A CORRELATIONAL STUDY OF NURSING INSTRUCTOR USE OF BEST PRACTICE WITH MULTIPLE CHOICE QUESTIONS

by

Diane Droutman

A Dissertation Presented in Partial Fulfillment

of the Requirements for the Degree

Doctor of Nursing

University of Phoenix

The Dissertation Committee for Diane Droutman certifies approval of the following dissertation:

A CORRELATIONAL STUDY OF NURSING INSTRUCTOR USE OF

BEST PRACTICE WITH MULTIPLE CHOICE QUESTIONS

Committee:

Donna Taliaferro, RN, PhD, COI, Chair

Anne Brett, Ph.D., RN, Committee Member

Gail Williams, PhD, RN, COI, Committee Member

_____
Donna Taliaferro

_____
Anne Brett

_____
Gail Williams

_____
Hinrich Eylers, PhD
Vice Provost, Doctoral Studies
University of Phoenix

Date Approved: _____

ABSTRACT

Multiple choice questions are used as evaluations in nursing schools. Nursing instructors and nursing book publishers develop exam questions. The specific problem addressed by the study was how best practice in multiple choice test items, item analysis, and revision of choice test items used by nursing instructors. Using a survey method, this correlational design research looked at the relationship of faculty use of best practice in test item construction, analysis, and revision of multiple choice test item in nursing programs in the United States. Even though a relationship was noted, the statistical effect level was minimal. There was no correlation between grading practices and the use of best practices in test construction, test analysis and test revisions. The research results provide insight into the use of best practices and the prevalence of the inconsistencies in test construction, item analysis, and revision by nursing instructors. A gap in literature was noted on the use of best practice with developing and evaluation of nursing examinations. The data reviewed did not have a statistical correlation between the demographic variables and the use of best practices in test construction, item analysis, and revision of multiple choice questions. This study identified current practices of nursing instructors developing, analyzing and revising multiple choice questions. Nursing educators can use the information to help develop plans for consistent grading practices in the future and prevent future grade inflation.

DEDICATION

This is dedicated to those who have been a constant source of support and encouragement during the challenges of my doctoral program. I thank my mom and step-father, Jane and L. Bruce Clark, for their support during these hard years through my personal and educational trials and tribulations. I appreciate the support you provided my children, I am forever grateful. My mom, my personal editor, was always available for me to talk to or work through problems. Her values and belief in education has been the foundation that grew into a love of learning. My Aunt MaryLou, Uncle Ed, Aunt Joan and Uncle Dick for encouragement and support during the summers so I could have uninterrupted writing time. Thank you for feeding me when I would forget the time.

To my friends; thank you for your understanding and encouragement and many, many moments of crisis. Your friendship means so much to me. Thank you for the venting and ranting sessions, along with the well-intended advice. Jan Real and Pat Hagan, I will always appreciate all you have done and I especially for thank you for your help and support.

The completion of this would not have been accomplished without the support of my children, Daniel, Marissa and Amanda. You expressed interest and asked me how it was going, even listening to me when you had no clue what I was talking about. When this journey began, we sat at the table and did our homework together. You let me know how proud you were of me and you were there the day I took my competency with sushi and special pictures for me to have. I love you all with my heart and you are my inspiration and the drive behind all I do.

ACKNOWLEDGMENTS

I would like to express my deep and sincere gratitude to Dr. Donna Taliaferro, my chairperson, my instructor, and faculty guide at my residency. She gave me opportunities to do research and guided me throughout my program. She believed in me and supported, encouraged, and motivated me to perform the research in nursing education. She made me think and consider issues from different angles. It was a great honor to work and learn from her. She believed in me through this process and I am very thankful for her assistance.

I want to thank Dr. Anne Brett and Dr. Gail Williams for their assistance and recommendations on my research work. I valued their support as committee members. They were willing to work with me researching nursing education. I appreciate their assistance, insightful comments, and hard questions. I extend my sincere appreciation for the learning opportunities provided to me by my committee.

I also would like to acknowledge Paula Williams who took the time to sit down and discuss my educational direction. Her encouragement was appreciated.

My sister and brother in law, Deborah and John Rice, I appreciate your support and encouragement. John Kuchenbrod for being there while I was waiting for my competency results. You were very supportive, and you helped me connect the "statistical" dots. Thank you, Andrew Droutman, for your support and understanding in the beginning of my educational career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Nursing education and accompanying evaluation of educational strategies have evolved significantly from the early 20th century to today. The academic performance of students is determined by formative and summative evaluation (Talebi, Ghaffari, Eskandarzadeh, & Oskoue, 2013). Formative evaluations are administered as a classroom assessment, which is defined as any classroom activity during the course that provides information about student learning (Ferrara, 2014). Summative evaluations help determine student learning at the end of the course (Quinn & Novotny, 2012). O'Halloran and Gordon (2014) suggested that assessments are influenced by many factors that result in assessment practices that do not always evaluate a student's understanding of course material. A study by Bowen, Grant, and Schenarts (2015) suggested that unclear grading policies lead to increasing grades or grade inflation, where grades do not coincide with the abilities of the students. The findings in studies by Bowen, Grant, and Schenarts (2015), Reynolds (2015), and Sowbel (2011) caused concern that new graduates may not be competent as nurses on the patient care units and patient safety, therefore, may be at risk.

In order to properly evaluate learning, assessments should be developed based on the course learning objectives. A test plan, or test blueprint, is used to identify the content for assessments (Oermann & Gaberson, 2014; Quinn & Novotny, 2012). Instructors can identify the specific areas of a course where content needs greater emphasis or clarity when reviewing the assessment data (Khoshaim & Rashid, 2016). In nursing, multiple choice questions (MCQs) are used to assess student learning and questions are designed

to assess students' critical thinking skills (Zaidi, Grob, Monrad, Kurtz, Tai, Ahmed, Gruppen, & Santen, 2018). Multiple choice questions (MCQ) are developed by instructors or some instructors will use test banks developed by nursing book publishers. In a recent study, 73% of respondents modified or used MCQ items from the textbook or test bank (Bristol, Nelson, Sherrill, & Wangerin, 2018). However, MCQs are chosen to use, higher level cognitive questions must be used for adequately assessment of student abilities. Caution must be taken when using test bank questions as the difficulty and discrimination information from item analysis is not available and the quality of the question could be poor (Oermann & Gaberson, 2014). Students can acquire access to these test banks online from internet sources, which compromises the results of the examination (Madara, Resha, Krol, Lacey, Martin, O'Sullivan, & Smith, 2017).

Some of the questions that arise when reviewing the literature include how nursing instructors handle grading multiple choice exam questions, what percentage of instructors use evidence-based testing practices, how do instructors determine the worthiness of a multiple choice exam question, how do instructors respond when a test question is deemed to be poor, and what are the actions on how they choose to grade exam questions (Killingsworth, Kimble, & Sudia, 2015). The way instructors grade exam questions has consequences that affect whether students obtain the grades needed to continue or graduate from the program. For example, once an exam question is deemed to be poor, instructors have the following options: (1) full points can be given to all students for the question, (2) full points can be given to those who got the question correct, while those who got the question wrong can be given partial points, or (3) the poor question grades can be maintained, unchanged (Phelps, McDonough, Parker, & Finks, 2013).

Options 1 and 2 can lead to grade inflation, meaning the grades students receive are higher than earned, given their understanding of the subject being tested (Phelps et al., 2013). Docherty and Dieckmann (2015) reported that grade inflation, which can be related to the evaluation of test questions, raises the concern that students are not passing a course because of their actual knowledge, but rather because grade inflation pulls the scores to a passing level. Grade inflation related to test question evaluation may misrepresent the students' true scores making it challenging for proper assessment of an individual student's progress and assessment of the class as a whole (Oermann & Gaberson, 2014).

Chapter 1 reviews the background of reviewing multiple choice questions, grade inflation, and the ethical question raised by grading practices. The problem, purpose, and significance of the research study is presented. The nature of the study and research questions are stated and the terms used in the study will be defined. Finally, the limitations, delimitations, and scope of practice of the study will be identified.

Background of the Problem

The National League for Nursing's (NLN) research priorities for 2016 included studies in teaching and practice that focus on behavior and the use of ethical codes. The NLN Fair Testing Guidelines for Nursing Education's first general guideline focuses on testing that is supported by evidence and is fair to all test takers (National League for Nursing, 2002; Oermann & Gaberson, 2014). Then scoring and results of an exam can be assessed; however, this evaluation needs to be consistently performed (McDonald, 2014). Multiple studies reported not all nursing instructors use evidence-based practice and test analysis when scoring exams (Killingsworth, Kimble, and Sudia, 2015; Oermann et al.,

2009; Bristol et al., 2018). Oermann and colleagues (2009) reported the most important thing nursing instructors look at is the pass rates of the NCLEX®. More concerning is that half of the faculty surveyed had not considered researching evidence-based practice for testing practices. Killingsworth (2013) and later, Killingsworth and fellow researchers (2015), reviewed decisions about best practices in constructing, analyzing, and revising tests and reported the use of the NCLEX® test plan and peer review of test items as components of test development used less frequently.

Educational Background

Many nursing instructors are clinical practice experts but lack expertise in curriculum development and assessment of learning (Bristol et al., 2018). Studies have been done showing many test items are not appropriate due to writing flaws, poorly written questions, or inappropriate difficulty levels (Baig, Ali, Ali, & Huda, 2014; Billings & Halstead, 2016; Gajjar, Sharma, Kumar, & Rana, 2014). Exam questions and test development procedures are frequently passed down from senior to junior instructors and often do not follow evidence-based practice. Without education and training, item writers develop low-quality test items that test Blooms remembering or understanding levels of thinking (Tarrant & Ware, 2012). Many test items do not encourage students to apply the knowledge that prepares them for competent practice. Without understanding educational strategies, instructors use nursing educational practices based on tradition rather than educational theories of learning, assessment, and evaluation (Bristol et al., 2018). Clinical expertise does not prepare nursing instructors for the classroom as education and nursing are different theoretically (Booth et al., 2016). Lack of educational strategies results in difficulty formulating assessment items. Nursing programs primarily

use multiple choice questions and require the necessary scholastic knowledge and training to develop high-quality questions (Tarrant & Ware, 2012). Furthermore, nursing instructors have been found to teach to the test, providing the education needed to be successful in examinations (Tarrant & Ware, 2012).

Ethical Practice

Standards are set at individual institutions of higher learning for passing academic evaluations and clinical practice (Docherty & Dieckmann, 2015). These standards should be built on evidence-based practice and ethical fairness. O'Flynn-Magee and Clauson (2013) found the commitment to fair grading at many institutions was strong, but even at these institutions there are inconsistent grading practices. Consistency in grading practices is needed for ethical practices and fairness (National League for Nursing, 2012). Nursing instructors have an ethical and professional responsibility to assess student learning fairly and in a reliable manner, as this is frequently the only means to determine competency.

Salminen and colleagues (2017) reported instructor ethics includes respect for the student's privacy, treatment of the students equally, and accept responsibility for assessing the student learning outcomes. Fair and honest assessment of students' learning is an important job for educators. Pazargadi, Ashktorab, and Khosravi (2012) noted that students felt there were inconsistencies between instructors and evaluations given by the same instructor at different times. In the study by Salminen and fellow researchers (2017). students reported injustice in student assessments with inconsistencies noted in grading from one student to another. Evaluation of student learning should focus on determining the level of understanding learned along with the ability to apply that

knowledge. Grade inflation related to test question evaluation may misrepresent the students' true scores making it challenging for proper assessment of an individual student's progress and assessment of the class as a whole (Oermann & Gaberson, 2014). Researchers point at deficiencies in nursing instructors' education in grading practices as a major cause for this issue (Docherty & Dieckmann, 2015; Salminen et al., 2013).

Developing Assessments

When considering assessment development processes, nursing instructors must understand their ethical obligations to nursing students and the nursing professional standards. Learning assessments should be based on evidence-based practices. Oermann and Gaberson (2014) emphasized that each step of test development requires the nursing instructors to make a decision based upon the purpose of the test and the population taking the classroom test. In the past, assessments of nursing students were based on patient care interactions and basic skills. With advances in scientific knowledge and technology, nursing practice has progressed from a bedside technical position to a profession with higher level of education. Testing of nursing students must incorporate practice and ability to critically think about situations.

Formative and summative assessments should be used to understand the student's level of understanding. A formative assessment is used during the course to measure student learning and the summative occurs at the end of the course, to evaluate overall learning at end of course. Learning assessments should be developed based on the course learning objectives. A test plan can be used to map out the course and assessment needs. Tarrant and Ware (2012) described a test blueprint as a tool that looks at the course objectives and content to determine the number of test questions from each content area.

