more individual instructors than any other module, and (c) students were familiar with the testing practices in the doctor of pharmacy curriculum given that the module occurred in the third year of the curriculum. One hundred eighty-seven test items from four examinations were deidentified

to assure faculty anonymity. The two researchers evaluated each of the items based on adherence to item-writing guidelines proposed by Haladyna et al. (2002). Pate and Caldwell (2014) categorized each item as adherent or nonadherent. The researchers used a consensus process that included discussion of the perceived item-writing violations until agreement was reached. Nonadherent items were counted once regardless of the number of violations included in the item. Findings from Pate and Caldwell's study indicated that 17 of the 31 guidelines were violated, with 142 IWFs identified. The most common IWFs included avoiding AOTA ($n = 24$) and "minimize reading" ($n = 24$). Another common violation noted in the study involved writing the options, representing 76 of 142 violations. Similar to other studies, Pate and Caldwell (2014) suggested that nonadherence to item-writing guidelines may adversely affect student test performance (Board & Whitney, 1972; DiBattista & Kurwaza, 2011; Downing, 2005; Pham, Besanko, & Devitt, 2018, Reichert, 2011).

Multiple-choice assessments are used in continuing education programs to ensure that health care professionals maintain their knowledge and skills throughout their career as a healthcare professional. Stagnaro-Green and Dowling (2006) examined 40 MCQ tests from the *New England Journal of Medicine (NEJM)*, which provides its physician readers with an opportunity to earn weekly continuing medical education (CME) credits. Physicians answered a total of 40 MCQs. The MCQs were evaluated using 20 of 31 item-writing guidelines developed in the taxonomy developed by Haladyna et al. (2002). Findings from this study showed that each MCQ reviewed contained at least three IWFs, with a total of 203 IWFs represented in 40 items (Stagnaro-Green & Downing, 2006). The most common IWFs identified in this study included the use of verbose text ($n = 40$), unfocused stem ($n = 40$), and window dressing ($n = 29$). As with other studies on IWFs, this study revealed the *NEJM* used flawed MCQ in its weekly CME

program, which can introduce construct irrelevant variance (CIV) to the assessment, leading to the inaccurate interpretation of the scores from the assessment (Haladyna & Downing, 2004). Because the study is based on a nonrandom sample of CME items from the *NEJM*, the generalizability of the results from Stagnaro-Green and Dowling's (2006) study are limited.

The Effect of Item-Writing Flaws on Student Achievement

Research indicates that IWFs can have a significant impact on student achievement (Caldwell & Pate, 2013; Pate & Caldwell, 2014, Tarrant & Ware, 2008). The presence of IWFs on examinations can affect student achievement by making items more or less difficult to answer (Downing, 2002; Grolund, 2006; Haladyna & Rodriguez, 2002; Tarrant & Ware, 2008). Furthermore, the presence of IWFs on an exam can falsely inflate or deflate student performance on an exam, regardless of whether the student has the content knowledge (Breakall, et al., 2019). Well-written, multiple-choice exam questions can produce meaningful test scores and measure student achievement (Collins, 2006). Downing (2005) suggested that the lack of careful editing may lead to errors in the test that adversely affect students who have test anxiety. Furthermore, errors can be distracting to students, causing them to score lower than they would if the error was not present (Haladyna & Rodriguez, 2013).

Pate and Caldwell (2014) compared mean item difficulty between items that adhered to item-writing guidelines and those that did not adhere to item-writing guidelines and found that there was a significant difference between them. Seventeen guidelines were violated and 142 examination items were flawed (Pate & Caldwell, 2014). According to Pate and Caldwell, violating guidelines may have an adverse effect on student achievement. Furthermore, Pate and Caldwell (2014) indicated that the percentage difference in student scores between guideline adherent and guideline nonadherent items could account for a letter grade on any given exam.

However, there was no statistically significant difference in the mean discrimination between items that were written according to item-writing guidelines and those that were nonadherent (Pate & Caldwell, 2014). Therefore, the results of the study suggest that test items that are nonadherent to item-writing guidelines may adversely affect student achievement without providing significant discrimination between higher and lower performing test takers.

Caldwell and Pate (2013) used *t*-tests to examine student performance between guideline-adherent items (standard scale) and guideline-nonadherent items (nonstandard scale). Caldwell and Pate used three item-writing guidelines that received mixed endorsement in the educational measurement literature and research conducted by Haladyna et al. (2002). The three item-writing guidelines were selected for analysis because it was suggested that the nonadherent items would increase item difficulty and have a negative impact on student performance (Caldwell & Pate, 2013). The guidelines used in the study included: wording the stem positively and avoiding negatives such as “not” or “except,” developing as many plausible options as possible, and using none of the above guardedly. The researchers developed two sets (nonstandard and standard) of 15 items to test the effects of the guidelines on student achievement and item performance. The 15 pairs of items were separated into standard and nonstandard scales and appended to the end of the examination. Two examinations (standard and nonstandard forms), each with 115 questions, were alternately distributed at each seat in the testing auditorium. Students were able to self-select seats in the testing auditorium. Using 109 students who took the mile marker exam administered at the end of the spring semester of the first pharmacy year, the researchers found that students scored higher on the items that were written in accordance to item-writing guidelines than those that were not written using item-writing guidelines. Caldwell and Pate (2013) could not ensure that students fulfilling the standard scale were similar to those

completing nonstandard scale, since the students selected their seats for testing. However, there was no significant difference in student characteristics between students fulfilling the standard scale and those completing the nonstandard scale. A small sample size was used, and scale reliabilities were not calculated. There are similar perspectives concerning the effect of IWF on student achievement (Caldwell & Pate, 2013; Pate & Caldwell, 2014). This study reported similar results to Pate and Caldwell (2014) which suggest that nonadherence to item-writing guidelines can adversely affect student performance by nearly a letter grade with a corresponding increase in item discrimination.

Experts agree that IWFs can affect student achievement on examinations (Downing, 2005; Pate & Caldwell, 2013, Tarrant et al., 2006; Tarrant & Ware, 2008). Particularly, IWFs that use absolute terms (i.e., always and never), use all of the above, make the correct option the longest or most detailed, or use logical clues in the stem as to the correct answer make items less difficult for test takers to answer (Downing, 1989; Gronlund, 2006; Haladyna & Downing, 1989; Haladyna et al., 2002; Harasym et al., 1998; Tarrant & Ware, 2008). Researchers also suggested that using negatively worded items, unfocused or unclear stems, unnecessary information in the stem, and none of the above can make test items more difficult by making the items unnecessarily confusing for students (Chiavaroli, 2017; Crehan & Haladyna, 1991; Haladyna, et al., 2002; Tarrant & Ware, 2008).

Multiple-choice question examinations are an integral part of every health sciences professional's academic training. The results of MCQ examinations provide important information about a student's progress toward meeting educational outcomes for the degree program. One of the central purposes of test construction is to develop an examination that will accurately measure student knowledge and abilities in the specific content areas (De Champlain,

2009). Item-writing flaws can have a significant effect on high-stakes health sciences examinations (Downing, 2005; Pais et al., 2016; Pham, Besanko, & Devitt, 2018; Tarrant & Ware, 2008). Tarrant and Ware (2008) conducted a study to examine the impact of IWFs on student achievement in high-stakes assessments in a nursing program at a university in Hong Kong. All of the test items were reviewed by a four-person consensus panel. Items were reviewed for the presence or absence of 32 item-writing guidelines. Items were classified as flawed if the item contained at least one IWF, and items that did not include an IWF were classified as standard. Two separate scales were developed including a total scale, which represented the characteristics of the test as it was administered, and a standard scale, which represented the characteristics of a hypothetical test that included only unflawed items (Tarrant & Ware, 2008). Tarrant and Ware examined 10 test papers administered to 121 nursing students and found that 47.3% of all items were flawed. A total of 401 item-writing violations were identified in 314 (47.3%) flawed items. The most common item-writing violations identified in this study were unfocused stem/negative stem, unnecessary information in the stem, one correct answer, implausible distractors, greater detail in the correct option, and word repeats. Tarrant and Ware (2008) found that fewer nursing students passed the standard scale which included unflawed items when compared to the number of students who passed the total scale. The findings from the Tarrant and Ware (2008) study suggest that the mean difficulty scores show that flawed items were not substantially more or less difficult over the tests than were standard items. In addition, findings from the Tarrant and Ware (2008) study suggest that students were able to pass the total scales compared with the standard scales. The results of this study are not consistent with the findings from other studies where students perform worse on flawed items than unflawed items. Unlike other studies, Tarrant and Ware determined that IWFs did not

disadvantage students who were on the borderline of passing the test because the results suggest the students would have passed if the flawed items were removed from the exams. However, this study suggests that high-achieving students were more likely to be adversely affected by IWFs. Tarrant and Ware (2008) suggested that high-achieving students are more likely to rely on knowledge and reasoning rather than test-wiseness to answer high-stakes assessments. Test-wiseness is a skill which allows students to guess the correct answers on an item without knowing the content, thereby increasing their test scores (Downing, 2002). Students who are test-wise look for mistakes in test construction, search for any unintentional clues that can be found in a test, and make guesses.

Downing (2002) assessed the impact of sets of flawed items included on an educational achievement test. It was found that IWFs were associated with more student failures than comparable items without flaws (Downing, 2002). Using Haladyna's taxonomy for item writing, an item was classified as flawed if it violated at least one of the guidelines (Haladyna et al., 2002). In another study, Downing (2005) randomly selected four basic science tests given to medical students and found that 36% to 65% of the items on the four exams were flawed. In addition, standard items were easier than flawed items. Downing (2005) also found that there were more students who passed the standard items than those who passed the flawed items. Findings from these studies suggest that greater effort must be placed on assuring the quality of MCQs in health sciences education by providing faculty with adequate professional development and resources on best practices for writing MCQs. Furthermore, test items must be subject to peer review prior to administration along with a thorough review of an exam's item analysis report to ensure that assessments accurately measure student achievement.

Pais et al. (2016) evaluated the prevalence of IWFs using each of the Haladyna taxonomy categories in a clinical anatomy course. The most common IWFs in this study were related to writing the stem ($n = 150$) and writing the choices ($n = 166$). In this study, items were classified as standard or flawed. Standard items did not violate any of the item-writing guidelines referenced in the study. An item was considered flawed if it violated at least one item-writing guideline referenced in the study. Pais et al. (2016) found that the “writing the stem” and “writing the choices” categories had a negative impact on the psychometric indices of the MCQs, which represented a higher difficulty and lower discrimination indices. The two categories that had a negative effect were “writing the stem” ($n = 150$) and “writing the choices” ($n = 166$). Rules about “content concerns,” “style concerns,” and “content concerns without rule 4” had no impact no impact on the psychometric indices of MCQs. This study suggested the presence of IWFs on an exam could make items more difficult for some students, thereby affecting students’ performance on an exam. Even though Pais et al. (2016) used a small sample size (i.e., two medical students), the findings point to the need to improve assessment quality through enhanced item-writing practices and faculty development.

It is evident from the extant literature that IWFs can have an effect on student achievement on high-stakes health sciences examinations (Caldwell & Pate, 2013; Downing, 2002, 2004, 2005; Pais et al., 2016; Pate & Caldwell, 2014; Rush et al., 2016; Tarrant & Ware, 2008). A conclusion based on an analysis of all the studies reviewed suggest it is the responsibility of academic institutions that employ faculty to teach and develop assessments to provide adequate professional development and resources to enable faculty to develop psychometrically sound assessments. Furthermore, all of the studies reviewed emphasized the

importance of following best practices research in constructing items for multiple-choice assessments.

The Use of Item Analysis to Improve Assessment of Student Achievement and Validity

Many faculty who construct items for examinations should be concerned about the quality and validity of MCQ items that are used to measure student achievement. Additionally, faculty should be concerned about how students who take the examinations respond to the items. This is where statistical analysis can provide important evidence about the validity of exam content and the construction of the questions included on examinations (Chiavaroli & Familiar, 2011). Item analysis is essential in improving test items which will be used again in later tests. Quaigrain and Arhin (2017) suggested that instructors need to know how good the tests are and whether the test items are able to reflect the student's actual knowledge of course content included on the exam. Item analysis is a statistical process used to examine student responses to individual test items to assess the quality of those items and the test as a whole (Quaigrain & Arhin, 2017). Item analysis is useful in improving items for use on later examinations and for eliminating ambiguous, unclear, or misleading items. In addition, item analysis is valuable for enhancing a faculty member's skills in item writing, test construction, and identifying specific areas of the course content which need greater emphasis or clarity. Crisp and Palmer (2007) found that many faculty are disinclined to engage in item analysis, possibly due to the fact that many faculty are not specialists in educational theory. According to Crisp and Palmer (2007), validation of exams and their results are often based on "academic acumen rather than quantitative evidence" (p. 89). However, item analysis allows faculty to observe the characteristics of a particular item, which can be used to ensure that items meet an acceptable

standard for inclusion on the examination or improve the quality of the test item (Quaigrain & Arhin, 2017).

Item analysis assesses the reliability and validity of an examination by examining student performance of each MCQ and calculating psychometric data to determine whether the item should be reviewed, retained, or eliminated (Kheyami, Jaradat, Al-Shibani, & Ali, 2018). Item analyses can aid faculty in improving MCQs by providing evidence that the items have an acceptable or high discrimination index and difficulty index and an excellent distractor efficiency (Kheyami et al., 2018). Common item analysis indices include the difficulty index and discrimination index. It is important to keep in mind that an item may show low discrimination for examinations that have a wide variety of content and differing Bloom's taxonomy levels. With dichotomously scored items (i.e., items scored as incorrect or correct), item difficulty is typically indicated as a p -value, which reflects the proportion of students who answered the item correctly. Lower p -values indicate fewer students responded correctly, which may suggest the item was more difficult. Conversely, a higher p -value indicates more students answered the question correctly, which may suggest the item was easier. Quality control is important for test development (Quaigrain & Arhin, 2017). Criterion-referenced tests (CRTs), with their emphasis on mastery of criteria or outcomes, will have p -values of .9 or above (Professional Testing, Inc., n.d.). Norm-referenced tests (NRTs) are designed to be harder overall and to show a greater spread of the students' scores. Thus, many of the items on an NRT will have difficulty indexes between .4 and .6 (Professional Testing, Inc., n.d.). Therefore, it is essential that faculty use item analysis data to guide the decision-making process about items that should be retained, revised, or discarded to increase the reliability of the test.

Summary of the Review of Literature

Assessments that include IWFs have been found to negatively impact the ability of students to pass exams (Downing, 2002; Pate & Caldwell, 2014; Tarrant & Ware, 2008). In addition, IWFs on high-stakes health sciences examinations resulted in more difficult and fewer correct responses, indicating that it is important for faculty to follow best practice guidelines for developing test items (Caldwell & Pate, 2013; Downing, 2002, 2004, 2005; Pate & Caldwell, 2014; Rush et al., 2016; Tarrant & Ware, 2008). However, there is a gap in the literature with regard to the most common IWFs in the clinical therapeutic module sequence of courses in a doctor of pharmacy curriculum. Based on the literature, it stands to reason that the incidence of IWFs in clinical therapeutics module examinations may misrepresent the accuracy of pharmacy students' assessed content knowledge. The goal of this study is to investigate the frequency in which IWFs occur on high-stakes summative assessments for the clinical therapeutics module sequence of courses at an academic health center in the southeastern United States and to determine the impact of the most common IWFs on item difficulty and item discrimination.

Pharmacy educators carry a significant responsibility for assuring student pharmacists are prepared for practice. In Section I (i.e., Educational Outcomes) of the *ACPE Standards* (2016), Standard 24 requires C/SOP to develop, provide resources for, and implement an assessment plan to measure student achievement of educational outcomes at specific milestones during the doctor of pharmacy program to assure that students are prepared to enter practice. This mandate challenges pharmacy faculty to use assessment measures for continuous curricular improvement that are appropriate for student pharmacists and evaluate educational outcomes identified by the CAPE. Pharmacy faculty can use the findings of this study to improve item and test performance for clinical therapeutics module examinations. Pharmacy faculty may be able to decrease the

incidence of IWFs on assessments from the findings of the study by using established item-writing guidelines for MCQ test construction. This study can also help health professional faculty compare the most common IWFs in the clinical therapeutic module sequence of courses with those in other courses, such as pharmacy communications, pharmacoeconomics, and pharmacotherapy labs (Caldwell & Pate, 2013; Pate & Caldwell, 2014; Plaza, 2007).

It is critical to hold pharmacy educators accountable to the high standards of evaluation and assessment as outlined in Section I of the ACPE Standards (i.e., Educational Outcomes). When C/SOP create a culture of continuous curricular improvement, faculty and administration can be assured that MCQ examinations are accurately measuring student content knowledge and achievement of learning outcomes. Given that multiple-choice assessments are the most common method of measuring student achievement in pharmacy education, it is imperative that assessment results accurately reflect students' knowledge and competence. The resulting scores on pharmacy examinations can have lasting consequences for students, specifically decisions made about academic progression and career pathways (e.g., residencies, fellowships, hospital pharmacy, community pharmacy). Therefore, to assure academic integrity, pharmacy educators must be cognizant of the best practices for item writing and use item-writing guidelines to improve the validity and reliability of assessments to accurately measure student achievement.

Definition of Terms

All-of-the-above (AOTA) represents a multiple-choice question that has three or more distractors with “all of the above” or a variation of “all of the above” as one possible distractor (Haladyna, 2002).

Average item answer time is the mean amount of time students used to respond to an individual test item represented in minutes and seconds. (McBrien, 2018).

Difficulty index is the percentage of test takers who answered the item correctly, represented in values from 0.00 to 1.00 (Paniagua & Swygert, 2016).

Discrimination index is a measure of item performance that distinguishes how well an item differentiates between low and high performing students (Haladyna & Rodriguez, 2013).

Distractors are the incorrect options (Paniagua & Swygert, 2016).

Item refers to an entire question, including the stem and options. In multiple-choice testing, it is customary to speak of test “items” rather than questions because items may be presented in the form of statements rather than questions (Paniagua & Swygert, 2016).