It is a process that is crucial to accurately determine learning assessment needs (Oermann & Gaberson, 2014).

Once a blueprint is developed, instructors have a guide to creating the formative and summative assessments (Oermann & Gaberson, 2014). The majority of assessments in nursing use multiple choice questions. MCQs are developed using a stem, which is the body of the question, followed by potential answers. The possible answers are called distractors. Distractors should be credible, but not fully correct answers (Kaur, Singla, & Mahajan, 2016). Questions should be written at an appropriate cognitive level. Using Bloom's taxonomy (Agarwal, 2019), there are six levels starting with remember and understand. The next step is application and analysis. The final levels are evaluate and create. Nursing exam questions are recommended to be written at the application and analysis level. This provides information that shows instructors that the students understand the content and can use it to form decisions and apply to nursing practice.

Multiple choice questions that assess students' abilities to apply and analyze course material are not easy to develop. There is little literature in nursing regarding the format, structure, validity, and reliability of MCQs. Most literature found on MCQs appears in medical education, psychometric testing, and psychology literature. Once an examination is administered, nursing instructors can perform an item analysis. An item analysis provides statistical data on each test item for item difficulty, discrimination, reliability, standard deviation, and distribution of test takers' responses (International Test Commission, 2014). The item analysis should be reviewed for issues with test items such as correct wording, understandability, and consistency in students' choice of distractors. If an issue is found with an item, instructors must decide how to score that

item. Some instructors will remove a MCQ item if 50% of the students get the question wrong; however, if the question was answered correctly by the top scoring students, the question is considered a good question (Sagendorf, 2013).

Decisions regarding what to do with poorly performing test questions affect scores and course grades (Phelps, McDonough, Parker, & Finks, 2013). Instructors have several choices when an item analysis reveals a poor question. The examination can be scored using multiple correct answers, keeping the question, or eliminating the question. When scoring learning assessments, professional standards and fairness must be considered. When rescoring, the instructors risk inflating student grades. Nursing students need to learn the foundation of knowledge, skills, behavior, and attitudes relevant to practice and patient safety. When grade inflation occurs, students' grades show a higher knowledge base and a higher ability to apply the knowledge learned. Faulty must be aware of the danger of grade inflation and the possibility of passing students that have not gained the knowledge needed to practice competent safe care. Instructors have an obligation to the public to graduate safe practitioners.

An increase in public demand for accountability of educational outcomes and newer federal government regulations require educational institutions to provide evidence that learning outcomes are being met (Billings & Halstead, 2016). Nursing program evaluations for program accreditation and the state board of nursing standards use the first time pass rates of the National Council Licensure Examination (NCLEX®) as a measurement. The NCLEX® exam is a licensure exam that measures basic competency of individuals graduating from a nursing program. The licensure examination for entry-level graduate nurses was developed to ensure public safety (National Council of State Boards

of Nursing, 2016). Nursing programs are under pressure from accreditation agencies to have NCLEX® pass rates within the national average range (Bristol et al., 2018). This has resulted in high stakes testing to assure that students can practice safely upon graduation. A high stakes test is any examination used for tracking or determining promotion or graduation (Tagher & Robinson, 2016).

Nursing programs are using high-stakes standardized tests as a basis for progression or graduation from their program. Nursing students face the potential loss of time and investment in their education if they do not earn target scores (Tagher & Robinson, 2016). The NLN does not support the practice of using high stakes standardized tests as a basis for program progression or graduation (Tagher & Robinson, 2016). The NLN explained their position by writing *The Fair Testing Imperative in Nursing Education* in 2012 to guide nursing programs in making more balanced decisions regarding nursing student competence (Sullivan, 2014). Another concern with high stakes testing arises when the focus of teaching/ learning is moved primarily to the content of the examination. Students focus on preparing for an examination and the acquisition of life nursing practice skills decreases (Kumandas & Kutlu, 2015), when nursing instructors focus instructional activities on lower cognitive skills based on testing results (Kumandas & Kutlu, 2015). Instructors can fall into the practice of teaching to the test, which does not support critical thinking skills.

Teaching to the test and focusing on memorization for high stakes testing does not produce graduate nurses ready for practice. Nursing students need to develop problem-solving critical thinking skills to provide safe and high quality care (Günüsen, Serçekus, & Edeer, 2014). Properly developed MCQs can determine if students can assess, apply

and evaluate information (Bauer, Holzer, Kopp, & Fischer, 2011). Baig and colleagues (2014) and Bush (2105) reported well written MCQs properly test the student's ability to apply knowledge, comprehension, application, and analysis. Nursing instructors need to review test item analysis to determine how well each test question performed in a testing situation. Strong testing MCQs can then be used to accurately assess student learning.

Problem Statement

While teaching undergraduate studies, nursing instructors are charged with determining if the student has mastered the material and can apply the information that has been learned (National Advisory Council on Nurse Education and Practice, 2010). There are various learning assessment tools, but the commonly used test comprises multiple choice questions (Bristol et al., 2018; Oermann & Gaberson, 2014). Many nursing instructors do not evaluate poor multiple choice questions in a consistent manner. This causes an ethical dilemma with inconsistencies in grading and potential grade inflation. To address this dilemma instructors can review the scores of the cohort and remove a poorly written question from an exam. Test item analysis can be used to determine if the question is poorly written or has poor distractors. An item analysis uses the data regarding how questions were answered, which distractors were chosen, and which student groupings answered the question correctly. This provides evidence for instructors to make a decision on the validity of a question. Research by Killingsworth (2013), Killingsworth, Kimble, and Sudia (2015), and Reynolds (2015) confirmed that test item analysis is not performed by all nursing instructors.

Nursing instructors are educated as nurses not always including educational pedagogy, so many are not prepared to develop higher level thinking assessments (Bristol

et al., 2018). Advanced academic preparation, many times, focuses on the clinical area of practice rather than gaining knowledge into evidence-based research and practice, teaching methods, and curriculum design and development that are encompassed in the foundation of academic practice (Booth, Emerson, Hackney, & Souter, 2016). The lack of understanding of assessment practice has translated into inconsistency in grading practices, which in turn, causes grade inflation (Bristol et al., 2018; Phelps et al., 2013). Students that pass courses due to grade inflation might not possess the ability to provide safe quality care. Grade inflation poses a safety threat to patients of nursing graduates who may be less competent than their academic records suggest. As educators, the primary goal is to have graduates prepared and competent to provide such care.

The specific problem the study addressed was how best practice in classroom test construction, item analysis, and revision used by nursing instructors. The research looked at correlations between factors in instructor demographic and teaching background. In Oermann, Saewert, Charaika, and Yarbrough's study (2009), only half of respondents looked at research when performing classroom test construction, item analysis, and revisions. More concerning is that half of those surveyed had not considered researching evidence-based practice for testing practices (Oerman et al., 2009). Bristol and colleagues (2018) research revealed inconsistencies in testing practices and a lack of evidenced-based standards in test development. A review of testing practices will provide insight into how nursing instructors are making decisions about testing and grading practices.

Purpose of the Study

The purpose of the correlational study was to examine the relationship between best practice and the reality of practice in classroom test construction, item analysis, and

revision in nursing programs in the United States. The study provided information exploring instructor's use of best practice when composing evaluations and grading multiple choice questions using the Best Practices in Test Development Instrument (Killingsworth, 2013). The instrument identifies information about participants' demographic data and teaching background. Correlations in various demographic data present insight into factors relating to different grading practice and the use of best practice.

Population and Sample

In research studies, the population refers to the group of subjects with characteristics the researcher wants to study (Boswell & Cannon, 2014). The population for this study was nurse educators teaching in undergraduate registered nursing programs in the United States. The lists of registered nursing programs in the United States was obtained from individual state's board of nursing web sites and The American Association of Colleges of Nursing. Email addresses for the deans and directors of programs were compiled from colleges, universities, and professional schools' websites.

For a research study, a sample of the population can be used to represent all members. A predetermined number, or sample, of research subjects can be used to provide information concerning that population (Malone, Nicholl, & Coyne, 2016). The data obtained from a sample of participants selected from the larger population can be examined and inferences can be made about the entire population (Hayat, 2013). Hayat (2013) reported determining the appropriate size for a sample of the population is an important consideration. A sample size too small does not have sufficient power to statistically detect relationships, while samples too large could be considered unethical,

wasteful of resources, or make the study impractical to conduct (Malone, Nicholl, & Coyne, 2016). An appropriate sample size provides information needed to make a statistical judgement about the results. The sample size determined for this study was 382 participants.

Significance of the Study

The research results can provide insight into the use of best practices and the prevalence of the inconsistencies in test construction, item analysis, and revision by nursing instructors, specifically focusing on assessments using multiple choice questions. A review of current literature has shown a gap in knowledge on the use of evidence-based practice and the development and evaluation of nursing examinations (Bristol et al., 2018; Killington, 2013). Information collected in the study can provide evidence regarding how instructors grade formative tests. The data collected provided information on whether different practices are common to educational variances in instructors' own educational levels, as well as whether they are teaching in associate degree programs and/or baccalaureate programs. The data also offers information about instructor use of best practice in test construction, item analysis, and revision of multiple choice questions. The information gathered will identify current practices and help educators formulate plans for consistent ethical grading practices in the future and prevent future grade inflation.

Nature of Study

A quantitative approach was used to investigate data from nursing instructors regarding the use of best practices and the prevalence of the inconsistencies in test construction, item analysis, and revision by nursing instructors. Quantitative methods

review large quantities of facts and allow generalizations based on statistical analysis (Roberts, 2010). The Best Practices in Test Development Instrument used to collect data from nursing instructors (Killingsworth, 2013). The population for the study includes nursing instructors from registered nursing educational institutions located in the continental United States. A power analysis was performed using the 2016 data from the Bureau of Labor Statistics to determine the number of participants needed for study. Data were analyzed and compared for relationships. Pearson correlations were used to determine relationships between instructors and the use of best practices when constructing test items, item analysis, and revisions. Pearson correlations is used to determine relationship and strength of the relationship between two items (Vogt, 2007).

Research Questions/Hypotheses

There are no clear common guidelines in nursing education to guide test construction, item analysis, and test item revision There is a gap in the knowledge of how nursing instructors make decisions regarding the use of multiple choice questions (Bristol et al., 2018; Killington, 2013; Killington et al., 2015).

Research Question 1

What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions?

Null hypothesis: There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

Alternative hypothesis: There is a relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

Research Question 2

What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

Null hypothesis: There is no relationship between factors in nursing instructor using best practices in test analysis, and the educational preparation of the educator?

Alternative hypothesis: There is a relationship between factors in nursing instructor demographics and educational background and nursing instructors' use of best practice in test construction, test analysis and test revisions.

Theoretical Framework

The theoretical framework used to guide this study was the Constructivism Theory, focusing on how individuals acquire knowledge and learn (Bada, 2015). The approach to classroom assessments is based on tests items designed by teachers that are reviewed and revised according to test analyses performed (Graue, 1993). Nursing student assessments are developed to determine if knowledge has been acquired. Nursing instructors develop test questions to evaluate student learning. Once the test items are administered, the items are analyzed for appropriate distractors and for effectiveness. Nursing instructors can take the information from the analysis and make necessary changes to improve test items. This method allows instructors to learn and further develop test items.

The study is important to identify the use of best practice when developing assessment for undergraduate nursing students. Nursing students are evaluated in the educational programs. Nursing programs need to be able to accurately assess students to ensure safe and competent practice. The issue of unsafe nursing students progressing

through nursing education is problematic in pedagogical practice as well as a professional issue which could result in a decrease in the public's trust in nurses (Paskausky & Simonelli, 2014). In addition, grade inflation is a current problem in higher education. Nursing programs need accurate assessments to ensure students are competent to enter the nursing field upon graduation.

Definition of Terms

Multiple choice question is a two part question that has a stem, which is the information and question being asked followed by a number of possible answers (Bailey, Mossey, Moroso, Cloutier, & Love, 2012).

Item analysis is a statistical evaluation of test items that typically includes item difficulty index, item discrimination, test reliability, and mean test scores (Tarrant & Ware, 2012).

Distractors are plausible alternative answers that are not correct and ideally should appear similar in grammar, length, and complexity as the correct answer (Begum, 2012).

Distractor discrimination is the difference of the distractor chosen by high-achieving and low-achieving students (Tarrant & Ware, 2012).

National Council Licensure Examination (NCLEX®) is a test that evaluates graduates from nursing schools for competence to practice nursing (Sullivan, 2014).

High stakes testing is any examination used for tracking or determining promotion or graduation (Tagher & Robinson, 2016).

Test blueprint is a grid or table that maps the course objectives and provides an outline of the number of test questions that should be given for each content and cognitive level (Tarrant & Ware, 2012).

Code of Ethics is defined as the behaviors expected, built on mores of the culture, education, and religion (Yildiz, Icli, & Gegez, 2013).

Bloom's Taxonomy describes levels of learning arranged from basic cognitive processes to more complex processes of critical thinking (Agarwal, 2019).

Formative evaluation/assessments are assessment or activities that provide information about student learning (Ferrara, 2014).

Summative evaluation is an evaluation performed at the end of a course to determine the knowledge, values, and skills achieved during a course (McDonald, 2014).

Grade inflation is when student grades do not match their performance (Paskausky & Simonelli, 2014).

Assumptions, Limitations, Delimitations

Assumptions

Assumptions are factors that are accepted as being true. For the purpose of this study, the assumptions are that those filling out the surveys are reporting accurate information on the use of best practice in test construction, test analysis and test revisions. The respondents will also provide accurate information for all other questions in the survey.

Limitations

Limitations are restrictive conditions or weaknesses causing factors that cannot be controlled (Locke, Spirduso, & Silverman, 2014). This study was performed using The

Best Practices in Test Development Instrument (Killingsworth, 2013). The results are based on self-reported use and therefore could be reported as more or less than are actually used in practice. Vogt and colleagues (2017) reported individuals who respond to survey research provide truthful responses. Those who respond could have a higher interest in the use of best practice in test construction, item analysis, and revision which could skew the responses. Response rates for different instructor backgrounds might not be evenly distributed. Another limitation of this study is the correlational design. Using this method only detects possible associations and cannot link associations as a cause (Vogt et al., 2017). Since this study is identifying potential relationships between various nursing instructor factors, further studies can be performed with the relationships identified. A final consideration is data that is an outlier. An outlier is a data point that falls substantially above or below the others. This can change the direction and strength of the correlation (Privitera, 2017). If this occurs, different statistical analysis may be needed.

Delimitations

Delimitations define the parameters of the study and the populations from which generalized study results can be inferred (Locke, Spirduso, & Silverman, 2014). This study will examine grading practices of instructors teaching registered nursing students at colleges, universities, professional schools and junior colleges in the United States. General medical and surgical hospitals and technical and trade schools will not be included in the study. Results of the study provided data on the extent that nursing instructors use best practices when developing, analyzing and rewriting exam questions.

Chapter Summary

This study investigated how nursing instructors handle grading multiple choice questions. This study examined how nursing instructors determine the worthiness of a multiple choice exam question and the percentage of instructors using evidence-based testing practices. Various practices to assess student learning have been used. Multiple choice questions are used to assess students' abilities to apply and analyze course material, but these questions are not easy to develop. Most MCQs in exams are knowledge based and do not reach higher level assessment. After an exam is given, instructors typically review an analysis and look for poorly performing items. Those items can be removed, left as is, or more than one response can be accepted. This practice can lead to grade inflation. Grade inflation and poor assessment questions lead to questions regarding whether graduates are prepared to be safe practitioners of care. There are ethical issues surrounding consistency and fairness in the grading practices of nursing instructors. A survey was administered and the data were analyzed to determine if instructors use evidence-based practices when reviewing and grading multiple choice question exams. This data can be compared to demographic data to look for any correlations in exam evaluation practices with the educational levels of instructors and the level of the program where the instructors teach.

Chapter 2 will review the current literature on nursing education assessments, use of evidence-based practices and ethical issues surrounding grading practices. Grading practices will be discussed. Grade inflation, codes for testing practice, grading consistencies, high stakes testing, and nursing instructors' education will be reported. Multiple choice questions will be discussed. Test item analysis, poor performing multiple

choice questions, and grading ethics will be presented. The theoretical framework and significance of this study will be offered.

CHAPTER 2

Literature Review

Chapter 2 reviews the current literature on nursing education assessments, on the use of evidence-based practices, and on ethical issues surrounding grading practices. First, the process of the literature search is described. Articles on grade inflation, codes for testing practice, grading consistencies, high stakes testing, and nursing instructors' education are discussed. Multiple choice questions along with test item analysis, poor performing multiple choice questions and grading ethics are presented. The theoretical framework and significance of the study are offered. Finally, chapter 2 will conclude with overall observations from the literature.

Title Searches and Documentation

Chapter 2 addresses the literature relevant to the research questions, historical and current literature on nursing education assessments, and gaps in the literature. Literature was retrieved through the University of Phoenix Library databases, EBSCOhost, Proquest, Medline, Sage, and internet search engine Google Scholar for contribution of information, peer-reviewed journal articles, and books. The internet links from stand-alone websites such as the National League for Nursing, American Educational Research Association, American Psychological Association, and National Council on Measurement in Education provide historical and current practices.

Table 1
Summary of Literature Search Results

	Peer reviewed articles	Books	Dissertations	Edited texts	Stand alone websites
Multiple choice questions	11				
Grading practices	9	2		1	
Grade inflation	5				
Item analysis	3				
Fair testing			1		2
Grading ethics	5				
Nursing instructor education	4				

Historical Context

A review of the development of nursing educational practice can help to provide a better understanding of the current educational practices. Looking back at the 1900s, nurses worked on the patient care units and had on the job training (Roux & Halstead, 2009). Nursing education moved to being taught at universities and focused education on nursing theory and patient centered care following the 1923 Goldmark Report (Goldmark, 1923). Advances in science and technology required advanced knowledge and skill to care for complex treatments and more critically ill patients (Roux & Halstead). By the late 1920s, 25 nursing programs were established in university settings (Keating, 2014).

The change in student preparation for practice had several factors affecting it. First, the profession was growing and developing evidence-based practice. Diploma nursing programs started pairing with colleges and universities awarding graduates a bachelor's degree. This resulted in nursing students with a well-rounded education (Tobbell, 2014). Docherty and Dieckmann (2015) wrote that a combination of

knowledge, technical skills, and ethical conduct is key to preparing students for professional practice.

Secondly, the modification in focus and location of nursing programs affected the length of time in the clinical setting. Diploma programs were typically based in a hospital and the students worked on the units while learning in classroom and hospital patient units. Evaluation of nursing student learning was based on patient care interactions and demonstration of basic skills. The programs at educational institutions are based on credit hours including didactic and clinical hours. Students went with an instructor to healthcare facilities to learn the clinical portion of nursing for a certain number of hours, fewer than the hours a diploma nurse spent in the clinical setting. Today, evaluation of nursing students has a small component of patient care interactions and basic skill, but more emphasis is placed on critical thinking skills. University nursing programs included physical and biological sciences, social science, communication skills, and general education along with nursing content (Keating, 2014).

Reduction in clinical time and increased knowledge base needs have resulted in the changing focus of educational pedagogies and evaluation practices in nursing. Nursing education programs need to produce graduates ready to seamlessly step into practice, therefore, education must focus on developing critical thinking and critical reasoning to instill the competencies necessary for new graduate nurses (Theisen & Sandau, 2013). The National Council Licensure Examination (NCLEX®) was developed to test nursing students on basic nursing education competency to enter practice, and applicants are required to demonstrate critical thinking, reflection and problem solving skills (Roa, Shipman, Hooten, & Carter, 2011). Nursing instructors have tried to emulate

the type of testing given in the NCLEX® to prepare students for their licensure exam. The NCLEX® is designed to evaluate students to determine if graduates have a competence level that is safe to practice nursing. With reduced clinical time, nursing instructors must find other ways to help students develop critical thinking skills so they may be safe competent practitioners. Multiple choice exams are the main method nursing instructors employ to determine students' knowledge levels and ability to critically analyze clinical situations. In order to use this type of questions in exams, nursing instructors need to use best practices when developing questions. Unfortunately, the use of best practice in nursing education has only recently begun to receive attention (Booth et al., 2016).

Evaluation of Learning and Testing

Nursing education and accompanying evaluation of educational strategies have developed significantly from the early 20th century to today. The academic performance of students is determined by formative and summative evaluations (Talebi et al., 2013) often taking place as a classroom assessment, which is defined as any classroom activity that provides information about student learning (Ferrara, 2014). O'Halloran and Gordon (2014) wrote that assessments are influenced by numerous aspects that result in varying assessment practices with loose connections to student understanding of course material. In order to properly test learning, assessments should be developed based on the course learning objectives. A test plan, or test blueprint, is used to identify the content for assessments (Quinn & Novotny, 2012). Instructors can identify the specific areas of the course where content needs greater emphasis or clarity when reviewing the assessment data (Khoshaim & Rashid, 2016).

Assessments identify any student learning gaps and evaluate if students have achieved course outcomes (Quinn & Novotny, 2012). Assessments of nursing students should reflect high standards, which should be clear to students and the nursing instructors who are facilitating student learning (Smith & Fleisher, 2011). Nursing programs use multiple choice questions (MCQ) as they can be reliable, easy to administer to small or large groups, and are cost efficient (Begum, 2012; Davey et al., 2015). MCQ determine if the student has the ability to assess, apply and evaluate information (Bauer et al., 2011). Begum (2012) stated MCQs can evaluate several items in relation to a single topic. Students can be identified with strong or weak abilities according to their responses

Nursing programs use MCQs for formative, summative and standardized high stakes testing. The National Council on Measurement in Education (2011) defines this as a test whose results have consequences that affect both examinees and institutions. Multiple choice questions were developed by Edward Thorndike and were first used at the Kansas State Normal School in 1914 (Siddiqui et al., 2016). MCQs are developed using a stem of the question and then providing possible answers called distractor options (Begum, 2012; Maher, Barzegar, & Ghasempour, 2016). The question should focus on a concept from the course (Begum, 2012). The distractors need to be credible, not incorrect, but should not be too close to the correct answer (Kaur, Singla, & Mahajan, 2016). Begum reported the most time consuming and difficult part of developing questions are figuring out the appropriate distractors.

Well written MCQs measure knowledge, comprehension, application, and analysis (Baig et al., 2014; Bush, 2015; Kaur et al., 2016). MCQ have the advantage of providing rapid feedback and the ability to analyze test results (Bauer et al., 2011; Davey

et al., 2015; Romero et al., 2013). Baig and colleagues attributed their frequent use to a higher reliability, higher validity, higher ease of administration and scoring. Maher, Barzegar, and Ghasempour (2016) reported MCQs are objective tests that can standardize questions that are not as easy to guess at answers. Multiple choice questions are used by nursing programs for testing and to measure both formative and summative learning. Some programs have high stakes standardized exams at the end of each semester while others administer them at the end of the program.

Unfortunately, multiple choice questions are difficult to develop and depend on the use of distractors (Begum, 2012). Analyses of the exam results are needed to evaluate the reliability of the questions. Weak MCQs impede the interpretation of test scores, which negatively impacts student pass rate (Bauer et al., 2011). New computer programing allows for test scoring that looks at discriminators and other analysis of exam data (O'Halloran & Gordon, 2014). These data are only useful if a proper analysis is performed. Not all nursing instructors utilize the tools available to determine fair test questions therefore, not employing ethical testing practice.

An item analysis should be performed on MCQ test results (Baig et al., 2014). Item analysis is the assessment of the quality of test questions using the information from students' responses (Khoshaim & Rashid, 2016). O'Halloran and Gordon (2014) reported that item analysis could help determine the reliability of the test and generate a better analysis to determine different scholastic abilities among students. Item analysis can improve assessments and teaching methods by allowing nursing instructors to focus on weaker areas noted from the results (Talebi et al., 2013). Item analysis is used to determine validity and reliability of test questions. According to NLN Fair Testing

Guidelines (2012) nursing instructors have an ethical obligation to ensure testing is fair to all and supported by evidence. There is an absence of clear guidelines on testing which may lead to variations of grading (Reynolds, 2015).

Current Content

Grading standards are set at individual institutions for passing academic assignments and clinical practice (Docherty & Dieckmann, 2015). Billings and Halstead (2016) noted that nursing instructors are responsible for the evaluation of students. Assessment of student learning occurs in the clinical and didactic settings. The International Test Commission (2014) stated that guidelines should be used to increase the efficiency, precision, and accuracy of the scoring and analysis of tests. There are multiple guidelines for test development, including test blueprints, writing test items, and evaluating test-item (Hicks, 2011; Oermann & Gaberson, 2014). However, Killingsworth, Kimble, and Sudia (2015) found minimal research on the construction, analysis, and revision of classroom assessments. Booth and associates (2016) reported that given the significant role of nursing in today's health care system, nursing instructors should use best practices when preparing nurses for entry into practice.