Item-writing flaws are violations of accepted item-writing guidelines which can affect student performance on multiple-choice questions by the complexity or simplicity of the answer (Downing, 2005).

Key is the correct option (Paniagua & Swygert, 2016).

Multiple-choice question refers to a question that consists of a stem (i.e., a lead in question) followed by a series of choices, with one correct answer and anywhere from three to five plausible distractors (Paniagua & Swygert, 2016).

Negative phrasing is defined as a multiple-choice question that includes phrasing such as “all of the following are true except” or “which of the following is not” (Haladyna, 2002).

None-of-the-above (NOTA) represents a multiple-choice question that has three or more distractors with “none of the above” or a variation of “none of the above” as one possible distractor (Haladyna, 2002).

Options are all possible answers to the item, including the distractors (the incorrect answers to the item) and the key (the one correct, best answer to the item; Haladyna & Rodriguez, 2013).

Point-biserial correlation is a measure of item reliability. It correlates each student's response on a specific item with their overall performance on the exam (Rudolph et al., 2019).

Stem is the statement, question, chart, or graph portion of an item. The stem of the item should clearly present the central problem or idea (Haladyna & Rodriguez, 2013).

Test-wiseness is a student's capacity to utilize the characteristics and formats of a test and/or the test taking situation to receive a high score (Breakall et al., 2019).

CHAPTER 3: METHODOLOGY

The purpose of this study was to examine the frequency and nature of item-writing flaws (IWFs) on locally developed, high-stakes summative examinations for 12 clinical therapeutics module (CTM) sequence of courses at a pharmacy school located at a research-intensive academic medical center in the southeastern region of the United States using the Item Writing Flaws Evaluation Instrument (IWFEI; Breakall et al., 2019). The examinations in the CTM sequence of courses are an example of a criterion-referenced test, which are designed to measure a student's academic performance against some predetermined standard, learning goal, performance level, or other criterion (Haladyna and Roid, 1983). The scores of other students are not considered as they are with norm-referenced testing. The study examined the relationship between IWFs and the psychometric test properties, including item difficulty, item discrimination, and average item answer time. This chapter describes the methodology used, including the research questions, the research design, the item pool and sampling, instrumentation, procedure, data analyses, and limitations. The primary research questions (RQ) were:

RQ1. What are the most common item-writing flaws in the clinical therapeutics module sequence of courses at a school of pharmacy at a research-intensive academic health center in the southeastern United States?

RQ2. What percentage of items from locally developed summative examinations for 12 clinical therapeutics module courses contain one or more item-writing flaws?

RQ3. What is the relationship between the most common item-writing flaws in the clinical therapeutics module examinations and the psychometric indices of items, including item difficulty, item discrimination, and average answer time?

Study Context

Summative assessments used in the CTM at the school of pharmacy used in this study consisted primarily of conventional multiple-choice question (MCQ) examinations, which are criterion-referenced in nature. Multiple-choice questions were authored by faculty members who facilitated learning sessions related to the content covered in the CTMs. The course content was identified based on the course objectives and learning session objectives for each of the CTMs. The CTM course objectives were developed using the framework of the 2013 Center for the Advancement of Pharmacy Education (CAPE) Outcomes. The MCQs were submitted to the course coordinators at the start of the CTM. Clinical therapeutics module course coordinators collected and reviewed the multiple-choice items written by the faculty members responsible for lectures in the modules. The course coordinators were responsible for determining the number of items per content area for inclusion on examinations and providing guidance and resources to faculty on the best practices for item writing. The *Clinical Therapeutics Module Coordinator Guide* (2012) provided a checklist of best practices for item writing. The *Clinical Therapeutics Module Coordinator Guide*'s recommendations for test development suggests that 50% of the items included on each CTM examination are new items and not reused from previous examinations. Items could be flagged for replacement because of poor performance from the previous year or at random to assure the integrity of the exam. Items replaced at random generally support the overall integrity of the exam. The Office of Education and Assessment imported the MCQs into the exam bank, assuring the appropriate formatting, creating the

examinations, publishing the examinations to the students, providing the item analysis results to the course coordinators, coordinating student post exam review sessions, making adjustments to test items in the ExamSoft™ Administrator Portal, and pushing the final grades to Blackboard® Learning Management System. The secure computer-based testing platform from ExamSoft™ Worldwide included two products to facilitate the delivery of offline, computer-based examinations: (a) Examsoft™ Administrator Portal (EAP) and (b) Exemplify® (formerly SoftTest®). The EAP allowed for creation of selected response items and item banks that were tagged in an unlimited number of categories (e.g., learning outcomes, accreditation standards, Bloom's taxonomy, item author, disease states). In addition, the EAP maintained records of current and past reports of student performance by assessment and by individual items. The EAP provided a unique identification number for each item that is created and for each exam that is created. Furthermore, the EAP maintained a detailed log file that provided the date the item was created or modified, the creator of an item, the date an item was last included on an examination, and categories that were assigned to an item. Using exam data to provide students with feedback on their academic progress is important; however, using the data to improve test questions and exams as a whole is equally important. The EAP provided six types of reports that include data to track student, course, and even programmatic performance. The types of assessment reports included a summary report, item analysis, exam taker results, category reports, strengths and opportunities, and assessment performance reports. This study focused on the item analysis report, which provided data about the difficulty index, discrimination index, and the average answer time. The presence of IWFs on CTM examinations introduce systematic error that reduces reliability and validity and negatively affects the psychometric indices of the exam, including the difficulty index (p), discrimination index (d), and average answer time of the exam

items. Psychometric indices of test items included on CTM examinations may change due to the presence of IWFs.

Examplify was the platform that students downloaded and used to complete their exams. Examplify included a text highlighter, timer, calculator, reminder, notepad, and the ability to include images as part of the MCQ. Examplify also supported five question types, including multiple-choice, true or false, essay, fill-in-the-blank/matching, and hotspots. Second (P2) and third-year (P3) pharmacy students completed CTM examinations using Examplify. All examinations were offered during a 50-minute exam block from 8:00 a.m.–8:50 a.m. on specific days identified for P2 and P3 students. The password-protected examinations were released to students 24–48 hours prior to the exam session to download on their personal laptops. Although students downloaded the exam prior to the exam session, they were unable access the exam until the faculty member or test proctor gave them the second password at the start of the exam block. Randomized assigned seating charts for each exam were posted to Blackboard Learning Management System for students at 5:00 a.m. on the morning of the exam. Students reported to the designated classroom on the day of the scheduled exam, placing all notes, books, book bags, and cellphones in a designated area to support the integrity of examination administration. Students were given a paper copy of ExamSoft Notes Page with a place to sign the honor pledge. Instructions for completing the exam and the secondary password were displayed for students promptly at 7:55 a.m. At the end of the exam, students were instructed to show their ExamSoft exit screen and the signed ExamSoft Notes Page to the instructor(s) or exam proctor(s) before gathering their materials to leave the exam room. Students with accommodation letters for extended time or a quiet space to complete exams were required to provide copies of the accommodation letter from the Division for Academic Success to each of the course coordinators

and the director of education and assessment before the first day of the CTM. The director of education and assessment and the testing coordinator in the Division of Academic Success arranged for students with accommodations to report the University Testing Center prior to the start of the exam to complete the assessment. The item analysis reports were available to faculty as soon as all exams were uploaded by students. The exam review sessions were held within five days to provide students with feedback on the exam and their performance on the exam.

Research Design

I used a descriptive, correlational nonexperimental design to determine the prevalence of IWFs on CTM examinations and to determine the effect of IWFs on psychometric indices of CTM examination items. A descriptive, correlational nonexperimental research design is suitable for measuring the impact of the independent variable (IWFs) on the psychometric indices, which are the dependent variables (Leedy & Ormrod, 2016).

This study evaluated MCQs from P2 and P3 CTM examinations from the 2017-2018 academic year according to eight of the 15 criteria outlined in the IWFEI developed to identify the presence of IWFs on multiple-choice exams in general chemistry education. Criteria #9 was excluded from the data analysis given that all of the items included on the exams were developed based on the course objectives and learning objectives for each of the Clinical Therapeutics Modules. The second criteria (#10) was excluded from the data analysis because of the raters' limited knowledge of pharmacotherapy content, which limited the ability of the primary rater secondary rater to determine whether the options were plausible. The four remaining criteria that were excluded (12, 13, 14, and 15) relate to the overall test rather than individual items, which is outside of the scope of this study.

Item Pool and Sampling

The Doctor of Pharmacy program is a 4-year curriculum that includes 3 years of didactic coursework and introductory pharmacy practice experiences and one year of advanced pharmacy practice experiences. The CTMs include individual modules integrating the principles of medicinal chemistry, pharmacology, pharmaceutics, pathophysiology, and pharmacotherapy to the application of drug therapy in patients with diseases. There was an average of three exams per module across the 12 CTMs, with a minimum of two exams per module and a maximum of five exams per module. Each of the exams was designed to assess whether second- and third-year pharmacy students demonstrate specific clinical therapeutics knowledge and skills to assure readiness for Introductory Pharmacy Practice Experiences (IPPEs) and Advanced Pharmacy Practice Experiences (APPEs). A total of 1,373 test items from 34 locally developed summative examinations of the second- and third-year CTM sequence of courses during the 2017-2018 academic year comprise the item pool. All of the examinations included questions that were authored by full-time or affiliate faculty in the School of Pharmacy, based on course objectives, the inclusion of 50% new items on each exam, and all exams were completed using Exemplify®.

The Calculator.net Sample Size Calculator was used to determine the appropriate sample size for the study. The calculator required three sources of information to determine the appropriate sample size, including the confidence interval, margin of error, and the population. Results from the sample size calculation suggested that for a 95% confidence interval and an 5% margin of error, a sample of 313 items was needed.

A stratified random sample was used to assure proportionate representation of items from each course and that each item in the item pool had an equal chance of being selected for inclusion. The strata in the 1,373-item pool included each of the courses in the CTM sequence.

Proportionate stratification was used to ensure that the sample selected had a proportional number of items from each of the CTMs. Given that the maximum number of items included on an examination is 45, a random number generator, #Number Generator, was used to randomly select item number 11. As a result, a list including the unique ID for each item was generated for each of the CTMs. Proportionate stratification was calculated based on a sample size of 313. To achieve this, the desired sample size ($n = 313$) was multiplied by the proportion of units in each stratum. Therefore, to calculate the number of items in the cardiovascular CTM required in the sample, I multiplied 313 by 0.11 (e.g., $0.11 = 10\%$ of the item pool included items from cardiovascular module), which yielded a total of 33 items. Table 2 lists the twelve clinical therapeutics module examinations, the number of items included on each examination for each module, and the number of items that were sampled from each module. I selected every eleventh item in each of the modules to obtain the required number of 313 items.

Table 2

List of Clinical Therapeutics Module examinations with the number of items per examination and the Number of Sampled Items Per Module

Year / Course	Number of Examinations	Number of Items	Number of Sampled Items
P2 Fall 2017 Semester			
PHAR 544: Cardiovascular	4		33
Exam 1		36	
Exam 2		39	
Exam 3		36	
Exam 4		33	
PHAR 555: Endocrinology	3		24
Exam 1		30	
Exam 2		32	
Exam 3		35	
PHAR 603: Respiratory & Immunology	3		24
Exam 1		35	
Exam 2		35	
Exam 3		36	
P2 Spring 2018 Semester			
PHAR 604: Infectious Diseases	5		39

Exam 1		38	
Exam 2		38	
Exam 3		33	
Exam 4		36	
Exam 5		34	
PHAR 606: Nephrology & Urology	2		18
Exam 1		40	
Exam 2		38	
P3 Fall 2017 Semester			
PHAR 556: Neurology	4		36
Exam 1		42	
Exam 2		45	
Exam 3		40	
Exam 4		39	
PHAR 602: Psychiatry	3		24
Exam 1		34	
Exam 2		38	
Exam 3		36	
PHAR 605: Hematology & Oncology	3		24
Exam 1		40	
Exam 2		35	
Exam 3		40	
P3 Spring 2018 Semester			
PHAR 607: Dermatology & Ears, Nose, Throat	2		18
Exam 1		40	
Exam 2		40	
PHAR 618: Gastrointestinal & Nutrition	3		24
Exam 1		34	
Exam 2		35	
Exam 3		36	
PHAR 619: Women's Health Bone & Joint	2		22
Exam 1		39	
Exam 2		38	
PHAR 620 Toxicology & Critical Care	3		27
Exam 1		39	
Exam 2		39	
Exam 3		40	
Total Examinations / Total Number of Items	37	1,373	313

Instrumentation

The IWFEI was used to evaluate the 313 items included in the sample. Although the IWFEI was developed to analyze general chemistry exams for item-writing guideline violations, the instrument was used to determine if items from other disciplines were written in adherence to

item-writing guidelines published in the literature. The IWFEI was developed based on published item-writing guidelines (Downing, 2002; Frey et al., 2005; Haladyna, 2010; Town, 2014). The IWFEI takes into consideration the following item-writing guidelines in assessing the presence of IWFs:

- Use appropriate linguistic complexity;
- Minimize the amount of reading;
- State the central idea in the stem (avoid unfocused stem);
- Avoid negative phrasing in the stem;
- Use options that are plausible and discriminating;
- Include only one of the options is the right answer;
- Place options in a logical or numerical order;
- Keep options independent;
- Avoid using all-of-the-above and none-of-the-above;
- Avoid negative phrasing in the options, including True/False; and
- Avoid giving cues to the right answer, including
 - Keep length of options about equal;
 - Avoid specific determiners including always, never, completely, and absolutely; and
 - Keep options homogeneous in content and grammatical structure.

Using the IWFEI to determine how well MCQs in the CTM sequence of courses align with best practices for item writing will enhance the validity of the instrument. The IWFEI includes 15 criteria based on a review of item-writing guidelines outlined in the literature (Downing, 2002; Frey et al., 2005; Haladyna, 2010; Town, 2014). The IWFEI consists of 11

criteria that apply to individual test items and four criteria that apply to the overall exam. Given that the focus of this study was on IWFs, I used the eight IWFEI criteria (see Figure 6) that focus on the quality of exam items rather than those criteria to evaluate the overall exam.

Criteria	Guideline	Yes	No	Not Applicable
1	Is the test item clear and succinct?			
2	If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?			
3	If the answer choices are numerical, are they listed in ascending or descending order?			
4	If the answer choices are verbal, are they approximately the same length?			
5	Does the item avoid “all of the above” or “none of the above” as a possible answer choice?			
6	Does the item avoid grammatical or phrasing cues?			
7	Could the item be answered without looking at the answer choices?			
8	Does the item avoid complex K-type item format?			

Figure 6. Item writing flaws evaluation instrument. Adapted from “Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry,” by J. Breakall, C. Randles, and R. Tasker, 2019, *Chemistry Education Research and Practice*, 20, 369-382. Copyright [2019] by Jared Breakall. Adapted with permission.

Procedure

I obtained copies of the detailed item analyses for all CTM examinations administered to second-year (P2) and third-year (P3) pharmacy students during the 2017-2018 academic year. The majority of the examinations included multiple-choice test items, and one exam included two short answer questions. The short answer questions were eliminated as the focus of this study was not on constructed-response items. Item performance (psychometric) parameters were collected from the item analysis reports from the EAP, including the item ID, difficulty index,

discrimination index, and average answer time. The discrimination index was reported for each item using responses from the upper 27% and lower 27% of students, categorized by their performance on the entire examination.

Sampled Test Item Database Construction

To facilitate the data analysis, a database was created to house the psychometric information for each of the sample test items. I exported the item analysis report for each examination in the CTM sequence of courses from the EAP. The item analysis report allowed the selection of the following options:

- The question identification number (ID);
- The question text including the question title and answer choice text;
- The multiple-choice responses for all test takers;
- The multiple-choice responses for students in the upper and lower 27% of the total score range;
- The item difficulty index;
- The item discrimination index;
- A history of item performance from previous examinations; and
- The categories associated with each item.

Categories can be created in Examsoft™ to guide and improve curricular design and to provide students with meaningful feedback on their performance. Test items can be tagged with unlimited categories; for example, accreditation standards, learning outcomes, Bloom's Taxonomy level, or any other measure that will provide meaningful assessment feedback for student learning or curricular improvement. Additionally, based on the available reporting information, the question ID, multiple-choice response, upper/lower 27% of students, difficulty

index, discrimination index, and average item answer time were the most salient to the purpose of the study and were exported into a Microsoft Excel 2016® spreadsheet. Figure 7 displays the option available for the item analysis in the ExamSoft Administrator Portal. The spreadsheet included a column for the eight criteria from the IWFEI for each of the 313 items being evaluated. Figure 8 displays an example of a modified item analysis report from the ExamSoft™ Administrator Portal.

Data Verification

Data verification was conducted to ensure data were imported without errors by visually comparing the entries against the original electronic copies of the item analysis report, which included the item ID. There were four instances where discrepancies were found. I checked the electronic copies of the item analysis using the item ID to correct the error in the database. Given that item writers were instructed to write an item that included a stem and four response options, there was no missing data reported in the data analysis. Therefore, I did not conduct a missing data diagnosis to understand the pattern and randomness of missing data.