The academic performance of students is determined by formative and summative assessments (Talebi et al., 2013). A classroom assessment is defined as any assessment activity that provides information about student learning (Ferrara, 2014). O'Halloran and Gordon (2014) wrote that assessments are influenced by numerous aspects that result in varying assessment practices with only loose connections to student understanding of course material. To properly test learning, assessments should be developed based on the course learning objectives.

Assessing the student's achievement at the end of the course, or summative assessment is another goal of testing (Quinn & Novotny, 2012). O'Flynn-Magee and Clauson (2013) found when nursing instructors are making decisions about assessments, the most important consideration is the program's rates of students passing the licensure examination In the United States, the National Council of States Boards of Nursing licensure examination (NCLEX®) is designed to measure the ability to perform safe and effective practice (Docherty & Dieckmann, 2015). Course learning objectives and nursing program outcomes are focused on educating safe practitioners. Nursing program curricula should be academically rigorous, and nursing instructors must uphold academic standards (Billings & Halstead, 2016). Reynolds (2015) reviewed the literature and found that teaching in public versus private institutions and community versus four-year colleges may influence grading practices, but did not find any studies that discussed grading practice differences between associate degree and baccalaureate degree nursing programs. Research on testing has focused on grade inflation, grading consistencies, quality of MCQs, and test item analysis.

Grading Practices

The NLN Fair Testing Guidelines in 2012 stated nursing instructors have an ethical responsibility to develop tests based on evidence-based practices, consistent in all courses, and fair to all students. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education Standards (2011) released a Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices as a guide for ensuring fairness in all phases of testing; development of tests, administering and grading of tests, reporting

and analyzing of test results, and notifying test takers. There are multiple guidelines on test development, including test blueprints, writing test items, and evaluating test-items (Hicks, 2011; Oermann & Gaberson, 2014). Guidelines provide consistency and ensure high quality practice between empirical evidence and actual professional practice (Zumbo & Chan, 2014).

Oermann, Saewert, Ika, and Yarbrough (2009) analyzed the data from The Evaluation of Learning Advisory Council of the National League for Nursing survey of nurse educators. The purpose of the survey was to gather information on nursing instructors' evaluation of student learning and factors that influence their decisions about assessment and grading their students. The research tool was piloted with a small group of 15 nursing instructors. The actual survey included 1573 full and part-time registered nurse and master level nursing instructors. Of these participants, 72% held master's degrees, 12% were certified nurse educators, and 57% had taught more than ten years.

According to the survey results, 83% reported traditional and past practice guided how they assessed and graded students (Oermann et al., 2009). The most important factor considered in grading was the pass rates of the NCLEX®. Educational soundness of the assessment and educational standards were additional factors considered. Nursing instructors indicated that time was a factor in the decisions about assessment techniques. Half of the respondents had not considered researching evidence-based practice for testing practices. Limitations of the study included a lack of demographic information, inability to determine the response rate, and the reliability of the survey tool.

Oermann, Saewert, Ika, and Yarbrough (2009) analyzed the data from The Evaluation of Learning Advisory Council of the National League for Nursing survey of

nurse educators. A survey tool was developed using modified versions of Victor and Cullen's (1988) Ethical Climate Questionnaire and Oermann and colleagues' (2009) Evaluation and Testing in Nursing Education. The survey tool (Killingsworth, 2013) asked participants to rate best practice activities with test construction, test item analysis and test revision using a scale of one to seven, one being not at all and seven being all the time. Test construction components consisted of 12 components; course objectives, class or unit objectives, major content topics, specific content topics, test blueprint, NCLEX® test plan, peer review of test items, higher cognitive levels according to Bloom's, taxonomy, clinical context for test items, plausible distractors in multiple-choice test items, even distribution of correct answer in multiple-choice options, and use of various test item types. Test item analysis asked nursing instructors to identify if the information was obtained after the test was administered; the number of students who answered each question incorrectly, difficulty level, discrimination index, the frequency of distractor choices with each test question, distractor discrimination, and central tendency of the student grades on the test. Test revision activities included using item analysis data when determining to keep or eliminate test questions before finalizing test scores; comparing item analysis data for test questions repeatedly used from one term to another; using distractor discrimination to revise test items, using difficulty level of test items to revise test items; assessing for linguistic/cultural bias in test items; assessing for changes in domain content based upon new research data; assessing for outdated language used in test items; changing test items to ensure test security; changing test items to reflect emphasis on classroom content; and changing test items to ensure sufficient sampling of

content. A pilot study of 34 participants was conducted and this tool had appropriate internal consistencies.

The research by Killingsworth (2013) analyzed data from the survey developed and administered to 127 nursing instructors teaching in BSN programs for at least two years who participated in classroom test construction and evaluation. When reviewing the data, it was noted that nursing instructors thought they did a good job developing tests and reported using best practices in test construction, analysis, and revisions. On a Likert scale 1 (not at all) to 7 (all the time) to indicate frequency of use, participants rated best practices in test construction 5.3 out of 7, item analysis 5.5 out of 7, and test revision 5.6 out of 7 (Killingsworth, 2013). The areas nursing instructors reported using less frequently included the use of the NCLEX® test plan, 4.8, peer review of test items, 4.2, analyzing distractor discrimination, 5.3, cultural bias, 4.5, and research-driven changes in content, 5.1 (Killingsworth, 2013). The results of the study could be affected by participants self-reporting use of best practices. Half the participants were employed at public institutions and this could affect the outcomes.

In 2015, Killingsworth, Kimble, and Sudia studied nursing instructors using best practices for constructing, analyzing, and revising tests in baccalaureate nursing programs using a descriptive correlational study. A survey tool developed by Killingsworth (2013) was used. The sample consisted of 127 participants teaching for at least two years in BSN programs from 31 different states. The mean number of years teaching was 12.9. Participants reported frequently using best practices, on a scale of one to seven, with seven being the most used; 22 of the 26 best practice descriptive were scored six or higher. Peer review of test questions scored the lowest and the second lowest was the use

of the NCLEX® test plan. The results of the research revealed that nursing instructors are using best practices in test development, analysis, and revisions.

Nursing instructors reported frequently using item analysis to determine the difficulty level of test items but were less likely to analyze the distractor discriminator (Killingsworth, Kimble, & Sudia, 2015). This was a self-reported study, meaning that participants provided answers to questionnaires and could potentially answer as to how they should be practicing rather than how they were actually practicing. Nursing instructors participating in the study might have been drawn to respond if they had an interest in test construction and evaluation techniques. This could sway the results if the participants were interested in practicing best practice and were aware of current literature on this subject.

Hughes, Mitchell, and Johnson (2016) reported on an integrative literature review of current knowledge on instructors not failing students within undergraduate nursing programs. Twenty-four articles with moderate or good methodological rigor including qualitative and quantitative research, were reviewed using the Mixed Method Appraisal Tool. The five themes identified included difficult to fail a student, failing a student is an emotional experience, instructor self-confidence is required, a student with unsafe characteristics and when failing a student, academic institutional support is needed. The nursing articles reviewed addressed how instructors not failing students occurs. The researchers noted that most articles reviewed did not identify assumptions about the topic, which increases the trustworthiness and rigor in qualitative descriptive research (Hughes, Mitchell, & Johnson, 2016). Identification of these specific areas would be helpful in developing a plan for success for these students.

The negative consequences of multiple choice questions addressed in the literature are caused by the potential formation and support of false understanding when students answer this type of question with guessing, inaccurate rationale, and incorrect knowledge development. Bailey and colleagues (2012) explored the implications of the use of MCQs in nursing in a written evaluation of current literature. The concern the authors' reported was the development of false knowledge and the consequences for the learners to later deliver care to patients. Another concern is that nursing education continues to use primarily MCQs for assessment of learning outcomes instead of using the vast variety of evaluation strategies that can measure different perspectives of learning and learning styles. Further, there is a considerable amount of time and effort needed to prepare MCQs for an accurate assessment due to exam blue printing, item writing processes and principles, and psychometric analysis of test items.

The literature confirmed that the provision of feedback used to focus students on the positive effects of MCQs can reduce the negative concerns (Bailey et al., 2012). Appropriate feedback with rationales can clarify inaccuracies and strengthen the development of knowledge. Feedback is essential to promote positive learning results; however, the type and timing of feedback remain unclear (Bailey et al., 2012). Bailey and associates (2012) reported a lack of literature addressing the negative testing effects of MCQs and the creation of false knowledge. The continued use of MCQ assessments in nursing education without addressing the negative consequences is concerning for educational and practice settings.

Grading practices of nursing instructors have several factors influencing how grading occurs. There are guidelines for general educators, but no specific grading

guidelines were found with the exception of vague statements developed by NLN on fair grading practices. Several books are available on curriculum development and contain information on grading techniques. Oermann and colleagues (2009) reported the most important thing nursing instructors look at is the pass rates of the NCLEX®. More concerning is that half of those surveyed had not considered researching evidence-based practice for testing practices. Killingsworth (2013) and later, Killingsworth and fellow researchers (2015) reviewed decisions about best practices in constructing, analyzing, and revising tests and reported the use of the NCLEX® test plan and peer review of test items as used less frequently. The use of item analysis was used, but the use of analyzing distractor discriminator was less likely to be used. The research was performed using self-report use of practices, which could have respondents answering what should be done as opposed to what is done in actual practice; and if this is the case, nursing instructors are aware of best practices that should be used. Instructors have difficulty when working with students not performing at an appropriate level. Hughes, Mitchell, and Johnson (2016) identified instructor emotions, self-confidence, and support of administration were factors in making grading decisions. A difference in grades was noted between associate degree and baccalaureate degree nursing programs, public versus private institutions, and community versus four-year colleges

Codes for Testing Practice

The National League for Nursing (NLN) research priorities included studies in teaching and practice that focus on behavior and the use of ethical codes (2016). The NLN Fair Testing Guidelines for Nursing Education's first general guideline focuses on testing that is supported by evidence and is fair to all test takers (Oermann & Gaberson,

2014). The guidelines are broadly ranged recommendations with no information for specific practices and policies for developing or administration of nursing tests. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education Standards (2011) released a Code of Fair Testing Practices in Education prepared by the Joint Committee on Testing Practices as a guide for ensuring fairness in all phases of testing; development of tests, administering and grading of tests, reporting and analyzing of test results, and notifying test takers. The International Test Commission (2014) published guidelines, which contain more information on specific processes to increase proficiency and accuracy of the scoring, analysis, and reporting of tests. These codes provide guidelines on how to prepare and administer a test, but lack clear direction on how to analyze and grade examinations. This lack of clear guidelines on how to analyze exams may lead to variations of grading (Reynolds, 2015).

Grade Inflation

Student grades have been going up over the years. Caruth and Caruth (2013) stated the shift in grading has been a move toward higher grades without a matching increase in knowledge. King-Jones and Mitchell (2012) defined this practice of higher grades without an equivalent increase in gained knowledge as grade inflation and attributed the beginning of this practice to when lower grades could cause students to be drafted during the Vietnam War. Grade point averages have increased by 0.6 from 1967 to 2000 with the average at private schools being 0.3 higher than at public schools and private school grade inflation rates were 25% to 30% higher (Smith & Fleisher, 2011). Students are expected to spend 24 to 30 hours studying for courses, however, Caruth and

Caruth (2013) found the current student averages 27 hours, while in 1961 students studied 40 hours a week.

Caruth and Caruth (2013) examined grade inflation in higher education using 14 publications. Grade inflation was defined as the upward shifting of grades without corresponding increases in learning or performance. The grade-point averages at private colleges rose 7%, from 3.09 in 1991 to 2006. At public colleges and universities, the average grade-point average rose 6%, from 2.85 in the same time frame. Three common causes of grade inflation found in the literature reviewed were identified as student evaluation of professors, student teacher dynamics, and merit-based financial aid. Student study time was also found to be changing. In 1961 the average student spent 40 hours a week engaged in attending class and studying. In 2003, this time dropped to 27 hours. Although this data provides information on grade inflation, the validity of the study is questionable as search terms, assumptions, and potential biases were not identified. The consequences of grade inflation were identified and the authors proposed potential solutions for this problem.

Two retrospective studies on grade inflation were reviewed by King-Jones and Mitchell (2012). The first study examined nursing students' overall GPAs and clinical grades in one school, which revealed 4,500 nursing students' GPAs increased significantly overall over a 25-year period. The second study reviewed grades for theory courses and corresponding clinical courses for ten paired courses over a ten-year period. The researchers found a slight positive slope in the theory grades and a negative slope in the clinical grades. Normal distribution was found with theory courses whereas the distributions in clinical courses were atypical, showing that theory grades remained stable

while clinical grades were higher over time and the theory grades did not match the higher clinical grades. Factors influencing these findings included use of adjunct faculty, tenure systems, and institutional responses to economic challenges. A review of the literature revealed higher student grade point averages in classes taught by adjunct faculty. Other factors identified were fear of poor student evaluations, lack of faculty effort and lack of faculty training. This review has limits due to lack of support of further literature. The authors did not identify the search terms; however, the potential bias was described by authors as to the use of numerical grades for clinical courses, rounding up of grades, and the weight of quizzes in grading schemes. The need for further research of numeric grading and pass/fail grading systems in clinical courses was identified, as understanding of theory knowledge is directly related to clinical practice.

O'Halloran and Gordon (2014) examined the current literature on the issues surrounding grade inflation in higher education in the United States and cited the decrease in standings in the standardized testing of students in over 60 countries. There is evidence in the literature reviewed that students are spending less time on studies and other educational activities. Furthermore, the information in the literature reviewed showed students are not as engaged in course material, reading only material that they considered necessary. Looking at three major universities, one-third of students did not attend scheduled classes. The authors pointed out that academic publications report grade points are trending upwards, while time series studies based on a national collection of college transcripts have shown smaller increases. These reports can be skewed by students dropping failing courses. Another factor is the pressure placed on educational institutions by accreditation groups that review graduation rates, ability to get a job, and

furthering education in graduate school. The authors discuss potential reform of education to reduce the factors that influence grade inflation. Connecting learning objectives, curriculum materials, teaching strategies, and rewards so students can see the purpose and relationship is recommended. This is not a new approach.

Smith and Fleisher (2011) reviewed literature on current and past practices of the grade inflation at public and private universities. They report an increase 0.6 grade point average increase from 1967 to 2000 with private schools having 25% to 30% higher grade inflation rates than at public schools. The authors suggested the cause is related to stakeholders' expectation of higher grades and the thought that paying high costs for education should result in students receiving higher grades to obtain better job opportunities. Adjunct faculty members were found to grade higher than tenured or non-tenured faculty members. Grading could be influenced by lack of teaching experience and teaching skills, or teaching could be more effective due to better training coupled with more motivation (Smith & Fleisher, 2011). They hypothesized that non-tenured and part-time faculty submit higher grades than do tenured faculty members, since merit, tenure, and promotion decisions are based, for the most part, on an instructor's teaching performance, as measured by course evaluations.

The clinical grade discrepancy score, a new measurement of grade inflation, was used in a descriptive correlational study to examine the relationship between exam and clinical grades. Paskausky and Simonelli (2014) studied the relationship between licensure exam-style final exams grades and faculty assigned clinical grades of 281undergraduate students for evidence of grade inflation. This study used secondary data of final exam grades and corresponding clinical grades. SPSS was used to determine

descriptive statistics and correlation analysis. Clinical grades were B+ or better while exam grades ranged from 59 to 93 out of 100 points. The result suggested grade inflation was present with a correlation at 0.357, which is considered moderate to low. Clinical grades were B+ or better while exam grades ranged from 59 to 93 out of 100 points. A ninety-eight percent positive discrepancy score showed that grade inflation was present with 70% of grade discrepancy between final exam and clinical grades being at least one full letter grade.

Grade inflation is an issue with how nursing instructors are grading. Studies reveal increases in student grades with students achieving the same levels of knowledge (Caruth & Caruth, 2013). Student grades are increasing while student study time is decreasing (Caruth & Caruth, 2013; King-Jones & Mitchell, 2012). Students are spending less time studying and other scholarly activities, reading only material that they deem needed (Caruth & Caruth, 2013; O'Halloran & Gordon, 2014). Researchers found discrepancies between clinical and theory grades, with clinical grades showing higher abilities than correlates with the theory grades (O'Halloran & Gordon, 2014; Paskausky & Simonelli, 2014; Smith & Fleisher, 2011). The inflation of grades poses a real concern for public safety. Nursing educators are responsible for ensuring students are prepared and will provide competent quality care. Gross incompetence in clinical is identified, but subtle weakness often go undetected (Paskausky & Simonelli, 2014). Grade inflation in nursing programs and the relationship to patient safety is an area that is lacking in research.

Grading Consistencies

Grading consistencies are not always found in nursing programs. O'Flynn-Magee and Clauson (2013) performed a qualitative study and found commitment to fair grading was strong, but there are inconsistent grading practices. Consistency in grading practices is needed. However, Pazargadi, Ashktorab, and Khosravi's (2012) descriptive qualitative study noted that students felt there were inconsistencies between instructors and between evaluations given by the same instructor at different times. In a cross-sectional data and content analysis study by Salminen and fellow researchers (2016) students reported feelings of nursing faculty as authoritative and were found to treat students unequally by applying rules differently to different students. Grading students in inconsistent ways was noted to be the second most commonly perceived unethical behavior (Yildiz, Icli, & Gegez, 2013). Billings and Halstead (2016) stated students have a right to be treated fairly, consistently, and objectively. Communicating expectations reduces misunderstandings and provides clear expectations of the course.

O'Flynn-Magee and Clauson (2013) conducted a qualitative study using informal focus groups with 13 faculty members. This was 33% of the faculty working at the university where the study was conducted. The faculty were from undergraduate and graduate programs with varying ranks and teaching experiences. Three main questions for the focus group looked at beliefs and values regarding grading practices, perceptions of effective grading strategies, and approaches for consistent grading. Ethical and relational practices related to grading practices were the two main themes identified with thematic analysis. Key words like equity, confidentiality, anonymity, consistency, and objectivity led to the ethical practices theme. Relational practice key words and phrases

included respect, promotion of self-esteem, caring, sharing of power, and communication.

During the focus groups, grading consistency was reported easier to achieve when using grading systems, standards, and tools (O'Flynn-Magee & Clauson, 2013). Several faculty spoke about commitment to supporting learning and success of students. Feedback was considered to be central to effective teaching and the timing and nature of feedback were an important consideration. Potential bias when the student's name is known was acknowledged, so anonymous grading is sometimes preferable. Grade inflation was not a key concern during the discussions although the researchers believe that grade inflation is linked to inconsistent grading practices. Researchers commented on how faculty did not realize their use of power with students. Educators felt their decisions on grading were fair and were unaware of how they used their power to determine whether they would discuss changing grades, including not discussing the grades if they did not feel it necessary.

The researchers concluded consistent, fair grading practices are a professional responsibility and should be based on policies and guided by principles (O'Flynn-Magee & Clauson, 2013). Inconsistent grading practices along with the grade inflation trends have affected the ability to predict student success on NCLEX®, poor student critical thinking and problem-solving skills and the inability of new graduates to practice as a professional nurse effectively. Nursing educator beliefs affecting student abilities include beliefs about satisfactory performance and relevant grades, subjective clinical grading, failure to fail in clinical, discrepancies in clinical and theory grades, the preponderance of

part-time and casual clinical instructors; and discrepancies between theory and clinical course grades.

The limitations of O'Flynn-Magee and Clauson's (2013) research included what faculty chose to discuss in the groups. Information that was not going along with policy or practice was not discussed. Opposing views of faculty might not have been expressed during focus groups since it could make working together difficult. Student views were not included in this study. O'Flynn-Magee and Clauson recommended further research on the relationship between grading practices, meaning of grades and grade inflation. Supportive strategies of grading, writing workshops, grading rubrics, and a blind review process for written work were recommended. This research supports the need for specific guidelines and policies for fair, consistent grading practices.

Yildiz, Icli, and Gegez (2013) researched the code of ethics for faculty using a questionnaire of general ethical guidance for academics provided by The American Association of University Professors (AAUP) and refined by The Academy of Management (AOM) and The American Marketing Association [AMA] (Yildiz, Icli, & Gegez, 2013). A 5-point scale was used to identify how strongly faculty felt about 32 statements regarding unethical behaviors. Data obtained from questionnaires were analyzed using SPSS© software. The questionnaire was given to 100 faculty members from public and private universities. The sample included faculty with different academic ranks and experience levels. All participants had doctoral degrees. Faculty have several roles, but the study was focused on the teaching aspect. Unethical behavior was defined as conduct not allowed or tolerated in professional practice and the ethical behavior that is expected is built on mores of the culture, education, and religion of the society as a

whole. Code of ethics has been referred to as codes of practice, code of ethics, and codes of conduct; all mean a compilation of some ethical standards or rules on how to behave.

Yildiz and colleagues (2013) analyzed the responses to the questionnaires and those with a mean score of more than 3.0 out of 5.0 and a standard deviation less than1.0 were deemed to be considered unethical behaviors. Twenty statements were identified as unethical by at least 78.7% of participants. Further analysis using t-test revealed that no statistical significance was seen comparing response from males and females. The correlation analysis did not show a statistically significant relationship between identified unethical statements and ages, years of experience in education, or academic title. All participants rated engaging in unbecoming behavior with students is unethical. The second highest identified behavior was grading students inconsistently. Statements on not explaining grades to students and misleading students also scored as unethical behaviors. Interestingly, faculty felt to disclose grades to administration that do not need to know grades is unethical, but it is acceptable to tell grades to parents.

Yildiz, Icli, and Gegez (2013) discussed the need for a code of ethics for academics and the concerns surrounding a formalized code. Some professors are concerned that the establishment of a code could interfere with professional autonomy and leave them exposed to student complaints. The research by Yildiz and colleagues provided interesting insight into what this group of faculty views as unethical behaviors. As with all studies relying on participant statements, responses could be affected by what participants perceive they should answer. The study had a sample of 100 faculty members from different institutions. Although participants were from different institutions, all were from the business divisions.

Grading consistencies are not always found in nursing programs. Research in this area focused on beliefs and values regarding grading practices. During a faculty focus group, researchers discovered faculty felt they graded consistently and were unaware of the use of power with students; such as not discuss changing grades or the specific grades depending on how the faculty member felt. Another study surveyed faculty and found that inconsistent grading of students was the second highest rated unethical behavior (Yildiz, Icli, & Gegez, 2013). Salminen and fellow researchers (2016) reported students felt instructors applied rules differently to different students, noting an inconsistency between instructor evaluations and even evaluation by the same instructor at different times. The inconsistency with grading along with inflating of grades has caused difficulty for predicting student success with NCLEX® along with students with decreased critical thinking and problem solving skills (O'Flynn-Magee & Clauson, 2013), adding to the concern for patient safety once students enter professional practice.

Grading Ethics

Ethics form the standard that determines what is right and wrong based on moral principles, standards, and rule of conduct (Yildiz, Icli, & Gegez, 2013). Professional ethics include principles, standards, and rules and are impacted by the philosophy, history and societal expectations of individual professions (Yildiz, Icli, & Gegez, 2013). Research focusing on the nurse educator's professional code of ethics is limited (Salminen et al., 2016). Docherty and Dieckmann (2015) reported that grade inflation relates to the ethical code of nurse educators. There is a moral and professional responsibility to fairly and reliably assess students' performance (Vasiliki et al., 2015).

Failure to perform the item analysis and to address the findings of this analysis can result in unethical assessment of students' examinations.

A literature review by Fowler and Davis (2013) found less than 10% of the nursing journal articles identified as ethical in nature discussed nursing education ethical concerns with a small amount of those discussing grading and evaluating students. The majority of ethical articles addressed the teaching of ethics in nursing schools, with single subject articles focusing on faculty author rights, cheating or prejudicial grading. The original purpose of the review was to look at the nature and extent of ethical issues in nursing education. The research shows a need for more information needed on systematic and comprehensive research on ethical issues in the context of nursing education.

Salminen and fellow researchers' (2016) descriptive study on professional ethics of nurse educators focused on the application of ethical principles within the teaching profession and nursing practice. Surveys were sent to nine nursing education programs. Along with survey questions, nurse educators and nursing students were asked two open ended questions on the three main ethical principles that guide the work of nurse educators and what ethical issues are faced by nursing faculty and nursing students. Both groups identified justice and equality as key issues with educators naming honesty and students identifying professionalism as the third. There was a response rate with 202 students and 342 nurse educators with diverse socio-demographics in both groups. The response from both groups shows a lack of understanding on identifying ethical topics. The students placed many issues in the category of professionalism, leading the authors to wonder if the students have a clear understanding of what concepts are included in professionalism.