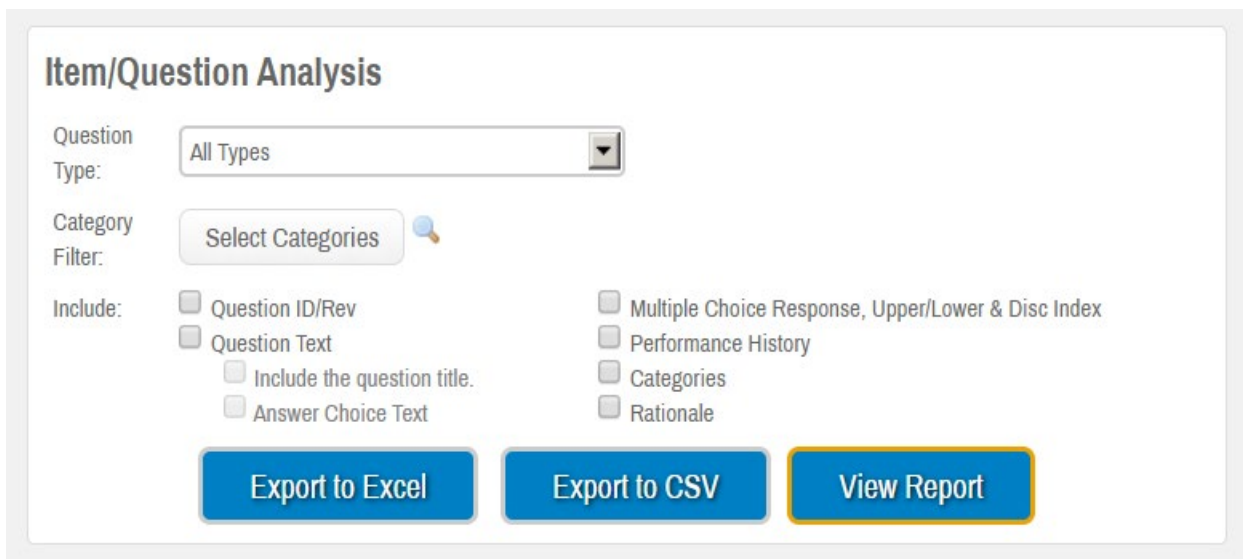


Figure 7. Example of the various options available for an item analysis report from the ExamSoft Administrator Portal.

Question Analysis (Multiple Choice)																					
Exam Takers = 120		KR20 = 0.63		Stdev = 10.19		Mean = 82.25 (82.25%)			Median = 82.82			Min = 54.50		Max = 98.50		Total Pts = 100.00					
Question #	Correct Responses			Disc. Index	Point Biserial	Correct Answer	Response Frequencies (*Indicates correct answer)											Avg Answer Time			
	Diff(p)	Upper	Lower				A	B	C	D	E	F	G	H	I	J	Unanswered				
1	0.75	87.50%	59.38%	0.28	0.26	D	22	4	4	*90	-	-	-	-	-	-	-	0	01:27		
Question ID / Rev: 6893 / 5							% Selected	18.33	3.33	3.33	75.00	-	-	-	-	-	-	-	0.00	-	
							Point Biserial (rpb)	-0.19	-0.05	-0.19	0.26	-	-	-	-	-	-	-	-	-	
							Disc. Index	-0.22	0.00	-0.06	0.28	-	-	-	-	-	-	-	-	-	
							Upper 27%	0.09	0.03	0.00	0.88	-	-	-	-	-	-	-	-	-	
							Lower 27%	0.31	0.03	0.06	0.59	-	-	-	-	-	-	-	-	-	
2	0.85	96.88%	71.88%	0.25	0.25	D	0	12	6	*102	-	-	-	-	-	-	-	0	01:23		
Question ID / Rev: 6894 / 7							% Selected	0.00	10.00	5.00	85.00	-	-	-	-	-	-	-	-	0.00	-
							Point Biserial (rpb)	0.00	-0.21	-0.12	0.25	-	-	-	-	-	-	-	-	-	
							Disc. Index	0.00	-0.16	-0.09	0.25	-	-	-	-	-	-	-	-	-	
							Upper 27%	0.00	0.03	0.00	0.97	-	-	-	-	-	-	-	-	-	
							Lower 27%	0.00	0.19	0.09	0.72	-	-	-	-	-	-	-	-	-	-

Figure 8. An example of a modified item analysis report from the ExamSoft™ Administrator Portal.

Interrater Reliability

Using the IWFEI, I analyzed a sample of 313 items from the CTM sequence of courses. I confirmed the item ID for the items included in the sample, and reviewed the stem and responses for each of the items. I responded to the questions on the form with “yes”, “no”, or “not applicable” for each of the eight criteria included on the IWFEI. A “yes” response suggests that there was adherence to the item writing guideline, a “no” response suggests that the item writing guideline was violated, and “not applicable” suggests that the item guideline did not apply to the item. The ratings were entered into the Microsoft Excel spreadsheet with the item ID. To provide a measure of accuracy and consistency of the item coding, a second rater evaluated a subset ($n = 92$) of the sample items. I used the List Randomizer application from Random.org to randomly select the 92 items that were evaluated by the secondary rater. I selected the secondary rater based on their educational background and professional experience with item-writing, test development, and the EAP. Both myself and the secondary rater possess a Master of Education (MEd) and have a combined 30 years of experience providing faculty development in teaching

and learning in an academic health sciences environment. Prior to evaluating the test item sub-sample, the secondary rater and I completed an orientation using the Item-Writing Flaws Evaluation Instrument Guide, which provided clear guidance, good examples, and poor examples of each of the criteria (See Appendix A). Both raters evaluated a common set of 10 items to ensure consistency in the interpretation and application of the IWFEI guide. Any disagreements were discussed and resolved. Following this process, the second rater independently coded 92 items.

Interrater Reliability Results. Interrater reliability was calculated using the Cohen's kappa (κ) statistic. Cohen's kappa (κ) agreement is an adjusted form of percentage agreement that takes into account chance agreement. The percentage of agreement is the simplest measure of interrater agreement. A common set of 92 items were double coded. Overall, there was very good agreement between the two raters' judgements, $\kappa = .83, p < .0005$. The results of the interrater reliability analysis indicate the proportion of agreement over and above chance agreement. The interrater reliability was further examined to determine agreement according to type of item-writing flaw identified. The results are shown in Table 3.

Results of the interrater reliability analysis showed that there was significant agreement of responses between the two raters in 7 out of the 8 item writing guidelines in the CTM sequence of courses which include (1) "Is the test item clear and succinct?" ($p < 0.001$); (2) "If the item uses negative phrasing such as "not" or "except", is the negative phrase bolded or capitalized?" ($p < 0.001$); (3) "If the answer choices are numerical, are they listed in ascending or descending order?" ($p < 0.001$); (4) "If the answer choices are verbal, are they approximately the same length?" ($p < 0.001$); (5) "Does the item avoid "all of the above" or "none of the above" as a possible answer choice?" ($p < 0.001$); (6) "Does the item avoid grammatical or phrasing cues?"

($p < 0.001$); and (7) “Could the item be answered without looking at the answer choices?” ($p < 0.001$). The significant p -values ($p < .0005$) means that the kappa (κ) coefficient is statistically significantly different from zero, meaning there is agreement. Investigation of the κ coefficient showed that there is almost perfect agreement in the responses in (1) “Is the test item clear and succinct?” ($\kappa = 0.82$); (2) “If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?” ($\kappa = 0.83$); and (3) “If the answer choices are numerical, are they listed in ascending or descending order?” ($\kappa = 0.83$). An investigation of the κ coefficient showed that there is substantial agreement in the responses in (3) “If the answer choices are numerical, are they listed in ascending or descending order?” ($\kappa = 0.68$); (5) “Does the item avoid “all of the above” or “none of the above” as a possible answer choice?” ($\kappa = 0.66$), and (7) “Could the item be answered without looking at the answer choices?” ($\kappa = 0.76$) between the two raters. On the other hand, there is a moderate agreement in the responses in (1) (6) “Does the item avoid grammatical or phrasing cues?” ($\kappa = 0.49$). The lower value means that the two different raters had moderate agreement on their rating of whether the item “avoided grammatical or phrasing cues.” Overall, the interrater reliability data analysis indicates a sufficiently high level of coding agreement providing evidence of the accuracy of the ratings of the full sample of 313 items.

Table 3

Interrater Reliability Statistics of Response on Most Frequently Occurring IWFs and Combinations of Flaws between Two Raters

Variable Code	Most Frequently Occurring IWFs and Combinations of Flaws	Kappa Value	Asymptotic Standard Error	Approximate T	p
Clear	1. Is the test item clear and succinct?	0.82	0.06	7.95	0.00*

Phrasing	2. If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?	0.83	0.12	10.11	0.00*
Numerical	3. If the answer choices are numerical, are they listed in ascending or descending order?	0.68	0.13	7.22	0.00*
Verbal	4. If the answer choices are verbal, are they approximately the same length?	0.83	0.06	10.66	0.00*
Aota_nota	5. Does the item avoid “all of the above” or “none of the above” as a possible answer choice?	0.66	0.23	6.71	0.00*
Cues	6. Does the item avoid grammatical or phrasing cues?	0.49	0.31	5.47	0.00*
Choices	7. Could the item be answered without looking at the answer choices?	0.76	0.69	7.28	0.00*
Ktype	8. Does the item avoid complex K-type item format?	0.00	No statistics are computed because rater 2 is a constant		

Data Analysis

The independent variable is the presence of IWFs on CTM examinations. The three dependent variables are the psychometric indices of the exam, which are the difficulty level, the discrimination index, and the average item answer time. A series of descriptive statistics were conducted to examine RQ1 (What are the most common IWFs in the clinical therapeutic module sequence of courses?) and RQ2 (What percentage of items from locally developed summative examinations for twelve (12) clinical therapeutics module courses contain more one or more item-writing flaws?). Descriptive statistics are most appropriate for measuring the most frequently occurring item-writing flaws, combinations of flaws, and percentage of items with one

or more flaws (McMillan, 2016). Frequencies and percentages for the most common IWFs are discussed in Chapter 4. Specific descriptive statistics were examined for each IWF identified and included the mean, minimum, maximum, and standard deviation. Microsoft Excel 2016 and IBM SPSS 27[®] were used to analyze the data in this study.

I conducted a series of Spearman's rho correlations to examine RQ3 (What is the relationship between the most common item writing flaws in the clinical therapeutics module examinations and the psychometric indices of items, including item difficulty, item discrimination, and average answer time?). The independent variables, item-writing flaws, are measured on an ordinal scale with 0 = N/A, 1 = Yes, 2 = No, which reflect an increasing degree of IWFs. The dependent variables, item difficulty, item discrimination, and average item answer time, are measured on a continuous scale with item difficulty having a *p*-value range of 0.00 to 1.00 and a discrimination index value ranging from -1.0 to 1.00. I conducted three Spearman's rho correlations to examine the strength and direction of the relationship between most common item-writing flaws and each of the psychometric indices of CTM MCQs. A Spearman's rho correlation was conducted to examine the relationship between the presence of IWFs and the difficulty index. A Spearman's rho correlation was conducted to examine the relationship between IWFs and the discrimination index. A final Spearman's rho correlation analysis was conducted to examine the association between the presence of IWFs and average item answer time.

CHAPTER 4: FINDINGS

This chapter presents the findings specific to each research question. The purpose of this quantitative study was to examine the frequency and nature of IWFs on locally developed, high-stakes summative examinations for 12 CTM sequence of courses at a pharmacy school located at a research-intensive academic medical center in the southeastern region of the United States using the Item-Writing Flaws Evaluation Instrument (IWFEI). The study examined the relationship between IWFs and the psychometric properties of 313 randomly selected test items, including item difficulty, item discrimination, and average item answer time. The following research questions guided the study implementation:

RQ1. What are the most common item-writing flaws in the clinical therapeutics module sequence of courses?

RQ2. What percentage of items from locally developed summative examinations for 12 clinical therapeutics module courses contain one or more item-writing flaws?

RQ3. What is the relationship between the most common item-writing flaws in the clinical therapeutics module examinations and the psychometric indices of items, including item difficulty, item discrimination, and average item answer time?

The chapter is organized into three sections. The first section presents the descriptive statistics summaries of psychometric indices of items to address RQ1. The second section presents the summaries of the responses on the most common IWFs in the CTM sequence of courses to address RQ2. The third section presents results of the Spearman's correlation analysis to determine the relationship between the most common IWFs in the CTM examinations and the psychometric indices of items, including item difficulty, item discrimination, and average item answer time to address RQ3.

Descriptive Statistics Summaries of Psychometric Indices of Items

Descriptive statistics summaries of the three psychometric indices for the 313 items from the 2017-18 CTM sequences of courses using the IWFEI were calculated. The purpose of the descriptive statistics is to describe the basic features of data or the summary statistics of the psychometric indices for the 313 items from the 2017-18 CTM examination based on the responses of the representative population of CTM items. The psychometric indices include item difficulty, item discrimination, and average item answer time. The descriptive statistics summaries of the psychometric indices are shown in Table 4. For item difficulty, the mean score was .83 ($SD = 0.15$). The highest item difficulty index among the 313 sample items was 1.00 and the lowest was .18. For item discrimination, the mean index score of item discrimination was .20 ($SD = 0.16$). The highest item discrimination among the 313 sample items was .94 and the lowest was -.04. For average item answer time, the mean average item answer time was 1.23 minutes ($SD = 0.54$ minutes). The longest average item answer time among the 313 sample items was 4.17 minutes and the shortest was 0 minutes.

Table 4

Descriptive Statistics Summaries of Psychometric Indices

Study variable	<i>N</i>	Minimum	Maximum	<i>M</i>	<i>SD</i>
Item difficulty	313	0.18	1.00	0.83	0.15
Item discrimination	313	-0.04	0.94	0.20	0.16
Average item answer time (minutes)	313	0.00	4.17	1.23	0.54

Analysis for RQ1: Summaries of Survey Responses on Most Common Item-Writing Flaws in the Clinical Therapeutics Module Sequence of Courses

I conducted descriptive statistics analysis to address RQ1 to determine the most common IWFs in the CTM sequence of courses and also the percentage of items from locally developed

summative examinations for 12 CTM courses that contain one or more IWFs. It should be noted that RQ1 and RQ2 were addressed using a similar analysis by summarizing the responses on the sample items on the IWFEI. A “yes” response suggests that there was adherence to the item-writing guideline, a “no” response suggests that the item-writing guideline was violated, and “not applicable” suggests that the item-writing guideline did not apply to the item. The responses were transformed into different variables, with “yes” recoded to “1,” “no” recorded to “2,” and “not applicable” recoded to “0.” Specifically, frequency and percentage summaries of the responses on the IWFEI were obtained. The summaries of the responses on the IWFEI are shown in Table 5.

Table 5

Frequency and Percentage Summaries of Response on Most Frequently Occurring IWFs

IWFEI Criteria	<i>n</i>	%
1. Is the test item clear and succinct?		
Yes	248	79.2
No	65	20.8
2. If the item uses negative phrasing such as “not” or “except,” is the negative phrase bolded or capitalized?		
NA	303	96.8
Yes	6	1.9
No	4	1.3
3. If the answer choices are numerical, are they listed in ascending or descending order?		
NA	279	89.1
Yes	29	9.3
No	5	1.6
4. If the answer choices are verbal, are they approximately the same length?		
NA	21	6.7
Yes	276	88.2
No	16	5.1
5. Does the item avoid “all of the above” or “none of the above” as a possible answer choice?		
Yes	310	99.0
No	3	1.0
6. Does the item avoid grammatical or phrasing cues?		
Yes	309	98.7
No	4	1.3
7. Could the item be answered without looking at the answer choices?		
Yes	226	72.2
No	87	27.8
8. Does the item avoid complex K-type item format?		
Yes	304	97.1
No	9	2.9

Most Prevalent Flaws

In terms of violating the item-writing guidelines as defined by the eight criteria on the IWFEI, the three criteria with the highest frequencies of “no” responses, which indicated that the item-writing guideline was violated, were: (a) “Could the item be answered without looking at the answer choices?” (87; 27.8%); (b) “Is the test item clear and succinct?” (65; 20.8%); and (c) “If the answer choices are verbal, are they approximately the same length?” (16; 5.1%). The findings suggest that the most common IWFs in the CTM sequence of courses include: (a) item could not be answered without looking at the answer choices, (b) test item was not clear and succinct, and (c) the answer choices were not approximately the same length if the answer choices are verbal.

In terms of adherence to the item-writing guidelines using the eight criteria included in the IWFEI, the majority of the 313 sample items evaluated adhered to six out of the eight item-writing criteria. The three criteria with the highest frequencies of “yes” responses, which indicate adherence to the item-writing guideline were: (a) “Does the item avoid “all of the above” or “none of the above” as possible answer choices (310, 99.0%); (b) “Does the item avoid grammatical or phrasing cues?” (309, 98.7%); and (c) “Does the item avoid complex K-type item format?” (304, 97.1%). Overall, the findings suggest that clinical therapeutics module coordinators and the item writers may have utilized the guidance from the Clinical Therapeutics Module Coordinator Guide or other item-writing resources to write the test items.

Analysis for RQ2: Percentage of Items from Locally Developed Summative Examinations for 12 Clinical Therapeutics Module Courses Containing One or More Item-Writing Flaws

The descriptive analyses indicated that 116 of the 313 items (37%) violated item-writing guidelines according to the eight of the criteria from the IWFEI. Fifty of the items (16%)

violated at least one item-writing guideline. Fifty-three items (17%) violated two item-writing guidelines. Thirteen items (4%) violated three item-writing guidelines. For those items that violated two guidelines, roughly half (24 out of 53 items) shared problems associated with (a) “Is the item clear and succinct?” and (b) “Could the item be answered without looking at the answer choices?” criteria. Of the 13 items that violated three item-writing guidelines, four items included flaws for (a) “Is the item clear and succinct?” (b) “If the answer choices are verbal, are they approximately the same length?” and (c) “Could the item be answered without looking at the answer choices?”

Analysis for RQ3: Results of the Correlation Analyses Between Most Common Item-Writing Flaws in the Clinical Therapeutics Module Examinations and Psychometric Indices

A Spearman’s rank order correlation was used to address RQ3 and to examine the strength and direction of the relationship among the most common IWFs in the CTM items and the specific psychometric properties: item difficulty, item discrimination, and average item answer time. The Spearman’s correlation analysis is a nonparametric correlation analysis that measures the strength and direction of relationship between two variables measured on an ordinal scale. A Spearman’s correlation was used because of the ordinal nature of item rankings (0 = *N/A*, 1 = *Yes*, 2 = *No*), which reflect an increasing degree of IWFs. Given that the variables used in this study were transformed into ordinal variables, the Spearman’s correlation was the most appropriate statistical test for the data analysis. The correlation coefficient provides a measure of the strength and direction of the relationship among stated variables. Correlation coefficient values range from +1 to -1, which indicate a positive or negative association of the variables. A correlation coefficient value closer to zero indicates a weak association and a value closer to one

indicates a strong association. In the correlation test, I used a two-tailed test and .05 level of significance. Significant correlation between variables exists when the p -value of the r statistic for the correlation test is less than or equal to the level of significance set at .05. Table 6 summarizes the results of the Spearman's correlation analysis to address RQ3.