Docherty and Dieckmann (2015) conducted a cross-sectional, descriptive survey looking at participants' experience with teaching and grading, whether they had training in grading or teaching, and questions regarding factors that could influence grading practices. Participants responses revealed 43% awarded higher grades than merited, 17.7 % passed a failing written examination, and 15.2 % of grading practices were influenced by upcoming licensure examinations. Lack of instruction on grading nursing exams was reported. When nursing faculty were asked about formal training, half reported that no formal instruction on grading was offered. Twenty-six percent of participants reported that knowing the name of the students influenced how they graded and 57 percent stated they had given students the benefit of doubt during the grading process. More research on faculty perception of administration's role and fear of litigation influencing grading practices was recommended.

Salminen and associates (2013) researched professional ethics of nursing educators. The authors looked at the knowledge of ethical principles and 18 questions in fairness, respect, and treatment of nurse educators by society. The ethical actions required as teachers include recognizing students' learning needs as the primary focus; being supportive and encouraging, acting as a role model, respecting the students, and being fair. Students report these qualities are not always present when educators give assessments and feedback.

The Salminen and associates' (2013) study had 342 nurse educators ages 27 to 64 years. The graduate educational level for 232 was an academic degree with 194 having a master's degree and 40 had a Ph.D in Health Science. None of the participants had earned any continuing education credits on ethics in the previous year. The survey results

showed participants felt their knowledge of ethical principles as being good. Educators 46-56 years old had a higher educator ethics knowledge level than those educators younger than 45 years. The length of time as an educator affected the level of knowledge of ethics reported by participants. Educators with 10 to 20 years of experience rated their own ethical knowledge as higher than those educators with five or fewer years of experience. Participants rated their knowledge and self-reporting study results can be skewed by over-reporting of knowledge. The authors reported student beliefs in this report but did not evaluate this factor in the study.

The lack of ethic practice with grading is noted by several researchers (Fowler & Davis, 2013; Salminen et al., 2016). Many researchers point at deficiencies in nursing instructors' education in grading practices as a major cause for this issue (Docherty & Dieckmann, 2015; Salminen et al., 2013). As has been noted in several studies, grade inflation due to evaluation of test questions, raises the concern that students are not passing a course because of their actual knowledge, but rather because grade inflation pulls them to a passing level. This concern leads to the question of whether or not it is ethical to inflate grades. Grade inflation related to test question evaluation may misrepresent the students' true scores making it challenging for proper assessment of an individual student's progress and assessment of the class as a whole (Oermann & Gaberson, 2014).

Nursing Instructor Education

Tarrant and Ware (2012) reviewed literature on the use of MCQs in nursing education and found nurse educators lack the essential knowledge and training needed to develop high-quality questions. Theory instructors develop most of the tests used in

nursing or questions are chosen from test banks from textbooks. This practice frequently results in substantial deficiencies as only a small number of nurse educators have sufficient preparation and knowledge on developing high-quality MCQ assessments. The authors evaluated one nursing school's MCQs for cognitive levels with Blooms Taxonomy. The questions obtained from the nursing textbooks revealed 72.1% measured knowledge and comprehension, which are the two lowest levels of Bloom's Taxonomy. The quality of questions used at the school over a five-year period was evaluated and 91.1% of 2,770 MCQs were written at the knowledge and comprehension levels.

Recommendations by Tarrant and Ware (2012) for test development were provided to help guide nursing faculty in developing and improving their MCQs. Faculty education on writing test items is important. Studies have shown test items written by educators with training had higher quality. The next step in producing a valid and reliable test is developing a test blueprint. This is a plan of how many test questions should be given for each content section and at what cognitive level the questions should be written; both should align with course objectives. Next, faculty can start developing MCQs using cognitive levels of application and analysis. Test items should include clinical decision-making tasks and not just recall of facts. Questions should have students apply information learned to make a judgment. Plausible answer options are important to make high quality test items. The answer options should distract students unfamiliar with the content, but should not be misleading to those knowledgeable about content. The authors also recommend writing 50 to 60 test items for each exam. Items should be reviewed for common errors such as grammar consistency in question and distractors, consistency in

length and detail of all options, credibility of all options, and assurance that all questions have only one correct answer. It is good to have someone else proofread the questions.

After the exam is given, an item analysis should be performed (Tarrant & Ware, 2012). This is a statistical analysis designed to assess the test items. The analysis reviews response distribution, item difficulty, item discrimination, test reliability, and the mean exam score. The analysis can help faculty determine poorly functioning questions. These questions can be removed or edited for future use. The statistical analysis helps identify the MCQs that perform well and can be used in the future.

Booth and associates (2016) reviewed literature on the educational preparation of nursing educators. The authors found that clinical expertise does not translate into teaching expertise, since education and nursing are two distinct disciplines. Knowledge of evidence-based research and practice, teaching methods, and curriculum design and development is needed to have the basis for practicing in the academic setting. Currently, there are no standards for educational preparation of nurse educators other than having an advanced degree. Pedagogical competencies in curriculum development, teaching strategies, and evaluation methods are recommended. Graduate nursing programs have been changing over the past 40 years from a focus on administration or education to clinical specializations. With a lack of defined evidence-based teaching practice, the application of evidence-based practice in nursing education has only recently been recognized in the nursing field.

Schoening (2013) conducted a grounded theory study of 20 nurse educators in the Midwest. Nineteen of the participants were employed fulltime as nurse educators and 19 of the participants were women. Of the participants, three had a master's degree in

nursing education, four reported taking at least one elective in nursing education during their graduate program. The remaining 13 had no formal preparation for teaching. Data were collected in semi-structured, face-to-face interviews.

Schoening (2013) reported novice nurse educators have less pedagogy training than those in the past. In 1969, the American Nurses Association recommended graduate preparation focus on clinical specialties to advance nursing theory and science. In the years that followed, the attainment of educational degrees for teaching decreased and in the 1990s, only 4 percent of graduate students were preparing for nursing education. This research led Schoening to the identification of four phases of transition from nurse to nurse educator. The first phase is named anticipation/expectation. This is followed by a disorientation phase. Next is the information seeking phase concluding with the final phase of identity formation. Participants described a lack of formal orientation and mentorship along with disbelief that they were expected to teach without prior experience. Another issue a majority reported was a lack of training in curriculum development and teaching strategies. The study identifies issues that occur in nursing educational settings. Since the sample group was small, more research is recommended (Schoening, 2013).

Cooley and De Gagne (2016) identified similar issues in nursing education. A phenomenological qualitative study to gain insight about novice nursing faculty's experience in academia was conducted. The authors reported more specialized nurse clinicians entering academia, but new faculty lack knowledge and preparation for the role of nurse educator. Researchers examined perceptions of facilitators and barriers to nurse educators' practice competence. The data gathered were from seven faculty teaching in

private, four-year colleges and was analyzed using Moustakas' seven step process. The participants depicted an experience that was lacking in both key information and supportive guidance.

Cooley and De Gagne (2016) recommended internship programs to assist novices to acclimate them to the academic environments and to assist in developing their competency in educational practice. Participants expressed concerns about test development, item analysis, and effective ways to teach. Achieving the best learning outcomes was an additional concern. The lack of resources and mentors for guidance was noted. Participants developed as educators by considering teaching evaluations received from students. Themes for successful novice educators included dedication to the nursing profession, a sense of obligation and responsibility to the students, diligence to teaching students well and responsibly, and an understanding of the impact of their instruction. This study's findings of new faculty lacking knowledge, support, and time are consistent with the literature.

The research reviewed confirms the lack of training and instruction in curriculum planning, teaching strategies, and evaluation methods for nursing instructors. Furthermore, clinical expertise does not prepare nursing instructors for the classroom as education and nursing are different theoretically (Booth et al., 2016). Lack of educational strategies results in difficulty formulating assessment items. Nursing programs primarily use multiple choice questions and require the necessary scholastic knowledge and training to develop high-quality questions (Tarrant & Ware, 2012). The evaluation of MCQs used in a nursing school and textbooks revealed an alarming rate of poor quality

items. The need for evidenced-based nursing pedagogy has recently been recognized (Booth et al., 2016).

Multiple Choice Questions (MCQ)

Assessment tools are used to determine grades, identify any student learning gaps, and evaluate if students have achieved course outcomes (Quinn & Novotny, 2012). Nursing programs use MCQs as they can be reliable, easy to administer to small or large groups, and are cost efficient (Begum, 2012; Davey et al., 2015). MCQs determine if the student has the ability to assess, apply and evaluate information (Bauer et al., 2011). Multiple choice questions were first developed by Edward Thorndike and were first used at the Kansas State Normal School in 1914 (Siddiqui et al., 2016). In current practices, MCQs are developed using a stem of the question and then possible answers called distractors that are credible, correct, but not be too close to the correct answer (Kaur, Singla, & Mahajan, 2016). Well written MCQs measures knowledge, comprehension, application, and analysis (Baig et al., 2014; Bush, 2015). MCQs have the advantage of allowing instructors to give rapid feedback and the ability to analyze test results quickly (Bauer et al., 2011).

Namdeo and Sahoo (2016) evaluated the quality of MCQs using a difficulty index (DIF I), discrimination index (DI) as well as the number of non-functional distracter (NFD). The researchers analyzed 25 MCQs and 75 distracters for DIF I and DI and presence of numbers of NFD. The test group had 25 students in the higher ability level and 25 in the lower level group. The groups were determined by a pretest of 76 students. The 26 individuals in the middle group were not included in the study. Fifty–six percent of the test questions were judged to be in the acceptable range of difficulty. The DI

distinguishes the answers of high ability and low ability student. The results were 48% and 53.4% of distractors were considered ineffective. The high percentage of distractors considered ineffective could skew the results of the difficulty index. Even though the number of questions was very small, the results point out the need to review multiple choice items.

The questions had a single stem with one correct response and three distractors. The 50 question exam was given to 148 students with a 60 minute time limit. The researchers analyzed the results using simple proportions, mean, standard deviations, coefficient of variation and the unpaired t test. Twenty-four items had good to excellent DIF I (31 - 60%) and 15 had good to excellent DI (> 0.25). Mean DE was 88.6%. The researchers noted that the average score of students in this study was 33 of 100. The authors also reported that low or negative DI could be caused by an incorrect answer key, ambiguous framing of questions, or generalized poor preparation of students, with the last being identified as the cause for the low DI.

Researchers Gajjar, Sharma, Kumar, and Rana (2013) encouraged the use of item analysis to assess the quality of questions and determine items that needed revisions. The analysis can also identify items too difficult, which deflate scores, or too easy, which inflate scores and can lead to a decline in students' motivation. The authors noted that a good MCQ can evaluate cognitive, affective, and psychomotor domains and has the advantages of objectivity by minimizing bias and comparability. It has the ability to assess a wide variety of concepts in a short amount of time. It would be advantageous to know if the questions were revised or if the test was administered to a different group of students, and what the results of another sample were.

Kaur, Singla, and Mahajan completed another study, assessing the quality of MCQs (2016). The researchers examined 50 MCQs and 150 distracters given to 150 students by assessing the DIF I, DI, and DE. The results of the evaluation revealed 76% of items were in the acceptable range of difficulty, 62% had excellent discrimination index, and 82% of questions had functional distracters. Like other studies looking at the quality of test questions, this study used a small sample size and the test was only administered once. The results show the importance of analyzing test questions for their quality and to ensure faculty are testing with appropriate evaluation tools.

Maher, Barzegar, and Ghasempour (2016) examined the relationship between MCQ taxonomy and negative stem questions using a cross-sectional study with 2400 written multiple-choice questions. The study looked at level I taxonomy questions for negative language including not right, wrong, except, unless, but, least, not likely and forbidden and found 63.9% with negative stems. Level II and Level III Taxonomy questions only 36.1% had negative wording. The researchers questioned the validity of the test as the low-level taxonomy questions require only superficial learning and memorization of material. The test taker with low cognitive levels can guess answers with negative formats due to this format making it easier to eliminating distractors. Unfortunately, due to lack of difficulty index and discrimination index, the relationship between the quality of questions and the negative wording could not be determined.

Baig, Ali, Ali, and Huda (2014) used Buckwalter's modification of the Bloom's taxonomy to review the cognitive levels of MCQs. One hundred and fifty MCQs were reviewed by one subject expert and three medical educationists looking at quality, cognitive level, and item writing flaws. The cognitive review found 76% of the questions

were at recall level and 24% were at the interpretation level. Sixty-nine item writing flaws were identified with 30.43% implausible distracters, 27.54% unfocused stem, and 24.64% unnecessary information in the stem. The reliability and validity of the Buckwalter's taxonomy tool and item writing flaw tool were documented in the study. The items reviewed were used for the first time in the first year that the school was opened. The need for a test blueprint was identified. A test blueprint maps the course objectives and content and allows the user to calculate the percentage of test items for different cognitive levels as well as content to include in the assessment. The study did not include a difficulty index or discrimination; however, the results pointed out the need for MCQ items to be reviewed for content and cognitive levels. The authors recommended a faculty development program to assist with test development and alignment of test items to student learning outcomes. The research report did not include difficulty index or discrimination.

The need to use rule mining to improve MCQ assessments was presented by Romero, Zafra, Luna, and Ventura (2013). MCQs provided multiple data points, including students' answers, individual question scores, final assessment score, and execution times. Use of this data was not always obtained due to difficulty with the use of statistical information. This study involved 104 students taking a MCQ assessment. Only one instance of testing was analyzed. The researchers used two new data matrixes to analyze the results and compared them with the traditional score matrix. The relationship matrix evaluated the concepts taught in the course combined with the knowledge matrix, which measures up to five questions on the same concept. The score matrix calculates

zero for an incorrect answer or one for a correct answer, along with the total time taken by the student. The results of the scores were analyzed for potential relationships.

Romero, Zafra, Luna, and Ventura (2013) identified three patterns in the study's data. The item-time-score pattern revealed relationships between items, times, and scores. This information is useful to see the effect of how much quiz time is given to the results of the exam. The instructor can then increase or decrease the test time according to the results. The pattern of relationships between several questions is called the item- item pattern. This process can identify if one question is right, then several other questions might be right. This process can also identify concepts that are influential in an understanding of course material. Concept-score patterns provide information on the relationships between concepts and scores and identify the concepts that might need to be modified or extended for students to fully understand the information.

An overview of traditional guidelines based on current literature for writing effective MCQs was presented by Begum (2102). Begum defined multiple choice questions as objective tests where students choose a correct response and are used because they are easy to grade, large numbers of students can be tested at the same time, and can cover a large amount of course content in a short amount of time. Unfortunately, MCQs can be difficult to develop, students are not able to demonstrate original or creative thinking and allow the test taker to guess correct answers. Questions with a single best answer format allow instructors to test knowledge, problem solving, judgment and perception as the student must select the most correct or appropriate answer from all correct responses provided. These questions can be difficult to develop and a flawed MCQ can cause the question to be inaccurate and will not test the students' competency.

Begum recommended that test developers be skilled in effective test item writing and that questions should be reviewed with feedback to the writer.

Multiple Choice questions can be used to assess student learning gaps and evaluate student learning outcomes. Nursing programs use this style question for formative and summative assessments. Although difficult to write, MCQs can measure higher cognitive learning and multiple concepts at the same time (Gajjar et al., 2013; Romero et al., 2013). MCQs need to be analyzed for effectiveness. The research showed various ways to evaluate test items and areas to be considered when developing test items. Analyzing the test items was a first step and studies reinforce the need for test blueprints and peer review of items. In order to write appropriate MCQs, nursing instructors need education on writing techniques and training in item analysis.

Test Item Analysis

An item analysis should be performed on MCQ test results (Baig et al., 2014). Item analysis is the assessment of the quality of test questions using the information from students' responses (Khoshaim & Rashid, 2016). O'Halloran and Gordon (2014) and Quinn and Novotny (2012) reported that item analysis could help determine the reliability of the test and generate a better analysis to determine different knowledge levels among students. Item analysis can improve assessments and teaching methods by allowing faculty to focus on weaker areas noted from the results (Talebi et al., 2013).

Reliability and validity of test items can be determined using item analysis. The responses from students can be broken down to review how the overall student population answered the question. Looking at the difficulty level of the question and how often each answer was chosen helps to assess the quality of each MCQ (The International

Test Commission, 2014). The use of this analysis is an important part of course assessment and determining the quality of the overall exam (Talebi et al., 2013). New computer programing allows for test scoring that looks at discriminators and other analysis of exam data (O'Halloran & Gordon, 2014). This data is only useful if a proper analysis is performed. Unfortunately, test item analysis is not performed by all nursing faculty, and when the analysis is performed, there is no uniform way to use the information, as this is often left up to the individual grading the exam (Killingsworth, 2013).

Nickerson, Butler, and Carlin (2015) purported test analysis can be performed for a variety of reasons. The research was performed with 130 participants taking a 50 questions multiple choice test. The first group took the test by choosing the correct response. The second group took the test assigning points towards the probability the answer was correct. This type of testing allows the individual to test by what they believe to be true. The analysis provides feedback on how well the material was presented and what parts of the concepts have been recognized by all students. This analysis is achieved by using the spherical gain rule found by taking the points assigned to a response and divided by the square root of the sum of the squares of the weights from the other responses in the question. When reviewing the exams using probability points, only 29% of the questions had all the points placed on a single answer. The probability group scored 11 points higher on the exam than those who took it the conventional way. Twenty-four of the 28 students using the probability testing method reported preferring this type of testing over the conventional testing.

The research of Nickerson and colleagues (2015) revealed probability testing allowed demonstration of understanding the subject being tested. The error caused by guessing was completely removed. The arguments over correct answers are dismissed as students assign points to each possible response, making students take responsibility for their answers. Probability testing encourages reflection on the material being tested. This testing provides the instructor clear feedback on the understanding of the material covered, which can be used to give remediation if necessary. This testing does take more time to score and can be longer to administer to students. Students must also understand how the test is to be taken and how the examination is scored. This can be time-consuming as well. None of the subjects in the research study reported difficulty understanding examination instructions.

The use of software programs to evaluate the difficulty of multiple choice questions was researched by Vasiliki and associates (2015). The software program was designed to evaluate the difficulty and discrimination of 220 multiple choice questions that had been written by a faculty member and reviewed and approved by two other faculty members. The data from a group of 497 students taking the exam were imported into the software program. The assessment of the difficulty level of the questions revealed only 16.8% were appropriate level and 33.2% were of excellent discrimination, while 53 (24.1%) were of bad discrimination. Additionally, there was an association found concerning very low or very high difficulty and the poor discrimination of high versus low performing students. The software program was quicker and identified flaws not detected by those who reviewed the questions.

The conclusion of the Vasiliki and associates (2015) researchers showed that review of multiple choice questions by the software program was more accurate than the judgments of the faculty when developing an objective assessment of students' learning. When researchers removed the flawed items from the test, 10-15% additional students would have received a passing grade. The study showed that objective analysis is needed and that software programs can provide the analysis faster and more accurately than a judgment of multiple choice questions. The faculty used judgment to determine the appropriateness of questions. The questions were chosen from a bank of questions used by the faculty. The researchers did not provide detailed information on how questions were chosen by the group. The results could be different if the faculty chose by content information or by using item analysis information gathered from previous examinations.

Talebi and colleagues (2013) performed a three-year investigation to study the effect of item analysis to improve assessment and teaching quality. The final exam in two semesters was analyzed. A new final exam with improved MCQs was administered the next semester, followed by improved teaching and an improved final exam the last semester of the study. This four stage study analyzed 40 MCQs used in four semesters. The final exam for the course was first performed following routine teaching forty different, but equivalent, MCQs were prepared for the final exam of each group (totally 80 MCQs for both groups). Item analysis software was used to analyze the items and the researchers developed new questions based on the descriptive item analysis. The criteria for deleting and changing questions included the difficulty index, poor or negative discrimination, poor distractors or items that did not have appropriate content.

The results of the research by Talebi and colleagues (2013) had no difference in the first two groups, the third group showed some improvement in scores (difficulty index 0.54) and the fourth group showed great improvement (difficulty index 0.65). The study supported the use of item analysis and the effectiveness of the improved teaching method. Unfortunately, the study had a small number of participants and courses were taught by different faculty members. Further studies with larger sample size are needed to generalize the results to the whole population.

Item analysis of multiple choice questions is important to determine reliability, knowledge level of students, and gaps in concepts (O'Halloran & Gordon, 2014; Quinn & Novotny, 2012). Item analysis can be performed a variety of way. Software programs have been developed providing an objective item analysis (Vasiliki et al., 2015). Item analysis can be performed over time with nursing instructors changing ineffective items to produce a more accurate student assessment.

Poor Performing MCQs

Once an exam question is deemed to be poor, faculty have the following options: (1) full points can be given to all students for the question, (2) full points can be given to those who got the question correct, while those who got the question wrong can be given partial points, or (3) the poor question grades can be maintained, unchanged. Options 1 and 2 can lead to grade inflation, meaning the grades students receive are higher than they should have earned, given their understanding of the subject being tested (Phelps, McDonough, Parker, & Finks, 2013).

Different theories of scoring multiple choice exams were explored by Barnard (2013). Four theories were reviewed for the impact of missing responses on the score of

the examination. Classical test theory is based on using the correct responses and the amount of errors to determine the score. This process includes reviewing the item's difficulty and discriminative values. The difficulty is the ratio of the number who answered the item correctly to the number of the test group. Item discrimination indicates the quality of the test item. A second theory is The Item Response Theory, which uses a statistical approach to provide estimates of the abilities of those being tested. The third theory is the Rasch's measurement theory. This theory used the observed score and the difficulty values of questions to determine the abilities of those taking the exam. The last theory reviewed was the Choice Probability Theory. This theory included measurement error to help determine if the response was a guess by looking at the probability of each option being the answer. The Choice Probability Theory started with the assumption that most students do not guess but do use elimination of distractors. The research looked at how test results can be interpreted to determine if guesses are made and if they are based on facts. It should be noted that the Choice Probability Theory was developed by the author.

Phelps, McDonough, Parker, and Finks (2013) explored how decisions on how test items are scored affect grading practices. Item analysis of multiple-choice questions is an important tool to measure if learning is assessed or if it is a deceiving question. The authors point out that there are inconsistencies in grading among faculty. Some faculty use the accepted process for analyzing questions. Grading practices vary between faculty at the same institute causing inconsistencies. Some will score examinations by keeping poorly performing questions, counting multiple answers correct, or by eliminating the entire question from the test, with subsequent readjustment of student scores. The authors

gave examples that show how test scores significantly change depending on how poorly performing questions are graded.

Phelps and colleagues (2013) argued that when faculty analyzes a question and it has a negative point biserial, adjustments should be made. Sometimes, a point biserial confirms a good question, but faculty may delete the question or take multiple answers. One example given is an examination with 25 questions with three performing poorly with 40-60% responded correctly. However, the point biserial of +0.09 to +0.30 indicates a good question. The faculty removed the three questions. A student who had those questions right and had 21 of the 25 questions right would have an 84% score. With the change in the number, the score would change to 18 of the 22 or 82%. However, the student that obtained 21 of the 25 (84%) right and had the three questions omitted wrong, the grade would change to 21 of the 22, which would score at 95%.

A second example given by Phelps and colleagues (2013) was rekeying the answers to give credit for poor questions. Using the same test example of 25 questions with three poorly performing questions with biserials, between +0.09 and +0.30, the student who correctly answered the three poor question would receive 21 of 25 questions or an 84% before and after the rekeying. The student who had the three questions incorrect, started with a 21 of 25 (84%) and with the rekeying, receive 24 of 25 (95%). Only the student that incorrectly answered the questions has an improved grade, which does not reflect the knowledge level. Faculty might not be aware of the effect of raising the score.

Diedenhofen and Musch (2015) performed an experimental research study comparing the number right scoring method and the empirical option weighting for

improved reliability and validity. A sample of 120 students was used where one group received full information, one group had partial information, and the third group did not receive any useful information. Responses were scored using number-right scoring and then using empirical option weighting. This method was then used on a second group of participants who were randomly assigned to groups with 196 receiving no information, 191 receiving partial information and 180 given complete knowledge. Exam scoring was performed twice, once using standard dichotomous scoring with no partial credit for incorrect responses and the second time, empirical option weighting using the point-biserial correlation between choices and total score to determine weight of each answer. The empirical option was shown to have increased reliability on multiple-choice tests. The analysis allows for different knowledge levels to be identified by responses given. This type of scoring grants partial credit for partial knowledge and identifies gaps in knowledge base. Instructors can review the concepts identified by the knowledge base gaps.

Bauer and colleagues (2011) identified gathered data from six end of term exams and used different scoring algorithms for multiple correct answer multiple-choice assessments. Three algorithms were reviewed. Dichotomous scoring was defined as one point all true answers and no wrong answers were chosen. A partial credit algorithm 1was set up as one point for 100% all true answers; 0.5 points if 50% or more true answers chosen, no points for less than 50% correct answers, and no point deduction for wrong choices. Partial credit algorithm 2 was set up with a fraction of one point depending on the total number of correct answers and no point deduction was taken for wrong choices.