Table 6

Results of Spearman's Correlation Analysis of Relationship Between Most Common IWFs in the CTM Examinations and Psychometric Indices

Criteria	Spearman's rho	Item difficulty	Item discrimination	Average item answer time (minutes)
1. Is the test item clear and succinct?	Correlation coefficient	.05	-.03	-.02
	Sig. (2-tailed)	.37	.60	.74
	N	313	313	313
2. If the item uses negative phrasing such as "not" or "except", is the negative phrase bolded or capitalized?	Correlation coefficient	.02	-.01	-.04
	Sig. (2-tailed)	.74	.85	.47
	N	313	313	313
3. If the answer choices are numerical, are they listed in ascending or descending order?	Correlation coefficient	.01	-.07	.15*
	Sig. (2-tailed)	.86	.23	.01
	N	313	313	313
4. If the answer choices are verbal, are they approximately the same length?	Correlation coefficient	.03	.02	-.05
	Sig. (2-tailed)	.56	.68	.41
	N	313	313	313
5. Does the item avoid "all of the above" or "none of the above" as a possible answer choice?	Correlation coefficient	.02	.00	-.01
	Sig. (2-tailed)	.75	.95	.91
	N	313	313	313
6. Does the item avoid grammatical or phrasing cues?	Correlation coefficient	.08	-.08	-.05
	Sig. (2-tailed)	.16	.14	.35
	N	313	313	313
7. Could the item be answered without looking at the answer choices?	Correlation coefficient	-.02	.07	.02
	Sig. (2-tailed)	.72	.21	.77
	N	313	313	313

8. Does the item avoid complex K-type item format?	Correlation coefficient	-0.16*	.10	.12*
	Sig. (2-tailed)	.00	.08	.03
	<i>N</i>	313	313	313

*Correlation is significant at the .05 level (2-tailed).

Item Difficulty

Item difficulty represents a percentage of students who answered the test item correctly. The difficulty index is the first indicator of how the question performed. It is a relevant factor in helping faculty determine whether students have learned the concepts being tested. I conducted a Spearman’s correlation to assess the relationship between the IWFs identified in the study and item difficulty. As described previously, the most common IWFs identified were: (a) “Is the item clear and succinct?” (b) “If the answer choices are verbal, are they approximately the same length?” and (c) “Could the item be answered without looking at the answer choices?” There was no statistically significant correlation between “Is the item clear and succinct?” and item difficulty, $r = .051, p = .371$. There was no statistically significant correlation between “If the answer choices are verbal, are they approximately the same length?” and item difficulty, $r = .033, p = .561$. There was no statistically significant correlation between “Could the item be answered without looking at the answer choices?” and item difficulty, $r = -.020, p = .721$. Although the results of the Spearman’s correlation analysis indicate a weak, non-statistically significant association, other studies suggest that item-writing flaws can affect item difficulty (Pais et al., 2016; Tarrant & Ware, 2008). Test items not written in adherence with best practices are often not clear and succinct and may not be answered without looking at the answer choices, which may increase or decrease the difficulty of an item (Rush, Rankin, and White, 2016.)

The results of the Spearman’s correlation analysis indicated a weak negative association between “Does the item avoid complex K-type item format?” and item difficulty. The weak

negative correlation means there is a lower index of item difficulty if there is a “no” response (coded as 2) in the question item “Does the item avoid complex K-type item format?” This means if there is violation in the item-writing guideline for avoiding the complex K-type item format, there is a lower level of item difficulty in answering the question item.

Item Discrimination

The discrimination index range value is from 0.0 -1.0. The target value for the discrimination index of an item should be above .20 for examination items, with the exception of items that were intended to be easy or difficult (Rush et al., 2016; Shete et al., 2015). Items with a discrimination index less than .20 or a negative discrimination index are unacceptable and should be discarded or revised. I conducted a Spearman’s correlation to assess the relationship between the most common IWFs identified in the study and item discrimination. There was no statistically significant correlation between “Is the item clear and succinct?” and item discrimination, $r = -.030, p = .603$. There was no statistically significant correlation between “If the answer choices are verbal, are they approximately the same length?” and item discrimination, $r = .024, p = .678$. There was no statistically significant correlation between “Could the item be answered without looking at the answer choices?” and item discrimination, $r = .071, p = .210$. Although the results of the Spearman’s correlation analysis indicate a non-statistically significant association, other studies suggest that item-writing flaws can affect item discrimination and impact clarity and consistency of exam items (Pais et al., 2016; Rush, Rankin, and White, 2016; Tarrant & Ware, 2008). To improve clarity of test items, faculty should follow best practices for item-writing or item-writing guidelines.

Average Item Answer Time

For the psychometric index of average item answer time, results of the Spearman's correlation analysis showed the response to "If the answer choices are numerical, are they listed in ascending or descending order?" was positively correlated with the psychometric index of average item answer time, $r = .15$, $p = .010$. The positive correlation means there is a longer average item answer time if there is a "no" response (coded as 2) to the question, "If the answer choices are numerical, are they listed in ascending or descending order?" This means if there is a violation in the item-writing guideline of listing numerical answer choices into ascending or descending order, there is a longer average item answer time in responding to the question item. Also, results of the Spearman's correlation analysis showed the response to "Does the item avoid complex K-type item format?" was positively correlated with the psychometric index of average item answer time, $r = .120$, $p = .034$). The significant positive correlation means there is a longer average item answer time if there is a "no" response (coded as 2) to the question, "Does the item avoid complex K-type item format?" This means that if there is a violation in the item-writing guideline of avoiding the complex K-type item format, there is a longer average item answer time in responding to the question item. In other words, test items that include prompts such as "which of the following is not correct" require significantly more response time than items that do not use this type of language.

In addition, I conducted a series of point-biserial correlations to further the results of the Spearman's rho correlations while treating the independent variable slightly differently to reflect adherence and non-adherence rather than as a continuum of adherence. The point-biserial correlation is a special case of the Pearson correlation in which I assessed the correlation between one dichotomous variable and one continuous variable. Adherence to the eight criteria

on the IWFEI was coded as “yes” and nonadherence to the criteria was coded as “no.” The results of the point-biserial correlation analyses were similar to the results of the Spearman’s correlations (see Table 7).

Item Difficulty

A series of point-biserial correlations were run to determine the relationship between the most common IWFs and item difficulty. There was no statistically significant correlation between “Is the item clear and succinct?” and item difficulty, $r_{pb} = .450, p = .431$. There was no statistically significant correlation between “If the answer choices are verbal, are they approximately the same length?” and item difficulty, $r_{pb} = .019, p = .743$. There was no statistically significant correlation between “Could the item be answered without looking at the answer choices?” and item difficulty, $r_{pb} = -.002, p = .966$. The results of the point-biserial correlation analysis indicated a weak negative association between “Does the item avoid complex K-type item format?” and item difficulty, $r_{pb} = -.224, p = .000$. The weak negative correlation confirms there is a lower index of item difficulty if there is a “no” response (coded as No) in the question item “Does the item avoid complex K-type item format?”

Item Discrimination

The results of the point-biserial correlation confirm the findings of the point-biserial correlation. There was no statistically significant correlation between “Is the item clear and succinct?” and item discrimination, $r_{pb} = -.027, p = .637$. There was no statistically significant correlation between “If the answer choices are verbal, are they approximately the same length?” and item discrimination, $r_{pb} = .011, p = .853$. There was no statistically significant correlation between “Could the item be answered without looking at the answer choices?” and item discrimination, $r_{pb} = .071, p = .213$.

Average Item Answer Time

There was no statistically significant correlation between “Is the item clear and succinct?” and item difficulty, $r_{pb} = -.047, p = .412$. There was no statistically significant correlation between “If the answer choices are verbal, are they approximately the same length?” and item difficulty, $r_{pb} = .002, p = .979$. There was no statistically significant correlation between “Could the item be answered without looking at the answer choices?” and item difficulty, $r_{pb} = -.038, p = .501$.

Table 7

Results of Point-Biserial Correlation Analysis of Relationship Between Most Common IWFs in the CTM Examinations and Psychometric Indices

Criteria	Point-Biserial Correlation	Item difficulty	Item discrimination	Average item answer time (minutes)
1. Is the test item clear and succinct?	Correlation coefficient	.05	-.03	-.05
	Sig. (2-tailed)	.43	.64	.41
	<i>N</i>	313	313	313
2. If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?	Correlation coefficient	-.03	-.00	-.04
	Sig. (2-tailed)	.60	.98	.47
	<i>N</i>	313	313	313
3. If the answer choices are numerical, are they listed in ascending or descending order?	Correlation coefficient	-.02	-.01	.03
	Sig. (2-tailed)	.71	.82	.64
	<i>N</i>	313	313	313
4. If the answer choices are verbal, are they approximately the same length?	Correlation coefficient	.02	.01	.01
	Sig. (2-tailed)	.74	.85	.98
	<i>N</i>	313	313	313
5. Does the item avoid “all of the above” or “none of the above” as a possible answer choice?	Correlation coefficient	.03	-.01	-.30
	Sig. (2-tailed)	.58	.97	.64
	<i>N</i>	313	313	313
	Correlation coefficient	.06	-.08	-.03

6. Does the item avoid grammatical or phrasing cues?	Sig. (2-tailed) <i>N</i>	.27 313	.18 313	.63 313
7. Could the item be answered without looking at the answer choices?	Correlation coefficient Sig. (2-tailed) <i>N</i>	-.01 .966 313	.07 .21 313	-.04 .50 313
8. Does the item avoid complex K-type item format?	Correlation coefficient Sig. (2-tailed) <i>N</i>	-.22** .00 313	.10 .08 313	.06 .30 313

**Correlation is significant at the 0.01 level (2-tailed).

Summary

The purpose of this quantitative study was to examine the frequency and nature of IWFs on locally developed, high-stakes summative examinations for 12 CTM sequence of courses at a pharmacy school located at a research-intensive academic medical center in the southeastern region of the United States using the eight criteria from the IWFEI. Using a descriptive, correlational nonexperimental design, the results from the data analysis indicated the majority of items evaluated for this study adhered to the eight item-writing criteria outlined in the IWFEI.

As stated, I conducted descriptive statistics analysis and Spearman's correlation analysis to address the research questions of this study. For RQ1, results of the descriptive statistics analysis showed that the most common IWFs in the CTM sequence of courses include: (a) the test item could not be answered without looking at the answer choices, (b) the test item was not clear and succinct, and (c) the answer choices were not approximately the same length if the answer choices are verbal. For RQ2, results of the descriptive statistics analysis showed for 116 of the 313 (37%) items evaluated, 50 of the items (16%) violated at least one item-writing guideline, 53 items (17%) violated two item-writing guidelines, and 13 items (4%) violated three item-writing guidelines.

For RQ3, results of the Spearman's correlation analysis showed that the response on question item "Does the item avoid complex K-type item format?" was significantly negatively correlated with the psychometric index of item difficulty. Results of the Spearman's correlation analysis also showed that the responses on question items "If the answer choices are numerical, are they listed in ascending or descending order?" and "Does the item avoid complex K-type item format?" was significantly positively correlated with the psychometric index of average item answer time. The results of the point-biserial correlation confirm the results of the Spearman's rho correlations.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

Introduction

The purpose of this study was to examine the frequency in which IWFs occur in locally developed high-stakes summative assessments for the CTM sequence of courses at a school of pharmacy using eight criteria included on the IWFEI (Breakall et al., 2019). The primary goals of the study were to: (a) identify the most common IWFs on examinations in the CTM sequence of courses; (b) determine the percentage of items included on the CTM examinations with one or more IWFs; and (c) examine the relationship between the most frequently occurring IWFs and test item psychometric indices including item difficulty, item discrimination, and average item answer time. This chapter summarizes the research findings as they relate to the overall objectives of the study, discuss the implications for pharmacy education, and offers recommendations for further research.

Discussion of the Findings

Pharmacy faculty have a responsibility to ensure that assessments are valid and reliable measures of students' learning (Hicks, 2014). Yet, there is significant evidence that IWFs are common among multiple-choice exams (Breakall et al., 2019; Nedeau-Cayo et al., 2013; Rush et al., 2016; Stagnaro & Downing, 2006; Tarrant et al., 2006), a common method of student assessment in pharmacy education. These IWFs can adversely affect student test performance (Board & Whitney, 1972; Breakall et al., 2019; DiBattista & Kurwaza, 2011; Downing, 2005; Pham, Besanko, & Devitt, 2018; Reichert, 2011). In addition, Pate and Caldwell (2014) noted, due to a lack of formal training in curricular assessment in pharmacy education, many pharmacy faculty are not familiar with item-writing best practices, which increases the risk of exam

questions having IWFs. Therefore, there is a need to evaluate the quality of exam items given to pharmacy students.

The current study addressed this gap in the literature by examining the IWFs in high-stakes summative assessments for the CTM sequence of courses at a school of pharmacy. Findings from the current study suggest IWFs are frequent in the exams, with over one third (37%) of items having at least IWF. The most common IWFs were: (a) the test item could not be answered without looking at the answer choices; (b) the test item was not clear and succinct; and (3) the answer choices were not approximately the same length if the answer choices were verbal rather than numerical. It is important to ensure the central idea being assessed is included in the stem of a test item. Additionally, the goal of writing any multiple-choice item stem is to be clear, succinct, and focused (McDonald, 2018). Test items that are clear, succinct, and focused eliminate extraneous information to make the item more direct and easier to understand, thereby decreasing the difficulty and increasing the reliability of the item (McDonald, 2018). The current study also found these IWFs were associated with poor psychometric indices.

The study's findings are consistent with previous literature, which indicates that IWFs are common (Tarrant et al., 2006) and may negatively affect exam psychometrics (DiBattista & Kurwaza, 2011; Pais et al., 2016; Tarrant & Ware, 2006). Considering both the existing literature and the current study, it appears colleges and schools of pharmacy should consider carefully evaluating exam items given to pharmacy students to determine the extent of IWFs and ensure the interpretation of scores from exams are valid and reliable.

RQ1. What are the most common item-writing flaws in the clinical therapeutics module sequence of courses at a school of pharmacy at a research-intensive academic health center in the southeastern United States?

In this study, findings indicated that three IWFs were the most common: (a) the test item could not be answered without looking at the answer choices ($n = 87, 27.8\%$); (b) the test item was not clear and succinct ($n = 65, 20.8\%$); and (c) the answer choices were not approximately the same length if the answer choices were verbal ($n = 16, 5.1\%$). The findings of the current study were somewhat consistent with studies conducted on the adherence to item-writing guidelines in other disciplines (Breakall, Randalls, and Tasker, 2019; Nadeau-Cayo, Laughlin, Rus, and Hall, 2013; Tarrant, Knierim, Hayes, and Ware, 2006). Ideally, students should be able to answer an item without looking at the response options (Towns, 2014). Test items that are clear, succinct, and focused eliminate extraneous information and make the item more direct and easier to understand, thereby decreasing the difficulty and increasing the reliability of the item (McDonald, 2018).

The three most common IWFs identified in the current study are often identified as common IWFs in other disciplines; however, the frequency of each type of IWF seems to vary depending on which kind of exam is under review. In a review of 43 general chemistry exams, Breakall et al. (2019) found that the most common IWF (29.3%; $n = 299$) was that the item could not be answered without looking at the answer choices. Breakall et al.'s findings are consistent with the current study in terms of which IWF was most common and the frequency at which the IWF was found. However, when examining exams from other fields, researchers found alternative IWFs were more common. Both Rush et al. (2016) and Stagnaro and Downing (2006) found the most common IWF was items that included awkward stem structure. Rush et al. (2016) found this IWF in 29% of the items on a veterinary exam, and Stagnaro and Downing (2006) found this IWF in 100% of the items among continuing medical education exams.

Though specific item-writing violations may vary across exams, there is overwhelming evidence that the IWFs identified in the present study are common. The current study adds to this growing body of literature by addressing IWFs in Doctor of Pharmacy programs. Previous research had yet to examine IWFs in pharmacy education; however, to meet the evaluation standards outlined by ACPE, the quality of multiple-choice exams given to pharmacy students needed to be examined. The current study takes the first step in understanding the quality of assessments for pharmacy students and raises awareness about the use of best practices in test construction to ensure the interpretation of scores from CTM examinations are valid and reliable.

RQ2. What percentage of items from locally developed summative examinations for 12 clinical therapeutics module sequence of courses contain one or more item-writing flaws?

An analysis of the data to determine what percentage of items from locally developed summative examinations for the 12 CTM courses contained one or more IWFs indicated that 62.9% of the test items ($n = 197$) were identified as free from IWFs. Approximately, 37% ($n = 116$) violated item-writing guidelines evaluated in this study according to eight of the criteria from the IWFEI. Of these, approximately 16% of the test items ($n = 50$) violated at least one item-writing guideline, 17% ($n = 53$) violated two, and 4% ($n = 13$) violated three item-writing guidelines. For those items that violated two criteria, the most common combination of flaws included “Is the test item clear and succinct?” and “Could the item be answered without looking at the answer choices?” For the items that violated three criteria, the most common combination of flaws included (a) “Is the test item clear and succinct?” (b) “If the answer choices are verbal, are they approximately the same length?” and (c) “Could the item be answered without looking at the answer choices?”

The findings of this study support other research on the prevalence of IWFs on multiple-choice examinations. Across a variety of disciplines, the rates of IWFs are high, with many studies indicating that approximately half of sampled questions had at least one IWF (Tarrant, Knierim, Hayes, and Ware, 2006). The results of the present study were consistent with those of Rush et al. (2016), who found 28.8% of test items ($n = 554$) were identified as free of IWFs, 33.9% of the test items ($n = 653$) included one IWF, and 37.3% ($n = 718$) were identified as having more than one IWF. Similarly, Tarrant et al. (2006) examined 2,770 test items from examinations that were administered over a five-year period from 2001 to 2005 and identified 46.2% of test items ($n = 1,280$) contained at least one IWF, 10.5% of test items ($n = 290$) contained two IWFs, and 1.4% ($n = 40$) contained three IWFs.

The results of the current study are in line with similar studies of multiple-choice examinations. The presence of IWFs on CTM examinations should concern faculty and administrators in colleges and schools of pharmacy. Though a majority of the questions do not have an IWF, over one third of the items did. A significant number of items in the exam have at least one IWF. As there is evidence these IWFs adversely affect student test performance (Board & Whitney, 1972; Breakall et al., 2019; DiBattista & Kurwaza, 2011; Downing, 2005; Pham et al., 2018; Reichert, 2011), it is possible that these flaws are hindering the reliability and validity of the exams given to pharmacy students. Colleges and schools of pharmacy may consider conducting thorough reviews of the assessments given to pharmacy students to ensure that the quality of the exams given are not adversely affecting the assessment outcomes. Additionally, the findings of the current study suggest, due to the high-stakes nature of CTM examinations in pharmacy education, pharmacy faculty should develop protocols to ensure examinations meet the professional testing standards outlined in the *Standards for Educational and Psychological*

Testing (American Psychological Association, 2014). These standards specifically address rules for test specifications, item development and review, procedures for administration and scoring, and test revisions.

RQ3. What is the relationship between the most common item-writing flaws in the clinical therapeutics module sequence of courses and the psychometric indices of items, including item difficulty, item discrimination, and average answer time?

In the final analysis of the current study, the psychometric indices (i.e., item difficulty, discrimination index, and average item answer time) of the CTM were examined to explore the connections between IWFs and item functioning. Psychometric indices provide important information on the validity and reliability of exam questions (Chiavaroli & Familiari, 2011). The difficulty index indicates how easy or difficult an item was, the discrimination index can identify items that may have been miskeyed, and the average item answer time may provide evidence about the difficulty of an item. It is considered to be a best practice to review these indices and remove or edit items that do not meet criteria per those indices (Kheyami et al., 2018; Quaigrain & Arhin, 2017; Rush et al., 2016; Shete et al., 2015). The results of the current study suggest that the average psychometric values of the CTM exam items are consistent with recommendations for acceptable difficulty levels and discrimination index values (Quaigrain & Arhin, 2017). Both the difficulty and discrimination indices for the current study were within recommended limits, suggesting that, overall, the psychometrics of the exam items are acceptable. Findings from the current student reveal opportunities to improve the psychometric values of the many multiple-choice tests. However, detailed analysis showed there may be questions from the exams that need to be reviewed or removed as they are outside of the recommended guidelines for the difficulty index and the discrimination index.

Previous researchers (DiBattista & Kurwaza, 2011; Pais et al., 2016; Tarrant & Ware, 2006) suggested multiple-choice questions with IWFs tend to be more difficult and have limited ability to discriminate among test takers. Although the current study found two specific IWFs were associated with poor psychometric indices or reduced item functioning, the results were not as consistent as previous research in this area. This may be due, at least in part, to methodological differences between the current study and previous research in this area (DiBattista & Kurwaza, 2011; Tarrant & Ware, 2006). The specific findings are reviewed in the context of the previous literature in the following sections.

Item difficulty. Item difficulty represents the percentage of students who answered a test item correctly. The difficulty index is the first indicator of how the question performed. Item difficulty is a relevant factor in helping faculty determine whether students have learned the concepts being tested. Item difficulty is a significant factor in the ability of an item to discriminate between students who know the content being tested and those students who do not (Kheyami et al., 2018). Study results indicated that the average item difficulty was .83 ($SD = .15$). The desirable range for acceptable item difficulty on norm-referenced assessments includes p -values of .30 to .70 (Oermann and Gaberson, 2021). Oermann and Gaberson (2021) suggest that when seeking a desirable range for acceptable item difficulty on criterion-referenced assessments, the difficulty level of test items should be compared between groups of students who met the criterion and those who did not. Additionally, MCQ can be interpreted within the context of the purpose of the exam and the learning objectives the exam assesses (Towns, 2014). For example, difficulty index values of 1.00 may be acceptable if the item is meant to measure mastery. A difficulty index value of 1.00 may not be appropriate for items meant for knowledge discrimination (Examsoft, 2019). Towns suggests that a rule of thumb for interpreting difficulty index values is that above

.75 if easy, between .25 and .75 is average, and below .25 is difficult. In this current study, the highest item difficulty among the 313 items was 1.00 and the lowest was .18. The findings suggest that there are items in the sample that need to be reviewed for difficulty as they are outside of the suggested p -value range (Quaigrain & Ahrin, 2017). Results of the current study suggest that there are both questions that are too easy and too difficult on the CTM examinations. Pharmacy faculty may consider the purpose of the test and review all questions for difficulty levels to ensure that each item falls in the acceptable range and appropriately measures students' knowledge and skills.

Discrimination index. The discrimination index indicates the relationship between success on an item and success on a test, and it is an indicator of test-item quality. Discrimination index refers to the ability of an item to differentiate between the high achieving students and low achieving students (De Champlain, 2010). The mean score of item discrimination was .20 ($SD = .16$). Among the 313 items, the highest discrimination was .94 and the lowest was -.04. This suggests that the overall discrimination index of the exams in the CTMs is consistent with published recommendations, suggesting that a discrimination value between 0.20 and 0.39 was acceptable; however, the item could be improved (McDonald, 2018). Chiavaroli & Familiari (2011) state that different thresholds for acceptable discrimination indices have been suggested (for example, 0.3 or 0.4 have also been suggested by Abdel-Hameed *et al.*, 2005, and McAlpine, 2002, respectively); however, consideration should be given to other psychometric attributes of the exam, such as the number of questions, score distribution, general level of difficulty, and overall homogeneity. There are several questions on the CTM exams that have poor discrimination. These exam questions should be reviewed and revised, or discarded as their respective discrimination index measure meets the criteria for a flawed or miskeyed item. A

negative discrimination value is unacceptable and indicates that the item should be revised or removed.

Average item answer time. Average item answer time is the mean amount of time students used to respond to test items, represented in minutes and seconds (McBrien, 2018). The average item answer time may be an indicator of how easy or difficult a test item was. The average item answer time should be used in conjunction with the item difficulty and item discrimination to make decisions about post-exam modifications.

The mean score of average item answer time is 1.23 minutes ($SD = 0.54$). Among the 313 items, the highest answer time was 4.17 minutes and the lowest was reported as 0.00 minutes. Considering this finding in the context of the item difficulty scores, the range of average answer times suggests that some items need to be reviewed or removed from the exam. However, both extremes may suggest an issue with exam items. For example, a significant number of questions with high answer times included complex case-based questions that required students to read two or more paragraphs to answer the question. McDonald (2018) suggests that while it is important to include all of the information that is needed to answer the complex case-based question in the stem, it is important to communicate the problem as efficiently and clearly as possible. Test items with unnecessary content in the stem extend the reading time and lengthen the time it takes to answer the item. Answering the test items should not be an assessment of reading comprehension for the student.

Item-writing flaws and psychometric indices. Item-writing flaws were found to be associated with poorer psychometric indices among exam questions in the CTM exams. The

current study found the use of complex K-type item format¹ was significantly associated with lower item difficulty. Additionally, there was a longer average response time for items that did not have numerical items listed in ascending or descending order. These findings provide some evidence that IWFs in the CTM are negatively affecting the reliability and validity of the exam items. The findings are consistent with previous literature that showed IWFs are associated with poor psychometric indices (DiBattista & Kurwaza, 2011; Tarrant & Ware, 2006). Items with an IWF have difficulty indices that are higher than the recommended values and have poor discrimination indices (DiBattista & Kurwaza, 2011; Tarrant & Ware, 2006).

Yet, the majority of IWFs examined in the current study were not associated with psychometric indices. Previous studies tend to find more consistent associations between IWFs that are associated with poorer item discrimination (Downing, 2005; DiBattista & Kurwaza, 2011; Tarrant & Ware, 2006). This may be due to several methodological differences. First, the current study correlated the presence of each type of IWF with each psychometric index. Previous research has correlated the presence or absence of at least one IWF with the psychometric indices (DiBattista & Kurwaza, 2011; Tarrant & Ware, 2006). It is possible that had the current study followed a similar method, similar findings may have been found. However, the current study contributes to the literature by examining the specific flaws that are associated with psychometric indices. This information can assist in making more concise recommendations for improving exam quality.

Second, previous studies examined undergraduate level exams not graduate level exams (DiBattista & Kurwaza, 2011; Tarrant & Ware, 2006). However, Pais et al. (2016) found the

¹ K-type format is a question that includes one answer that combines other answers (e.g., all of the above, both A and B).

associations between IWFs and psychometric indices was less consistent in clinical anatomy assessment, which is an advanced course. It is possible that graduate level students have more test-wisness compared to undergraduate students. Test-wisness is the ability to guess the correct answer to an exam question without knowing the content (Downing, 2002). There are many possibilities that make it likely for graduate students to have greater test-wisness compared to undergraduate students. For example, graduate-level students have taken more multiple-choice tests compared to undergraduate students, increasing their skills at answering questions they do not know the answer to or that have multiple flaws. Therefore, IWFs have less of an effect on item discrimination among graduate students because they are able to guess answers correctly and with greater ease when there are IWFs.

However, there is still some evidence in the current study that IWFs are associated with poorer psychometric indices. Therefore, colleges and schools of pharmacy should not ignore the impact of IWFs on the reliability and validity of exam questions based on the results of the demonstrated correlations between IWFs and psychometric indices. The analysis of the data in the current study shows there are exam questions that need to be reviewed, revised, or removed due to either multiple IWFs or poor psychometric indices. Even a few IWFs and exam items with poor psychometric indices can impact the validity and reliability of the exam (Breakall et al., 2019).

As a whole, the findings of the current study suggest that the reliability and validity of the exam items reviewed in the current study are acceptable; however, items could be improved. The average psychometric indices for the items are within recommended limits, but outliers raise concerns about individual questions. As some IWFs are associated with poorer psychometric indices, faculty should be concerned about the impact these IWFs have on the validity and

reliability of individual questions. It is possible that removing these items or revising the question to remove the IWFs could improve the reliability and validity of the exams. Towns (2014) suggests that using item-writing guidelines based on research allows faculty to strengthen test items that makes the inferences drawn from the scores more reliable and valid.

The presence of item-writing flaws on pharmacy CTM examinations may contribute to the inaccurate measurement of pharmacy student knowledge (Tarrant, Knierim, Hayes, & Ware, 2006, Rudolph et al., 2019) and threatens the validity of the inferences made, or conclusions drawn, on the basis of the examination scores. Thus, any decisions about the academic progression of pharmacy students made by administrators based on these examination scores are questionable, when item-writing flaws exist. Factors that interfere with the meaningful interpretation of assessment data are a threat to validity (Downing & Haladyna, 2004). Specifically, item-writing flaws and items that are too easy, too hard, or non-discriminating are indicators of construct-irrelevant variance, which introduces systematic error limiting the ability to interpret assessment scores accurately. The results of the current study revealed that 37% of the items sampled included at least one IWF. Pharmacy faculty should be mindful that IWF introduce error, which weakens the reliability and validity evidence for examinations and penalizes some students and calls into question the use of the scores (Downing, 2005).

Furthermore, the CTM examinations in pharmacy education are designed to assess student achievement of education outcomes outlined in the ACPE Standards and to assure that pharmacy students are well-prepared to assume the clinical responsibilities of the Introductory Pharmacy Practice Experiences (IPPEs), Advanced Pharmacy Practice Experiences (APPEs), and to enter pharmacy practice upon completion of the Pharm.D. program. It is critically important that pharmacy faculty strive to develop psychometrically-sound test items that will

prepare students for successful completion of the North American Pharmacist Licensure Examination (NAPLEX), which is a requirement to become a licensed pharmacist in the United States. The NAPLEX is a 6-hour exam that includes 250 computer-based questions with a scaled passing score of 75.

Limitations

The findings of the current study should be considered within the context of several limitations. First, the IWFEI was developed to evaluate general chemistry examination questions and has not been validated in other disciplines. However, criteria outlined in the IWFEI were developed from item-writing guidelines found in the literature (Breakall et al., 2019; Haladayna et al., 2010). Additionally, the findings of the current study are consistent with previous studies that have reviewed exams across multiple disciplines (Breakall et al., 2019; Rush et al., 2016; Stagnaro & Downing, 2006; Tarrant et al., 2006), suggesting that the IWFEI is applicable to other disciplines, including pharmacy.

Second, despite rigorous methods and multiple raters in evaluating items from the CTM examinations using the IWFEI, errors in coding or missed IWFs are possible. However, the rigorous method for conducting the coding and the background of the raters should make any of these errors unlikely. Both raters have an extensive educational background (e.g., a combined 30 years of providing faculty development), and the interrater agreement for the IWFEI was 83%, which is considered in the literature to be a good reliability rate (McMillan, 2016).

Another limitation of the study is sample size of the selected item pool. For example, the association between IWFs and psychometric indices could have been affected by the differences in cell size across the categories of the independent variable. Statistical power typically increases with increasing sample size. The sample size used for this study may not have been large enough

to detect the correlational relationships among the presence of item-writing flaws and the psychometric properties, including item difficulty, item discrimination, and average answer time. By broadening the sample to include test items from the 2018-2019 academic year, the sample size would have increased to more than 600 items. Finally, the association between IWFs and psychometric indices may have been influenced by factors not considered in the current study. Item difficulty, item discrimination, and average item answer time may be influenced by other factors outside of item-writing errors that were not measured by the current study. For example, the exams in the current study were timed. In a timed exam, it is possible that the average item answer time decreased and incorrect responses increased if students were rushing to complete the exam. Future research should consider accounting or controlling for these additional factors when examining the association between IWFs and psychometric indices.

Recommendations

In spite of the noted limitations, the findings of the current study can inform future research. First, future research should be conducted to validate the IWFEI for multiple disciplines including pharmacy. The results of the current study, in addition to previous research (e.g., Breakall et al., 2019; Rush et al., 2016; Stagnaro & Downing, 2006; Tarrant et al., 2006), suggest the IWFs identified by the IWFEI are common across disciplines and prevalent in pharmacy assessments. Future research may consider conducting full validation of the IWFEI for evaluating exams in the Doctor of Pharmacy curriculum. Validating the IWFEI for pharmacy examinations would provide faculty and administrators at colleges and schools of pharmacy an efficient and reliable tool for evaluating pharmacy assessments to ensure the interpretation of scores are valid and reliable. Furthermore, the current study could be replicated with a larger sample size to provide additional empirical evidence about the relationship between IWFs and

the psychometric indices, including item difficulty, item discrimination, and average item answer time.

Second, the current study was descriptive and correlational, which limits the ability for researchers and faculty to draw conclusions about the effects IWFs have on psychometric indices and student performance. Descriptive correlational research cannot determine cause and effect; therefore, the current study cannot determine if IWFs were the cause of the poor psychometric indices or their effect. Additionally, outside factors (e.g., timed exam) may influence psychometric indices, and the current study could not control for those outside factors. Therefore, future research should consider experimental designs to directly test the effects of IWFs on both item psychometric indices and reliability and validity of the examination. These experimental designs may also help illicit specific recommendations to improve an examination.

Finally, future research may consider comparing the quality of locally developed exam questions to the quality of exam questions developed by textbook publishers available to faculty. There has been some evidence that exam questions provided by publishers also suffer from the same IWFs and threats to validity and reliability as locally developed exam questions (Masters et al., 2001). However, the quality of publisher developed exam questions has not been compared to the quality of locally developed exam questions. Knowing which type of exam question (i.e., locally or publisher developed) are less likely to have IWFs may assist colleges and schools of pharmacy in creating an assessment strategy with higher quality examinations.

In addition to future research, findings of the current study can be used to inform current practices. First, due to the significant percentage of locally developed exam questions that had at least one IWF, colleges and schools of pharmacy may consider in-depth evaluations of the examinations provided to pharmacy students. In addition to locally developed exam questions,

Masters et al. (2001) conducted a review of the supplemental materials provided by textbook publishers and found these resources lacked evidence of best practices. To ensure that the exams given to pharmacy students meet the criteria for high quality assessments outlined by the ACPE, test items need to be evaluated or reviewed prior to administration. This can be done in a collaborative review of items across examinations, including a formal peer review of test items. Additionally, many faculty are unfamiliar with how to interpret data from item analysis reports to inform curricular improvement partly due to the fact that they have not been trained in educational and testing theory (Chiavaroli and Familiar, 2017). However, colleges and schools of pharmacy should dedicate resources to assuring that course coordinators receive sufficient faculty development, resources, and guidance to interpret and use the indices of item difficulty, item discrimination and average item answer time to improve the validity and reliability of test items. Quagrain and Arhin (2017) suggests that an item analysis is essential in improving items which will be used again in future tests; it can also be used to eliminate misleading items in a test. Items having average difficulty and high discriminating power with functional distractors should be integrated into future tests to improve the quality of the assessment (Quagrain and Arhin, 2017).

Second, Doctor of Pharmacy curricula and pharmacy residency programs should consider including courses in teaching, education, and/or assessment as a part of the curriculum. Item-writing flaws were common in the CTM examinations included in the current study. Additionally, Masters et al. (2001) and Pate and Caldwell (2014) noted a lack of formal teaching and assessment training in pharmacy education programs may contribute to faculty members' significant inadequacies in creating high quality exams. Including a course in education and

assessment as part of the core curriculum in Doctor of Pharmacy programs may assist in increasing the quality of future examinations given to pharmacy students.

Third, colleges and schools of pharmacy may consider incorporating multiple types of assessments into their pharmacy education programs. Multiple-choice exams are frequently used in pharmacy education because they can be graded quickly and efficiently using software programs. Additionally, multiple-choice exams provide objective score data for a large number of items and a large number of test takers (Epstein, 2007; Kane, 2006; McBrien, 2018). However, based on the findings of the current study and previous research, multiple-choice exams can suffer from flaws that hinder validity and reliability. Varying the assessments given to pharmacy students may assist in reducing the threats to validity and reliability in assessment that come from only using one type of assessment.

Conclusion

The purpose of this study was to examine the frequency in which IWFs occur on locally developed high-stakes summative assessments for the CTM sequence of courses at a school of pharmacy using eight criteria included on the IWFEI that was developed based on reviews of published literature on item-writing guidelines (Breakall et al., 2019). To answer the research questions outlined for this study, I used the IWFEI to evaluate test items from the second-year (P2) and third-year (P3) CTM examinations during the 2017-2018 academic year. Three hundred thirteen items were selected from a stratified random sample from a pool of 1,373 items to assure proportionate representation in each stratum

Findings of the current study suggest IWFs are common in the exams, with 37% of items having at least one IWF. Item-writing flaws continue to be a serious problem in multiple-choice

exams that affect the psychometric indices of exam questions and may possibly hinder student test performance (Board & Whitney, 1972; Breakall et al., 2019; DiBattista & Kurwaza, 2011; Downing, 2005; Pham et al., 2018; Reichert, 2011). The American Association of Colleges of Pharmacy (AACP) should consider creating a test development guide based on the *Handbook of Test Development* (Lane, Raymond, & Haladyna, 2016) and the *Standards for Educational and Psychological Testing* (APA, 2014) that is available to the 133 accredited colleges and school of pharmacy in the United States. Additionally, colleges and schools of pharmacy should conduct systematic evaluations of the quality of the multiple-choice assessments that are given to pharmacy students. The IWFEI is one possible method for conducting these systematic evaluations. Improving the examinations given to pharmacy students has the potential to improve student performance and is consistent with recommendations for assessment elements ensuring that students are prepared to enter pharmacy practice as outlined in Section I of ACPE Standard 24.

REFERENCES

- Accreditation Council for Pharmacy Education. Accreditation Standards and Guidelines for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree. https://www.acpe-accredit.org/pdf/S2007Guidelines2.0_ChangesIdentifiedInRed.pdf. Published 2007.
- Accreditation Council for Pharmacy Education. Accreditation Standards and Key Elements for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree.. <https://acpe-accredit.org/pdf/PoliciesProceduresJune2016.pdf>. Published 2016. Accessed September 2, 2016
- Albano, A. (2018). *Introduction to educational and psychological measurement using R*. Retrieved from <http://www.thetaminusb.com/intro-measurement-r/index.html#license>
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (5th ed.). Washington, DC: American Educational Research Association.
- Anderson, H.M.; Anaya, G.; Bird, E.; Moore, D.L., (2005). A Review of Educational Assessment. *American Journal of Pharmaceutical Education*, 69(1), p. 12.
- Angelo, Thomas (1995). Reassessing (and Defining) Assessment. *American Association for Higher Education* 48(3): 7.
- Attali, Y., & Bar-Hillel, M. (2003). Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable. *Journal of Educational Measurement*, 40(2), 109-128. doi:10.1111/j.1745-3984.2003.tb01099.x

- Baig, M., Ali, S. K., Ali, S., & Huda, N. (2014). Evaluation of multiple choice and short essay question items in basic medical sciences. *Pakistan Journal of Medical Sciences* 2014, 30(1), 3–6. doi:10.12669/pjms.301.4458
- Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*, 20, 369–382.
- Birkhead, S. F. (2015). *Nurse educator practices in the measurement of student achievement using multiple-choice tests in prelicensure programs in New York state* (Unpublished doctoral dissertation). The Sage Colleges, New York.
- Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, 77(4), 1-5.
- Camili G. & Shepard, L.A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.
- Case, S. M., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners.
- Chiavaroli, N., & Familiarì, M. (2011). When majority doesn't rule: The use of discrimination indices to improve the quality of MCQs. *Bioscience Education*, 17(1), 1–7.
doi:10.3108/beej.17.8
- Collins, J. (2006). Writing Multiple-Choice Questions for Continuing Medical Education Activities and Self-Assessment Modules. *RadioGraphics*, 26(2), 543-551.
doi:10.1148/rg.262055145

- Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian, 12*(1), 19-24.
doi:10.1016/s1322-7696(08)60478-3
- Creating Valid Exams and Exam Items through the Use of Psychometrics. (2019, October 24). Retrieved January 23, 2021, from <https://examsoft.com/resources/exam-quality-use-psychometric-analysis>.
- Crehan, K. & Haladyna T.M. (1991). "The Validity of Two Item-Writing Rules." *The Journal of Experimental Education 59* (2) 183-92.
- Crisp G.T. and Palmer E.J. (2007) Engaging Academics with a Simplified Analysis of their Multiple-Choice Question (MCQ) Assessment Results. *Journal of University Teaching and Learning Practice, 4* (2), 88-106
- Cross, K. J. (2000). *Cognitive levels of multiple-choice items on teacher-made tests in nursing education* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (AA19992186).
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medication education. *Medical Education, 22*, 109–117.
doi:10.1111/j.1365-2923.2009.03425.x.
- De Pew, D. D. (2001). *Validity and reliability in nursing multiple-choice testing and the relationship to NCLEX success: An internet survey* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses Database (3040905).
- Dew, J. R., & Nearing, M. M. (2004). *Continuous quality improvement in higher education*. Westport (Conn.): Praeger.

- DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, 2(2), 1–23.
- Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77(10), p. 103-104.
- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10, 133–143.
doi:10.1007/s10459-004-4019-5
- Downing, S.M. (2006). Selected -response item formats in test development. In S.M. Downing and T.M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.
- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333.
doi:10.1046/j.1365-2923.2004.01777.
- Downing S.M., & Halaydna T.M. (Eds.) 2006. *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.
- Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of Educational Measurement* (4th ed.). Sydney: Prentice-Hall of Australia.
- Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: what are textbooks authors telling teachers. *The Journal of Educational Research*, 83(5), 289–293.

- Epstein RM (2007). Assessment in Medical Education. *The New England Journal of Medicine*. 356:387-96.
- Ewell, P.T. (2002). "An Emerging Scholarship: A Brief History of Assessment." In Trudy W. Banta and Associates (Eds.), *Building A Scholarship of Assessment* (pp. 3-25). San Francisco: Jossey-Bass.
- Ewell, P. T. (2009, November). *Assessment, accountability, and improvement: Revisiting the tension* (NILOA Occasional Paper No. 1). Urbana, IL: University of Illinois and Indiana University, National Institute of Learning Outcomes Assessment. Retrieved from http://www.learningoutcomeassessment.org/documents/PeterEwell_005.pdf
- Frey, B.; Petersen, S.; Edwards, L.; Teramoto Pedrotti, J.; and Peyton, V (2005). Item-writing Rules: Collective Wisdom (2005). *College of Education Faculty Research and Publications*. 52.
- Gronlund, N. E. (2006). *Assessment of student achievement* (8th ed). Boston, MA: Pearson.
- Goodwin, L.D. (2002). Changing conceptions of measurement validity: An update on the new standards. *Journal of Nursing Education*, 41, 100-106.
- Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37–50.
doi:10.1207/s15324818ame0201_3
- Haladyna, T.M. (2004) *Developing and validating multiple-choice test items*. 3rd edn. New Jersey: Lawrence Erlbaum.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334.

- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- Harasym, P. H., Leong, E. J., Violato, C., Brant, R., & Lorscheider, F. L. (1998). Cuing effect of “All of the above” on the reliability and validity of multiple-choice test items. *Evaluation & the Health Professions, 21*(1), 120–133. doi:10.1177/016327879802100106
- Hicks, N.A. (2014), "Establishing the validity and reliability of the fairness of items tool". *Dissertations*. Paper 154.
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.
- Kean, J. & Reilly, J. (2014). Classical Test Theory. In Hammond, F., Malec, J., Buschbacher, R., & Buschbader, R. (Eds). *Handbook for clinical research: Design, statistics, and implementation*. Retrieved from <https://ebookcentral-proquest-com.proxy.library.vcu.edu>.
- Khan, M. U., & Aljarallah, B. M. (2011). Evaluation of modified essay questions (MEQ) and multiple choice questions (MCQ) as a tool for assessing the cognitive skills of undergraduate medical students. *International Journal of Health Sciences, Qassim University, 5*(1).
- Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple-choice questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal, 18*(1), 68.
doi:10.18295/squmj.2018.18.01.011
- Killingsworth, E. E. (2013). *Nursing faculty decision making about best practices in test construction, item analysis, and revision* (Unpublished doctoral dissertation). Mercer University, Atlanta, GA.

- Laerd Statistics (2018). Spearman's correlation using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from <https://statistics.laerd.com/>
- Lane, S., Raymond, M. R., & Haladyna, T. M. (2016). *Handbook of test development*. New York: Routledge, Taylor & Francis Group.
- Leedy, P. D., Ormrod, J. E., & Johnson, L. R. (2016). *Practical research: Planning and design*. NY, NY: Pearson.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- Masters, J. C., Hulsmever, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying textbooks used in nursing education. *The Journal of Nursing Education*, 40(1), 25–32.
- McBrien, S.B. (2018). Effects of Structural Flaws on the Psychometric Properties of Multiple-Choice Questions.
- McMillan, James H. *Fundamentals of Educational Research*. Pearson, 2016.
- McDonald, M. (2018). *The nurse educators guide to assessing learning outcomes*. Burlington, MA: Jones & Bartlett Learning.
- McHugh M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276–282.
- Medina MS, Plaza CM, Stowe CD (2013). Report of the 2012-13 Academic Affairs Standing Committee: revising the Center for the Advancement of Pharmacy Education (CAPE) educational outcomes 2013. *American Journal of Pharmaceutical Education*.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan

- Morrison, S., & Free, K. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education, 40*(1), 17–24.
- National Institute of Education (1984). *Involvement in Learning: Realizing the Potential of American Higher Education*. Washington, D.C.: U.S. Government Printing Office.
- Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development, 29*(2), 52–57. doi:10.1097/NND.0b013e31828d1108
- Novick, M. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Oermann, M. H. P. R. A., & Gaberson, K. P. R. C. C. (2019). *Evaluation and testing in nursing education*, sixth edition. ProQuest Ebook Central <https://ebookcentral-proquest-com.proxy.library.vcu.edu>
- Pallant, J. (2016). *SPSS Survival Manual A Step By Step Guide to Data Analysis Using SPSS Program* (6th ed.). London, UK McGraw-Hill Education.
- Pais, J., Silva, A., Guimaraes, B., Povo, A., Coehlo, E., Silva-Pereirqa, F., Severo, M. (2016). Do item-writing flaws reduce examinations psychometric quality? *BMC Research Notes, 9*, 1–7. doi:10.1186/s13104-016-2202-4
- Paniagua, M. A., & Swanson, D. B. (2016). *Constructing written test questions for the basic and clinical sciences*. Philadelphia, PA: National Board of Medical Examiners.
- Pham, H., Besanko, J., & Devitt, P. (2018). Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple-choice question. *MedEd Publish, 7*(4). doi:10.15694/mep.2018.0000225.1

- Pate, A., & Caldwell, D. J. (2014). Effects of multiple-choice item-writing guideline utilization on item and student performance. *Currents in Pharmacy Teaching and Learning*, 6, 130–134.
- Professional Testing, Inc. (n.d.). Retrieved from https://www.proftesting.com/test_topics/steps.php#9
- Quality Improvement Essentials Toolkit: IHI. (n.d.). Retrieved from <http://www.ihl.org/resources/Pages/Tools/Quality-Improvement-Essentials-Toolkit.aspx>.
- Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, 4(1). doi:10.1080/2331186x.2017.1301013
- Richman, H., & Hrezo, M. (2017). The Trouble with Test Banks. *Perspectives In Learning*, 16 (1).
- Ray, M. E., Daugherty, K., Lebovitz, L., Rudolph, M., Shuford, V. P., & DiVall, M. V. (2018). Best practices on exam construction, administration and feedback. *American Journal of Pharmaceutical Education*, 82(10), 7066; DOI: <https://doi.org/10.5688/ajpe7066>
- Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & DiVall, M. V. (2019). Best practices on exam item construction and post-hoc review. *American Journal of Pharmaceutical Education*, 83(7). DOI: <https://doi.org/10.5688/ajpe7204>
- Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, 16, 1–10. doi:10.1186/s12909-016-0773-3

- Sansgiry, S. S., Bhosle, M., & Sail, K. (2006). Factors That Affect Academic Performance Among Pharmacy Students. *American Journal of Pharmaceutical Education*, 70(5), 104. doi:10.5688/aj7005104
- Shete, A. N., Kausar, A., Lakhkar, K., & Khan, S. T. (2015). Item analysis: An evaluation of multiple choice questions in Physiology examination. *Journal of Contemporary Medical Education*, 3, 106-109. doi: 10.5455/jcme.20151011041414.
- Smith, W. Z. (2016). *The effects of scaling on trends of development: Classical test theory and item response theory* (Unpublished doctoral dissertation). University of Nebraska-Lincoln, Lincoln, NE.
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 160-169.
- Stagnaro-Green, A. S., & Downing, S. M. (2006). Used of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher*, 28(6), 566–568.
- Standards for educational and psychological testing*. (2014). Washington, DC: American Educational Research Association.
- Sullivan, G.M. (2011) A Primer on the Validity of Assessment Instruments. *Journal of Graduate Medical Education*: June 2011, Vol. 3, No. 2, pp. 119-120.
- Taber, K. S. (2017). The Use of Cronbach’s Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, 48(6), 1273–1296. doi: 10.1007/s11165-016-9602-2.

- Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 6(6), 354–363. doi:10.1016/j.nedt.2006.07.006
- Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42, 198–206.
- Towns, M. H. (2014). Guide to Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *Journal of Chemical Education*, 91(9), 1426-1431. doi:10.1021/ed500076x
- Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). New York, NY: Springer Publishing Company.
- Werking, R. H., Wilson, P., & Hassenger, R. (1986). "Integrity in the College Curriculum": Three Perspectives Integrity in the College Curriculum: A Report to the Academic Community. The Project on Redefining the Meaning and Purpose of Baccalaureate Degrees. *The Library Quarterly*, 56(2), 167-179. doi:10.1086/601723.
- Williams, K. (1998). Assessment and the challenge of skepticism. In D. Carr (Ed.), *Education, knowledge, and truth: Beyond the postmodern impasse*. London, UK: Routledge.

APPENDIX A - ITEM WRITING FLAWS EVALUATION INSTRUMENT GUIDE

APPENDIX 1
ITEM WRITING FLAWS
EVALUATION INSTRUMENT
(IWFEI)
USEAGE GUIDE

Jared Breakall, Dr. Chris Randles, and Dr. Roy Tasker

Table of Contents

Introduction to Multiple-Choice Item Format	Page 2
Is the Test Item Clear and Succinct?	Page 3
If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded?	Page 4
If the answer choices are numerical: Are they listed in ascending or descending order?	Page 5
If the answer choices are verbal: Are the answer choices all approximately the same length?	Page 6
Does the item avoid “all of the above” as a possible answer choice?	Page 7
Does the item avoid grammatical and phrasing cues?	Page 8
Could the item be answered without looking at the answer choices?	Page 9
Does the item avoid complex K-type item format?	Page 10
Is this item linked to one or more objectives of the course?	Page 11
Are all answer choices plausible?	Page 12
Are there six or less thinking steps needed to solve this problem?	Page 13
Does the exam avoid placing three or more items that assess the same concept next to each other?	Page 14
Does the exam avoid placing three or more difficult items next to each other?	Page 14
Is there an even distribution of correct answer choices?	Page 14
Does the exam avoid linking performance on one item with performance on others?	Page 15
References	Page 16

Introduction to Multiple Choice Item Format

The diagram shows a multiple choice question within a rectangular box. The question text is "9. How many moles of K⁺ ions are in 30 mL of 0.60 M K₃PO₄?" and the answer choices are (a) 0.054, (b) 0.042, (c) 0.036, (d) 0.018, and (e) 0.006. The choice (a) is bolded. Labels with arrows point to various parts: "Item" points to the entire box; "Stem of the Item" points to the question text; "Answer choices" points to the list of options; "Correct response (keyed answer)" points to the bolded choice (a); and "Distractors" points to the other four choices.

_____ 9. How many moles of K⁺ ions are in 30 mL of 0.60 M K₃PO₄?

(a) 0.054

(b) 0.042

(c) 0.036

(d) 0.018

(e) 0.006

Item

Stem of the Item

Answer choices

Correct response (keyed answer)

Distractors

Is the test item clear and succinct?

- The stem can only be interpreted as having one meaning.
- The stem doesn't include any extra information or wording (**Needed context is appropriate**).
- The answer choices don't include any extra information or wording
- There is clearly only one correct answer choice

Good Example

- _____ 1. The atomic weight of silicon is 28.0855. Round this number to 4 significant figures.
- (a) 28.0
(b) 28.08
(c) 28.09
(d) 28.086
(e) 28.1

The stem has only one interpretation with no extra information.

The answer choices don't include any extra information and only one answer choice can be interpreted as correct.

Poor Example

- _____ 6. It takes 19 days for a particular nuclide to decay 30% of its original activity. What is the half-life of this nuclide?
- (a) It would take 0.44 days
(b) It would take 11 days
(c) It would take 16 days
(d) It would take 27 days
(e) It would take 37 days

The stem could be interpreted as decaying from 100% to 70% or as decaying from 100% to 30%. This makes the question unclear.

The answer choices are not as succinct as possible. They include extra information/wording. "It would take" could be removed.

Poor Example

- _____ 3. Aspirin is a pain killer that has a density of 1.40 g/cm³. What is the amount (in moles) of aspirin, C₉H₈O₄, in a 325 mg tablet that is 100% aspirin?
- (a) 0.00180 mol
(b) 0.00325 mol
(c) 0.467 mol
(d) 1.80 mol
(e) 2.80 mol

The density of Aspirin is extra information that is not needed to solve the problem. This introduces student ability to determine needed information as a variable in student performance. The question is no longer just testing the intended chemistry content.

If the item uses negative phrasing such as “not” or “except”, is the negative phrase bolded or capitalized?

- The words “not” or “except” should be bolded or capitalized if included in the item.
- Avoiding “not” or “except” is ideal in most cases.

Good Example

- _____ 1. Which of the following contains a triple bond?
- ethylene
 - ethane
 - propene
 - benzene
 - propyne

The stem of this question doesn't contain negative phrasing. This is ideal for most items.

Good Example

- _____ 1. Which of the following does **NOT** contain a triple bond?
- Butyne
 - Pentyne
 - Hexyne
 - Benzene
 - Propyne

The negative phrase is capitalized.

Poor Example

- _____ 1. All of the following processes are exothermic **except**:
- Combustion of propane
 - Rusting of iron
 - Freezing of water
 - Melting of ice

The word 'except' is not bolded or capitalized.

If the answer choices are numerical:

Are they listed in ascending or descending order?

- Numerical answer choices should be listed in ascending or descending order. For example: 1,2,3 vs. 2,1,3

Good Example

_____ 12. What is the molality of a solution prepared by mixing 12.0 g benzene (C_6H_6) with 38.0 g CCl_4 ?

- a. 0.240 *m*
- b. 0.316 *m*
- c. 0.508 *m*
- d. 0.622 *m*
- e. 4.05 *m*

The answer choices are written in ascending numerical order.

Poor Example

_____ 12. What is the molality of a solution prepared by mixing 12.0 g benzene (C_6H_6) with 38.0 g CCl_4 ?

- a. 4.05 *m*
- b. 0.240 *m*
- c. 0.622 *m*
- d. 0.316 *m*
- e. 0.508 *m*

The answer choices are not written in ascending or descending numerical order.

This is *Not Applicable* if an item is K-type

Symbolic answer choices, such as electron configurations or chemical formulas are NOT considered numerical.

This criterion would be Not Applicable.

If the answer choices are verbal:

Are the answer choices all approximately the same length?

- An answer choice should **not** be substantially longer or shorter than any of the other choices. This may cue students to an answer without consideration of the item content.

Good Example

- _____ 19. What is the purpose of standardizing a solution?
- To determine its purity.
 - To determine its concentration.
 - To measure its volume.
 - To determine its molecular formula.
 - To determine the endpoint.

This item keeps all answer choices approximately the same length.

Poor Example

- _____ 19. What is the purpose of standardizing a solution?
- To determine its purity.
 - The purpose is to determine the concentration of the solution
 - To measure its volume.
 - To determine its molecular formula.
 - To determine the endpoint.

One answer choice is significantly longer than the others.

(This item includes phrasing cues as well (see page 8))

This is *Not Applicable* if an item is K-type

Symbolic answer choices, such as electron configurations or chemical formulas are NOT considered verbal.

This criterion would be *Not Applicable* if answer choices are symbolic.

Does the item avoid “all of the above” as a possible answer choice?

- Using “all of the above” as an answer choice can cue students to eliminate distractors.

Good Example

- _____ 1. Which of the following contains a triple bond?
- ethylene
 - ethane
 - propene
 - benzene
 - propyne

This item doesn't use all of the above or none of the above as answer choices

Poor Example

- _____ 1. Which of the following contains a triple bond?
- ethylene
 - ethane
 - propene
 - propyne
 - all of the above

The use of 'all of the above' is quickly eliminated when a student recognizes any molecule that doesn't contain a triple bond.

For K-type items, if an answer choice includes all of the possibilities, then it violates this guideline.

Does the item avoid grammatical and phrasing cues?

- A cue leads a student to the right answer or to eliminating a distractor.
- A grammatical cue is a difference in grammar between the stem and the answer choices or between answer choices.
- A phrasing cue is where a phrase from the stem is used in one distractor or in the correct answer.

Good Example (Grammatical cuing)

__ 19. Carbon has ____ proton(s).

- (a) One
- (b) Three
- (c) Six
- (d) Twelve

← This item keeps the grammar of the stem consistent with the answer choices.

Poor Example (Grammatical cuing)

__ 19. Carbon has ____ protons?

- (a) One
- (b) Three
- (c) Six
- (d) Twelve

← Answer choice A does not fit the grammatical structure of the stem. This may cue students to it being the incorrect answer.

Good Example (Phrasing cues)

__ 19. How many proton(s) does Carbon have?

- (a) One
- (b) Three
- (c) Six
- (d) Twelve

← This item gives no phrasing cues to the correct answer.

Poor Example (Phrasing cues)

__ 19. How many proton(s) does Carbon have?

- (a) One
- (b) Three
- (c) Six protons
- (d) Twelve

← This item using a phrase (protons) from the stem in the correct answer choice. This may cue students to choose this answer.

Could the question be answered without looking at the answer choices?

- It is important to write the stem of an item in a way that it could be answered without looking at the answer choices. This ensures that the central idea is included in the stem.

Good Example

__ 19. Hydrogen can have how many protons?

- (a) 0 or 1
- (b) 1
- (c) 1 or 2
- (d) 2

This item contains the central idea in the stem and can be answered without the answer choices.

Poor Example

__ 19. Hydrogen:

- (a) can have 0 or 1 protons
- (b) can only have 1 proton
- (c) can have 1 or 2 protons
- (d) can only have 2 protons

This item cannot be answered without looking at the answer choices. The stem doesn't contain the central idea.

Does the item avoid complex K-type item format?

- K-Type items have answer choices that contain combinations of other answer choices.
- K-Type Items have been shown to cue students to the correct answer
- Ordering items, such as the ordering of ion-size, are not considered to be K-type.

Good Example

- _____ 12. What is the molality of a solution prepared by mixing 12.0 g benzene (C_6H_6) with 38.0 g CCl_4 ?
- 4.05 *m*
 - 0.240 *m*
 - 0.622 *m*
 - 0.316 *m*
 - 0.508 *m*

This item avoids K-type format

Poor Example

- _____ 5. Which of the following properties influence the frequency of a molecular vibration, seen in infrared absorption spectra?
- Size (radius) of the atoms on each side of the bond
 - Strength of the bond between atoms
 - Mass of the atoms on each side of the bond
- i only
 - ii only
 - iii only
 - i and ii
 - ii and iii

This is an example of a K-type question

Good Example

- _____ 17. A double bond is composed of _____ bond(s) and _____ rotate.
- Two sigma; cannot
 - Two pi; cannot
 - One sigma and one pi; cannot
 - One sigma and one pi; can
 - Two sigma; can

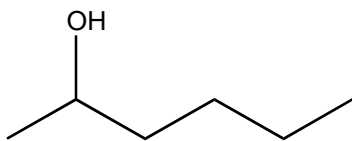
This is a fill-in-the-blank item.
This is NOT in k-type format.

Is this item linked to one or more objectives of the course?

- Test items should test one or more objectives of the course.

Good Example:

_____ 3. What is the correct formula for the organic molecule shown below?



- a. $C_7H_{14}O$
- b. $C_6H_{14}O$
- c. $C_7H_{13}O$
- d. $C_6H_{13}O$
- e. $C_4H_{10}O$

Hypothetical Course Objectives

Students Should Be Able to:

1. Interconvert between skeleton structures and chemical formulas.
2. Determine the number of atoms in a molecule based on various representations.
3. Draw Lewis Dot Diagrams from molecular formulas.

This item directly assesses course objective 1 and indirectly assesses objective 2.

Poor Example:

_____ 2. Alkenes by definition contain a _____.

- a. C=C bond
- b. C≡C bond
- c. C-C bond
- d. C=H bond
- e. C≡H bond

This item doesn't assess any of listed the course objectives.

Are all answer choices plausible?

- **All** distractors should be made by using common student errors or misconceptions. Even if only one distractor is not, then the item is in violation of this guideline.
- Each distractor should have been chosen by more than 5% of the students tested.

Good Example

- _____ 6. What kind of electromagnetic radiation is able to break bonds?
- a. Ultraviolet**
 - b. Infrared
 - c. Visible
 - d. Microwave
 - e. Radiowaves

All the answer choices are likely to be chosen. They are all viable forms of electromagnetic radiation

Poor Example

- _____ 6. What kind of electromagnetic radiation is able to break bonds?
- a. Ultraviolet**
 - b. Infrared
 - c. Visible
 - d. Microwave
 - e. The bonds of friendship are too strong to break.

Answer choice E is not plausible.

Are there six or less thinking steps needed to solve this problem?

- A thinking step is a small cognitive process that must be taken to solve a problem (Johnstone & El-Banna, 1986).
- The thinking steps should be based on the average student taking the exam.

The following is an example of the thinking steps that may exist in an item. Reproduced from (Johnstone & El-Banna, 1989) with permission from the Royal Society of Chemistry.

'What volume of molar hydrochloric acid would be exactly neutralized by ten grams of chalk?'

Thinking Steps:

1. chalk---calcium carbonate (recall)
2. calcium carbonate =-CaCO₃ (recall or deduce)
3. Formula weight of CaCO₃= 100 g (calculate)
4. When it reacts with hydrochloric acid, what are the products? (recall)
5. Write a balanced equation (transformation)
6. Recognize that 1 mole CaCO₃~ 2 moles HCl (deduce)
7. = 2 litres of molar HCl (recall)
8. 10 g CaCO₃ ~ 1/10 mole ~ 1/5 mole HCl (deduce)
9. ~ 1/5 litre molar HCl (recall) = 200 ml molar HCl

Because this item can be viewed as having nine thinking steps, it may be measuring working memory capacity along with the students understanding of chemistry. This negatively effects the validity of the item.

Does the exam avoid placing three or more items that assess the same concept or skill next to each other?

- Placing three or more similar questions next to each other may cue students to what the correct answer may be.
- A concept or skill is defined as the same learning objective.

Does the exam avoid placing three or more difficult items next to each other?

- A difficult item is defined as an item that you believe less than 50% of students will get correct.

Is there an approximately even distribution of correct answer choices?

- Correct answer choices should be approximately evenly distributed. No two distractors should have a difference of greater than two in frequency appearing in the key.

$$\text{Even Distribution} = \frac{i_t}{a} \pm 1 \text{ (for each answer choice)}$$

i_t = Total number of items in the exam

a = Answer choices per item

Good Example

Answer Key 4 choices per item

1) A

2) C

3) C

4) D

Poor Example

Answer Key 4 choices per item

1) A

2) C

3) C

4) C

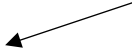
Does the exam avoid linking performance on one item with performance on others?

- Items should be independent of one another on an exam.

Good Example

1. What is the molar mass of $C_6H_{12}O_6$?
 - a) 168.2
 - b) 180.2
 - c) 200.2
2. How many moles of water are there in a 14.0 gram sample of H_2O ?
 - a) 1.00
 - b) 0.780
 - c) 0.550

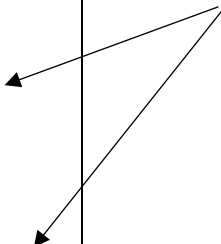
Students can do well on each item independent of each other.



Poor Example

1. What is the molar mass of $C_6H_{12}O_6$?
 - a) 168.2
 - b) 180.2
 - c) 200.2
2. Based your answer to question one, would a 0.5 mole sample of $C_6H_{12}O_6$ weigh more or less than 90.0 grams?
 - a) More
 - b) Less
 - c) Not enough information to tell

Students **cannot** succeed on item two if they don't succeed on item one.



References

- Johnstone, A., & El-Banna, H. (1986). Capacities, demands and processes - a predictive model for science education. *Education in Chemistry, 23*, 80–84.
- Johnstone, A., & El-Banna, H. (1989). Understanding learning difficulties—A predictive research model. *Studies in Higher Education, 14*(2), 159–168. <http://doi.org/10.1080/03075078912331377486>

APPENDIX B - CURRICULUM VITAE

VERONICA POWELL SHUFORD, M.ED.

cell: (804) 873-2170 e-mail: vpshufor@vcu.edu

EDUCATION

- **Doctor of Philosophy in Education – Research, Assessment and Evaluation (Candidate)**
Virginia Commonwealth University, Richmond, Virginia. June 2016 - present
Dissertation proposal title: Examining the Effect of Item-Writing Flaws on the Psychometric Parameters of Pharmacy Therapeutics Examinations
- **Post-Baccalaureate Certificate in Higher Education Assessment**
James Madison University, Harrisonburg, Virginia. May 2013
- **Master of Education in Adult Education – Human Resources Development**
Virginia Commonwealth University, Richmond, Virginia. May 1996
- **Bachelor of Science in Mass Communications – Public Relations**
Virginia Commonwealth University, Richmond, Virginia. May 1990

PROFESSIONAL DEVELOPMENT

- The Evaluators' Institute, Claremont Graduate University, Stephanie Evergreen Data Visualization, Washington D.C., July 2018.
- Lean Six Sigma Green Belt Training, Center for Professional Development, Virginia Commonwealth University School of Business, Richmond, Virginia, October 2016
- Basic Compliance & Ethics Academy, Society for Corporate Compliance and Ethics, Scottsdale, Arizona, February 2007
- The Instructional Developer Workshop, Darryl L. Sink Associates, Inc., June 2004
- VCU Leadership Development Program, Grace E. Harris Leadership Institute. Virginia Commonwealth University, November 2001
- Myers-Briggs Type Indicator Certified Practitioner, Consulting Psychologist Press, Inc., November 1997

EXPERIENCE

Virginia Commonwealth University – Office of the Provost and Senior Vice President for Academic Affairs, Richmond, Virginia
December 2018 – Present

Director for Program Development and Innovation (SCHEV Liaison)

- Guide VCU faculty and staff in developing academic degree program proposals, coordinating the advance of the proposals through institutional governance process and the State Council of Higher Education for Virginia (SCHEV).
- Collaborate with the Office of Academic Affairs staff on academic program review, policy development, regional accreditation, and learning outcomes assessment.
- Provide workshops to faculty and academic staff regarding degree proposal development, including criteria, processes, and timelines.

- Coach proposal authors on the development of high quality pre-proposals and proposals.
- Manage the academic program development workflow process for the preparation and approval of academic program proposals, including coordinating and monitoring documents for review and approval by university-level curriculum committees, University Council, President's Cabinet, the VCU Board of Visitors, and SCHEV.
- Develop and administer student demand surveys for proposed new academic program proposals.

**VCU School of Pharmacy – Dean's Office, Richmond, Virginia
November 2009 – December 2018**

**Director of Education and Assessment, Assistant Professor, June 2012 - present
Education Specialist, Assistant Professor, November 2009 - June 2012**

- Collected, analyzed, interpreted, and reported assessment data to assure the results are used effectively for continuous quality improvement purposes and to assist the school with its accreditation efforts.
- Developed and monitored key performance indicators (KPIs), Top 10 benchmarking report, and other metrics for continuous quality improvement.
- Provided administrative and operational support to the Executive Associate Dean for Academic Affairs for the planning, management, and coordination of the Doctor of Pharmacy curriculum and services for more than 80 fulltime faculty and 560 students.
- Designed, developed, and evaluated faculty development programs to support teaching, learning, and assessment. Topics include flipped classrooms, team-based learning, e-Learning, creating an inclusive learning environment, item-writing, curriculum mapping, computer-based testing, and the Learning Connections Inventory.
- Coordinated the Academic Affairs portion of the first-year pharmacy student orientation.
- Collaborated with the Division for Academic Success to arrange accommodations for computer-based assessments and student support with the Learning Connections Inventory.
- Developed and facilitated the School's New Faculty Orientation for all incoming faculty.
- Worked collaboratively with the senior leadership team, Curriculum Committee, Outcomes and Assessment Committee to develop and implement assessment policies and procedures to improve and advance the mission of the school.
- Served as a subject matter expert in current assessment practices related to assessment of student learning.
- Provided strategic curricular and programmatic support by collaborating with the Office of Admissions and Student Services, Vice Chairs of Education, Curriculum Committee, and Outcomes & Assessment Committee, including curriculum development, curriculum realignment, curriculum mapping, and other assessment tasks.
- Managed the course and instructor evaluation process for approximately 60 courses per year and 50 faculty.
- Developed and implemented methods to assess the effectiveness of the curriculum, including web-based testing, focus groups, student experience surveys, periodic curricular review, validation of assessment, and annual faculty course assessments.

- Led the planning, development, and implementation of assessment efforts across academic and operational areas within the school, including key performance indicator metrics and dashboard.
- Used quantitative and qualitative research methods to develop, analyze and interpret assessment data and prepare assessment and accreditation reports, including survey development and facilitating focus groups.
- Served as a resource on the Accreditation Council for Pharmacy Education (ACPE) accreditation requirements and criteria for assessing student learning outcomes.
- Served as the school administrator for Blackboard Learning Management System, ExamSoft Computer-Based Testing System, WEAVE Assessment Management System, AACP Central Survey System, and the AACP Assessment and Accreditation Management System.

Altria Client Services, Richmond, Virginia
Compliance & Integrity, Compliance Institute
October 2006 – October 2009

Learning Consultant

- Partnered with internal clients to develop or improve compliance training requirements for business units.
- Reviewed and evaluated compliance training requirements and recommended training solutions to enhance employees' knowledge and behavior related to corporate compliance issues.
- Provided project management guidance to internal clients in designing, developing, and executing compliance training.
- Advised clients on change management and communication strategies for new training implementations.
- Designed, developed, implemented, and evaluated face-to-face and web-based compliance training programs to help employees understand the legal and regulatory requirements that governed their work, including financial responsibility, Sarbanes-Oxley, political activity, animal care and use, anti-harassment, and monetary approval.
- Facilitated train-the-trainer sessions for subject matter experts on active learning, training techniques, presentation skills, and facilitation skills.
- Assessed the effectiveness of training programs by developing and conducting participant evaluations and interviews.
- Developed monthly reports including management summaries and status updates for the compliance training team and department.

VCU School of Pharmacy – Dean's Office, Richmond, Virginia
June 1999 – September 2006

Director of Academic Technology

- Developed an instructional and academic technology focus within the School of Pharmacy to support teaching and learning in the Doctor of Pharmacy, Non-Traditional Doctor of Pharmacy Pathway, and graduate programs.

- Coordinated the design and support of instructional facilities to expand the use of technology in teaching and learning.
- Designed, developed, and conducted student orientations and faculty development workshops on a variety of topics related to teaching, learning, technology, and assessment.
- Conducted best practices research to evaluate the effectiveness of various teaching and learning theories in pharmacy education.
- Planned and implemented a Student Laptop Requirement and Personal Digital Assistant (PDA) Initiative for 450 students and 70 faculty.
- Managed the network upgrade to the School's traditional wired network (10/100) and installation of a wireless network (802.11b) including 12 wireless access points, 200 data ports, and 200 power outlets in Smith Building Room 103 and other common areas of the building.
- Developed and maintained a helpdesk infrastructure to support the technology needs of 450 students, 80 faculty, and 45 staff.
- Managed a \$200,000 annual operational budget, \$90,000 annual student technology fee budget, and \$70,000 annual Capital Equipment budget.
- Managed the procurement of software and computer resources for faculty and students.
- Served as the technology consultant for the School's Inova Fairfax Medical Campus expansion.

University of Richmond, Richmond, Virginia
Division of Information Services – Academic Technology Services
February 1998 - June 1999

Instructional Technology Consultant

- Collaborated with faculty to integrate the Internet and other technologies into teaching and learning.
- Coordinated the development of workshops to help faculty enhance teaching and student learning.
- Provided academic technology training and individual consultations to faculty on the instructional uses of HTML, Adobe PageMill, Adobe Acrobat, Adobe Photoshop, Netscape Newsgroups, WebBoard, WebCT, and e-mail.

Crestar Financial Corporation, Richmond, Virginia
March 1996 – February 1998

Senior Instructional Design Consultant, Assistant Vice President
Corporate Training Department

- Conducted needs assessments to determine corporate training needs by business function. Used results to develop and deliver training and professional development programs for more than 2,300 employees.
- Designed, developed, implemented, and evaluated corporate training projects, including the SMART Link Entity computer-based training for the Commercial Lending Department, management development training, leadership development training, Myers-Briggs Type Inventory training, merger training, and diversity training to support corporate goals and objectives.

- Provided consultation to banking professionals and departments related to special corporate training projects and initiatives.

Teaching

- PHAR 652: Health Promotions & Communications in Pharmacy Practice, Health Literacy and Patient Education Materials (2 hours), September 2015 and September 2017
- PHAR 525: Pharmacy Communications, Health Literacy and Patient Education Materials Module (2 hours), September 2013, 2012, and 2011
- PHAR 691: Information Technology for the Health Professions, VCU School of Pharmacy, Spring 2006 and Fall 2004 (Course Co-Coordinator)
- Web Development with Microsoft FrontPage, Office of Non-Profit Programs, VCU Office of Community Outreach, Fall 2004 - Spring 2006 (Instructor)
- ISYS 201: Effective Use of Microcomputers, University of Richmond School of Continuing Studies (Hybrid course), Spring 1999 – Fall 2001 (Course Coordinator)

Awards

- 2020 AACP Assessment Special Interest Group Collaborative Publication Award. Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & Divall, M. V. (2019). Best Practices Related to Examination Item Construction and Post-hoc Review. *American Journal of Pharmaceutical Education*, 83(7), 7204.
- Virginia Commonwealth University Outstanding Practices in the Assessment of Student Learning, June 2016.
- Dave L. Dixon, Evan M. Sisson, Veronica P. Shuford, Virginia Commonwealth University School of Pharmacy, and Spencer E. Harpe, Midwestern University Chicago College of Pharmacy. “Use of video recorded clinic visits to improve assessment of student pharmacists’ clinical interviewing skills.” 2014 AACP Innovations in Teaching Award. American Association of Colleges of Pharmacy Annual Meeting, Grapevine, TX. July 2014.

Presentations

- “Student Perspectives on ExamSoft: The Good, The Bad, The Awesome.” ExamSoft Webinar, May 2015.
- “A Pathway to Continuous Improvement in Computer-Based Testing.” ExamSoft Webinar, March 2015.
- “Contrasting Approaches to Electronic Exams.” American Association of Colleges of Pharmacy Annual Meeting, Grapevine, TX. July 2014.
- “Creating Significant Learning Experiences for Adults.” Virginia Geriatric Education Center, Faculty Development Program. Newport News, Virginia. April 2015 and March 2014.
- “Creating Significant Learning Experiences for Adults.” Virginia Geriatric Education Center, Faculty Development Program. Richmond, Virginia. November 2014, November 2013, and January 2013.

- “Creating Significant Learning Experiences for Adults.” Virginia Geriatric Education Center, Faculty Development Program. Virginia Beach, Virginia. January 2012
- “Emerging Technology in Teaching and Learning.” Virginia Geriatric Education Center, Faculty Development Program. Newport News, Virginia. April 2015 and March 2014.
- “Emerging Technology in Teaching and Learning.” Virginia Geriatric Education Center, Faculty Development Program. Richmond, Virginia. November 2014, November 2013, and January 2013.
- “Emerging Technology in Teaching and Learning.” Virginia Geriatric Education Center, Faculty Development Program. Virginia Beach, Virginia. January 13, 2012
- “The Effective Use of Audiovisuals for Presentations” for the Department of Pharmaceutics fall seminars, September 2001, 2002, 2003, 2004, and 2005.
- “The Use of Technology in Teaching,” Pharmacy Resident Education Program, Richmond, VA, August 2001, 2002, 2003, 2004, and 2005.
- “Second Year Pharmacy Students Receive PDAs as Curriculum Resource.” Tompkins-McCaw Library Mobile Technology Fair, Richmond, VA, March 2006.
- “VCU School of Pharmacy’s PDA Initiative.” Tompkins-McCaw Library PDA-Special Interest Group, Richmond, VA, March 2005.
- “School of Pharmacy Implements Secure Testing with Blackboard”, Innovative Teaching Strategies for Faculty Using Blackboard Conference, Richmond, VA, April, 2005.
- “Cheat No More: The School of Pharmacy Implements Secure Testing with Blackboard.” VCU Emerging Technologies Day and Blackboard Conference, Richmond, VA, April 2004.
- VCU Emerging Technologies Day, Richmond, VA. Poster: “The School of Pharmacy Evaluates Software for Secure Online and Computerized Testing. Richmond, VA, April 2003.
- “Effective Presentation Strategies,” Department of Pharmaceutics Fall Graduate Seminars, September 2001.
- “WLAN Pilot Program in the School of Pharmacy.” VCU Instructional Development Center Seminars, Richmond, VA, September 2001.

Posters

- Shuford VP, Donohoe KL, Krista L. Donohoe, Kirkwood CK, Moret PM. Enhancing Pharmacy Student Professionalism by Creating Meaningful Co-Curricular Experiences. Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, Nashville, TN, July 17, 2017.
- Donohoe KL, Slattum PW, Peron EP, Powers K, Shuford VP. The Assessment of Changes in Student Pharmacists’ Knowledge, Skills, and Attitudes Toward Older Adults. Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, Nashville, TN, July 17, 2017.
- Kirkwood CK, Shuford VP, Frankart LM, Lockeman KS. The Integration of Interprofessional Education in an Established Pharm.D. Curriculum. Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, National Harbor, MD, July 11, 2015.
- Dixon DL, Sisson EM, Shuford, VP, Harpe SE. “Use of Video Recorded Clinic Visits to “Improve Assessment of Student Pharmacists Clinical Interviewing Skills.” Poster

presentation at the American Association of Colleges of Pharmacy Annual Meeting, Grapevine, TX. Poster: July 2014.

- Phipps LB, Shuford VP, Sicat BL, Kirkwood, CK. “Student Perceptions of a Peer Evaluation System in a Clinical Therapeutics Course.” Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, Chicago, IL. July 2013.
- Sicat BL, Shuford VP, Phipps LB, Kirkwood CK. Modification of the Peer Evaluation System in the Clinical Therapeutics Modules. Poster presentation at the 13th Annual Team-based Learning Collaborative Conference, San Diego, CA, March 1, 2013.
- Harpe S, Delafuente J, Shuford V, Sicat B, Venitz J. “Development of a Process to Validate the Assessment of Doctor of Pharmacy Course Objectives.” Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, San Antonio, TX. July 2011.
- Morgan LA, Shuford VP. “Students’ confidence in their abilities to achieve APPE hospital pharmacy practice competencies.” Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, San Antonio, TX. Poster: July 2011.
- Kirkwood CK, Shuford VP, Delafuente JD. “Integrating Clinical and Basic Sciences Throughout a Curriculum.” Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, San Antonio, TX. July 2011.
- Huynh CN, Willett RM, Sicat BL, Mayer SD, Polich SM, Shuford VP. “Medical-Pharmacy Interprofessional Education in a Medical Center Teaching Clinic.” Poster presentation at the Society of General Internal Medicine Annual Meeting, Phoenix, AZ. May 2011.
- Wright BA, Shuford VP, Purcell K. “A Collaborative Partnership to Support Community-based Pharmacy Preceptors’ Information Access.” Poster presentation at the Medical Library Association Annual Meeting, Minneapolis, MN. May 2011.
- Huynh CN, Willett RM, Sicat BL, Mayer SD, Polich SM, Shuford VP. “Medical-Pharmacy Interprofessional Education in a Medical Center Teaching Clinic.” Poster presentation at the Southern Society of General Internal Medicine Regional Meeting, New Orleans, LA. February 2011.
- Talluto BA, Besinque, KH, Cable GL, Kahaleh AA, Nemire R, Smith GB, Shuford VP, Henry, J. “ACCP APPE Project: Developing a Library of Resources for Preparing and Supporting the Practitioner Educator. Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, San Diego, CA. July 2006.
- Yunker NS, Shuford, VP, Kirkwood C. “Evaluation of a Web-Based Medical Terminology Module Incorporated into a Traditional Pharmacotherapy Course. Poster presentation at the American Association of Colleges of Pharmacy Annual Meeting, San Diego, CA. July 2006.
- Cheang, KI, Shuford VP. “Virtual Learning of Cardiovascular Hemodynamics in an Advanced Cardiovascular Pharmacotherapy Course.” Poster presentation at the American College of Clinical Pharmacy Spring Practice and Research Forum, Myrtle Beach, SC. April 2005.
- Shuford, VP, Smith WE. “VCU School of Pharmacy Requires Incoming Pharmacy students to Purchase Laptops. Poster presentation at the VCU Emerging Technologies Day and Blackboard Conference, Richmond, VA. April 2004.
- Calarco P, Hill LH, Wright BA, Shuford VP. “Digital Curriculum: Pharmacy and Library to Support the Pharm.D. Curriculum.” Poster presentation at the American Association of College of Pharmacy Annual Meeting, Toronto, CA. July 2001.

Publications

- Rudolph, M. J., Daugherty, K. K., Ray, M. E., Shuford, V. P., Lebovitz, L., & Divall, M. V. (2019). Best Practices Related to Examination Item Construction and Post-hoc Review. *American Journal of Pharmaceutical Education*, 83(7), 7204. doi:10.5688/ajpe7204
- Ray M.E., Daugherty K.K., Lebovitz L., Rudolph M.J., Shuford V.P., and DiVall D.V. Best Practices on Examination Construction, Administration, and Feedback. *American Journal of Pharmaceutical Education*, 2018 82:10.
- Lebovitz L, Shuford VP, Divall MV, Daugherty KK, Rudolph MJ. Creating an Arms Race? Examining School Costs and Motivations for Providing NAPLEX and PCOA Preparation. *American Journal of Pharmacy Education*, 2017 81(7).
- Sicat BL, Kreutzer KO, Gary J, Ivey C, Marlowe EP, Pellegrini, JM, Shuford VP, Simmons DF. Collaboration Among Health Sciences Schools to Enhance Faculty Development in Teaching. *American Journal of Pharmacy Education*, 2014; 78(5): 1-5.
- Sicat B, Shuford VS, Phipps LB, Kirkwood CK. TBL Trends: Modification of the Peer Evaluation System in a Longitudinal Course Sequence with Multiple Modules. *Team Base Learning Collaborative Newsletter* 2014;4(2):6-8. Available at: www.julnet.com/tblc/newsletter/tblc_newsletter_march2014.html#sicat.
- Fisher, E.J. (1995). Clinician's Guide to the Therapy of Adults with HIV/AIDS, 3rd Edition. Richmond: Virginia Commonwealth University, Medical College of Virginia. VCU HIV/AIDS Center and the Office of Faculty and Instructional Development. Veronica Shuford, Project Manager.
- Shuford, V. (Ed.) (1996). Computer Resource Guide for Medical Students. Richmond: Virginia Commonwealth University, Medical College of Virginia, School of Medicine, Office of Faculty and Instructional Development.
- Goodall, P., Hill, J., Perkins, B, & Powell, V. (1992). Integrated Leisure Options for Individuals with Traumatic Brain Injury (Special Topic Report). Richmond: Virginia Commonwealth University, Medical College of Virginia. Rehabilitation Research and Training Center on Severe Traumatic Brain Injury and the Chesterfield County Open Doors Project.
- Raines, S., Waaland, P., Powell. V. (1992). For Kid's Only: A Guide to Brain Injury. Richmond: Virginia Commonwealth University, Medical College of Virginia. Rehabilitation Research and Training Center on Severe Traumatic Brain Injury.

eLearning Modules

- Delafuente, J.C. (2003). Thrombosis Prevention and Management in the Older Patient: A Case-Based Educational Program. Richmond: Virginia Commonwealth University School of Pharmacy (audio narration).
- Goode, J.V. (2003). Online Course in the Advances in Community Pharmacy Practice and Therapeutics. Virginia Commonwealth University School of Pharmacy.
- SMART Training Curriculum (Sales Management and Relationship Tracking System) Crestar Financial Corporation (1997) (interactive computer-based instruction program).
- Entity Link Computer-Based Training Program. Crestar Financial Corporation (1996).

- Crestar Telephone BillPayer Computer-Based Training Program Crestar Financial Corporation (1996).
- Educational Coordinator/Instructional Designer, MCV Telemedicine Project: Blackstone Family Practice Center. The use of two-way interactive videoconferencing for the delivery of “live” patient consultations and educational programs to physicians in Blackstone, Virginia. August 1995 – February 1996. (research grant)
- Peng, T, Shuford, V., Stephens, C., Schlesinger, J. (1995). Clinical Simulations of Fetal Heart Rate Patterns in Labor. Richmond: Virginia Commonwealth University, Medical College of Virginia. School of Medicine, Office of Faculty and Instructional Development (interactive computer-based instruction program).
- Girerd, P., Shuford, V., Stephens, C., et al. (1995). APGO Quiz 95: Women’s Health Care. Washington, D.C.: Association of Professors of Gynecology and Obstetrics and Virginia Commonwealth University, Medical College of Virginia School of Medicine, Office of Faculty and Instructional Development (computer-based quiz program).
- Seibel, H., Seibel, W. Stephens, C., Shuford, V., Schlesinger, J. (1996). Review Questions in Gross Anatomy of the Head and Neck. Richmond: Virginia Commonwealth University, Medical College of Virginia. School of Medicine, Office of Faculty and Instructional Development (interactive computer-based instruction program).
- Returning to School Following Traumatic Brain Injury (1993). Richmond: Virginia Commonwealth University, Medical College of Virginia. Rehabilitation Research and Training Center on Severe Traumatic Brain Injury (videotape).
- Posttraumatic Epilepsy Following Brain Injury (1992). Richmond: Virginia Commonwealth University, Medical College of Virginia. Rehabilitation Research and Training Center on Severe Traumatic Brain Injury (videotape).

Committees and Advisory Groups

Virginia Commonwealth University Committees

- Health Sciences Classroom Study Advisory Group (Ad Hoc), 2015 – 2017
- Faculty Learning Community on Faculty Development, 2010 - 2014
- Tompkins-McCaw Library Mobile Technology Fair Planning Committee, 2005-2006
- University Assessment Council, 2010 – 2017
- VCU Information Technology Advisory Committee, 2000 – 2005
- VCU LAN Managers Group, 2002 – 2006
- VCUNET User Group, 2001 - 2004
- VCU Media Support Services Evaluation Committee for AV Services and Equipment, 2001
- VCU Student Computer Initiative Committee: Training Subcommittee, 2000-2001
- VCU Wireless Local Area Networking (WLAN) Committee, 2000-2001

VCU School of Pharmacy Committees

- ACPE Accreditation Self Study: Library Resources Subcommittee, 2000-2001
- ACPE Accreditation Self Study: Curriculum Subcommittee, 2012 - 2014
- Curriculum Committee, 2009 – 2018
- Information Technology Advisory Committee, 1999 – 2006 and 2010 - 2013

- INOVA Fairfax Medical Campus Planning Committee, 2005 - 2006
- Non-Traditional Pharm.D. Planning Committee, 1999 – 2006
- Outcomes and Assessment, 2009 – 2018 (Chair: 2010 – 2014)
- Skills Lab Renovation Committee (203 and 221), 2002 – 2005
- Smith Building 103 and 107 Renovation Committee, 2001-2005
- Software Review Committee, 2001 – 2006
- Strategic Planning Committee, 2005 – 2006, 2013-2014

University of Richmond Committees

- Information Services Web Development Committee, 1998 – 1999
- Information Services Faculty Development Web Committee, 1998 – 1999
- WebCT Planning and Implementation Committee, 1998 - 1999

Professional Organizations

- American Association of Colleges of Pharmacy (AACP), 1999 – 2006 and 2009 – 2018
- American Evaluation Association (AEA), 2015 - present
- Association for the Assessment of Learning in Higher Education (AAHLE), 2012 - present
- Association for Institutional Research (AIR), 2012 - present
- Educause, 2006 – present
- Team-Based Learning Collaborative, 2010 – 2018
- Society for College and University Planning, 2019 - present
- Virginia Assessment Group, 2009 – present