Data for the study Bauer and colleagues performed were obtained from 1,255 students with 180 MCQs used. The researchers reviewed answers and rescored the assessments. The results were analyzed for test reliability, item discrimination, and item difficulty. The two partial credit methods showed higher levels of psychometric results and are more accurate than dichotomous scoring. Along with providing more accurate assessments, the use of multiple correct answer multiple-choice assessments presents students with questions with representative clinical situations that can have more than one solution. When comparing the two partial credit scoring methods, the partial credit algorithm 1was noted to be slightly higher in reliability. The researchers concluded that partial knowledge should be awarded in MCQ exams. A strength of this study was the findings confirmed the results of the study by Ripkey and colleagues in 1996. The method of partial credit can separate the random guesses from the educated deductions. The authors recommended further study to determine the exact threshold for determining guesses verses deductions. Studies to determine the best number of distractors to use are a second recommendation from the researchers.

Another study in 2016 by Siddiqui, Bhavsar, and Bose reviewed different methods to score multiple response multiple choice questions, which evaluated the ability to recognize distractors in the question. One method can grant full credit only when the correct options are marked correctly. Another method is to give credit for each correct selection. The third option is to take away points when an incorrect response is chosen. The final option is to grant points for correct responses, delete points for wrong selections, but give no loss of points when correct selection is not chosen. The authors reported the last option provides the fairest grading method. The method is rational,

simple and does not reward students for guessing at answers. This evaluation method is good for formative evaluation where various aspects of concepts are being assessed. Multiple choice questions can be graded in different ways. Nursing instructors need to understand the various grading styles when scoring learning assessments. The studies reviewed provided valuable information on how the different techniques can change the scores of an exam. The various methods all affect how test items are scored and can therefore unwittingly raise the scores of students performing poorly and cause grade inflation (Phelps et al., 2013). Unfortunately, in many examples, the student performing well can also experience a change in score, but the score decreased for them. None of the researchers identified any of the methods of scoring multiple choice questions as best practice.

Theoretical Framework Literature

The theoretical framework used to guide this study was the Constructivism Theory. The constructivism theory developed from the educational constructivist learning theory which looks at how knowledge is built from previous experiences (Hunter & Krantz, 2010). The constructivist learning process is a process of taking in information, rearranging thoughts and feelings, and developing new ways to react to situations (Kickman, Neubert, & Reich, 2009). As an active learning process, constructivism requires inquiry and communication in which individuals discover differences, identify the meaning, and reform thinking (Hunter & Krantz, 2010).

Constructivism Theory is a scientific approach uses methodical processes to examine conditions and significances and includes experimental and instrumental applications to determine how perspectives link to practical consequences (Kickman,

Neubert, & Reich, 2009). There are five main principles of this theory. The first principle is learning environments that are complex and relevant (Almala, 2005). The second principle is the use of social interactions (Almala, 2005). The use of different perspectives and different methods of education is the third principle (Almala, 2005). The fourth principle is individuals are active participants in learning (Almala, 2005). The final principle is an individual's understanding and knowledge creation (Almala, 2005). The assumptions of this theory are learning is based on previous knowledge, new ideas are integrated, knowledge is formed and not memorized, and knowledge is developed with reflection and built on known concepts (Hunt, 2018).

The constructivism approach to classroom assessments is based on tests designed by teachers that are reviewed and revised according to test analyses performed (Graue, 1993). Instructors develop the assessments and each time the assessment is administered, the instructor reviews each item using the data obtained from analysis. This helps the instructor determine the quality of the item. The question can be redone and improved to achieve the required assessment objective. The process should be performed each time the assessment is given to students.

Methodology Literature

The methodology used in this research was quantitative. The National League of Nursing surveyed nurse educators on testing practices. The National League of Nursing conducted a survey of nursing educators which was analyzed by Oermann, Saewert, Ika, and Yarbrough (2009). The data were studied for factors that influence the use of student learning assessments and grading practices. Oermann and colleagues (2009) reported 83% of the participants use traditional and past practice when grading students (Oermann

et al., 2009). Nursing educators considered the pass rates of the NCLEX®, educational soundness of the assessment, and educational standards when developing student learning assessments.

Killingsworth (2013) developed a survey tool for participants to rate best practice activities with test construction, test item analysis and test revision. One hundred twenty-seven nursing instructors teaching in BSN programs for at least two years participated in this study. Using a Likert scale, participants self-reported use of best practices 75.7 percent of the time with test construction, 78.6 percent of the time with item analysis, and 80 percent of the time with test revision (Killingsworth, 2013). In 2015, Killingsworth and colleagues surveyed baccalaureate nursing instructors using the same Likert scale survey tool. Twenty two of the 26 best practice were rated as being used at least 85.6 percent of the time.

Reynolds (2015) conducted a descriptive quantitative study using a survey method to review differences in grading practices between associate and baccalaureate degree program faculty, full time and part time faculty, and tenure and non-tenured faculty (2015). The Grading Attitudes Questionnaire (Olsen, 1995) was used to obtain data. Data revealed that student grades varied between associate and bachelor's degree programs, public versus private institutions, and community versus four-year colleges. A gap was noted in studies focusing on grading practice differences between associate degree and baccalaureate degree nursing programs, public versus private institutions, and community versus four-year college. This may influence grading practices. Reynolds reported 56.3 percent of full-time faculty answering the survey did not receive education

on grading. Adjunct faculty reported the lack of education on grading practices. Clear grading guidelines with formal training on grading were recommended.

Quantitative research on the quality of multiple choice question is more prevalent. Kaur, Singla, and Mahajan (2016) examined 50 MCQs and 150 distracters for quality by measuring difficulty, discrimination and functional distractors. The study results were 76% of items were in the acceptable range of difficulty, 62% had an excellent discrimination index, and 82% of questions had functional distracters. Namdeo and Sahoo (2016) analyzed 25 MCQs and 75 distracters for quality. Fifty–six percent of the test questions were in the acceptable range of difficulty. Half of the distractors were considered ineffective.

Maher, Barzegar, and Ghasempour (2016) examined 2400 MCQ and discovered that 63.9 percent of low taxonomy questions had negative while only 36.1 percent of higher level taxonomy questions had negative stems. The low level questions can be answered by guessing since the negative format is easier to identify the distractors. The researchers lacked the discrimination and difficulty index to determine the quality of questions with negative wording.

In 2016, Siddiqui, Bhavsar, and Bose reviewed different scoring methods for multiple response multiple choice questions. The study focused on the ability to recognize distractors in the question. The researchers tested three grading approaches and determined the fairest method that does not award the test taker for guessing. The research by Diedenhofen and Musch (2015) focused on grading options. Their experimental research compared grading methods and the empirical option weighting had improved reliability and validity. The empirical option weighting uses partial credit and

identifies knowledge gaps. Overall, the importance of analyzing the quality of test questions and the use of appropriate evaluation tools is important.

Research Design Literature

A quantitative approach with a correlational design was used to obtain information on the use of evidence-based practice in test construction, test item analysis, and test revisions. Correlational research designs were used in two of the research articles presented. Killingsworth (2013) evaluated the use of best practices in exam construction. A correlational study on nursing instructors and their decisions about best practices in constructing, analyzing, and revising tests was conducted by Killingsworth (2013) The survey tool developed by Killingsworth (2013) asked participants to rate best practice activities with test construction, test item analysis and test revision. A pilot study using the new tool had appropriate internal consistencies. The research data indicated that nursing instructors thought they did a good job developing tests and reported using best practices in test construction, analysis, and revisions. Limitations of this study included using self-reported surveys and half of the participants worked at public institutions.

Killingsworth, Kimble, and Sudia (2015) performed a correlational design research study on nursing instructors using best practices for constructing, analyzing, and revising tests in baccalaureate nursing programs. This study used the tool developed by Killingsworth (2013). The results of the research were faculty frequently used best practices in test development, analysis, and revisions. Limitations of this study included self-reported surveys, use of only BSN faculty, and potential participants drawn to study were interested in test construction and evaluation techniques.

Conclusion

Current literature revealed several issues with grading practices in nursing education. Nursing faculty have an ethical responsibility to be fair to all students, to demonstrate consistency in all courses and to develop tests using on evidence-based practices (National League for Nursing, 2012). Vasiliki and colleagues (2015) stated there is a moral and professional responsibility to fairly and reliably assess students' performance. O'Flynn-Magee and Clauson (2013) documented inconsistent grading practices. Salminen and associates (2016) reported that students feel nursing faculty treat students unequally by applying rules differently to different students. Grade inflation has been documented in the current literature. Grade inflation has been correlated to evaluation of test questions, which raises the concern that students are not passing a course because of their actual knowledge, but rather because grade inflation pulls them to a passing level (Docherty & Dieckmann, 2015).

Most nursing instructors lack the essential knowledge and training needed to develop high-quality questions (Tarrant & Ware, 2012). Fifty-six percent of full-time faculty had not received education on grading practices (Reynolds, 2015). Booth and fellow researchers (2016) concluded that clinical expertise does not translate into teaching expertise since education and nursing are two distinct disciplines. Multiple choice questions are not easy to write and training is needed to properly analyze and review test items. The majority of questions reviewed, both instructor-developed and those found in textbooks, were found to be at a subpar level. Nursing instructors need the proper training and evidence-based guidelines on writing and administering nursing

assessments. This practice could decrease the grade inflation occurring and allow proper evaluation of students entering into practice to provide safe patient are.

Current grading practice of nursing instructors has been shown to be inconsistent. Standards for this are not written out for all to follow. Therefore, it is wise to discover if instructors use best practices in test construction, test analysis, and test revision. There has been shown a lack clear direction on how to analyze and grade examinations. Understanding of current practices are important to recognize in order for nursing educators to make changes in personal practices and institutions to require testing guidelines.

Chapter Summary

Current literature on nursing education assessments, use of evidence-based practices and ethical issues surrounding grading practices were presented in chapter 2. Grading practices of nursing instructors are influenced by several factors. A lack of clear guidelines for assessment practices accompanied by the lack of knowledge on educational pedagogies is part of the issue. Student assessments are not always based on evidence-based practice for testing practices. The result is inflation of grades and a discrepancy in the grades received and the knowledge achieved. There is a concern that public safety is at risk when nursing students are not effectively prepared to care for them. Grade inflation has been correlated with a decrease in critical thinking and problem solving skills (O'Flynn-Magee & Clauson, 2013).

Multiple choice questions are used by nursing programs to assess student learning. Although these questions are difficult to write, MCQs can evaluate higher level thinking and multiple concepts at a time. Item analysis needs to be performed on all

MCQs to evaluate quality effectiveness (Talebi et al., 2013). Software programs can be used for more objective item analysis. When test items are analyzed over time, nursing instructors can develop more accurate assessments for students. Even with item analyzes, nursing instructors must understand the various grading methods available to prevent grade inflation from unintentionally happening due to the type of scoring used.

Chapter 3 will present the methodology of the study. Rationale for the appropriateness of the correlational design will be stated. The research questions will be presented. Population and sample information is reported. The identified ethical issues with informed consent, confidentiality, participants' withdrawal procedure, and data security will be identified. The data collection process and the Best Practices in Test Development Instrument will be described. The reliability and validity of the instrument will be discussed.

Chapter 3

Research Methodology

The study examined how nursing instructors grade multiple choice questions using the Best Practices in Test Development Instrument (Killingsworth, 2013). The instrument provides data measuring nursing instructors' use of best practices when developing, testing and revising multiple choice test questions. After analyzing a MCQ test item, nursing instructors can determine if the question is a not a good question or if any distractors are poorly written. If an exam question is deemed to be poor after the exam is administered, nursing instructor can (1) give full points to all students for the question, (2) give full points to those who got the question correct, while giving those who got the question wrong partial points, or (3) maintain the poor question grades, unchanged (Phelps et al., 2013). In the first two decisions, students' grades may be higher than they should have earned, given their understanding of the subject being tested. Most nursing instructors make different decisions depending on analysis, but evaluation of test items may not be done in a consistent manner (McDonald, 2014). Nursing instructors can be clinical practice experts without having experience in assessment of learning (Booth et al., 2016). Due to inconsistencies with grading and grade inflation, students can pass courses when they do understand the concepts taught. This is an ethical concern since the primary goal is assessing the knowledge learned (Watt & Winter, 2017).

The methodology of the study includes research method and design appropriateness, research questions, and hypotheses. The population and sample will be presented and includes the sample size and location. Ethical issues will be addressed with informed consent, confidentiality, participants' withdrawal procedure, and data security.

The data collection process and the Best Practices in Test Development Instrument will be described. Reliability and validity of the instrument will be discussed. Data collection processes and analysis will be described.

Research Method and Design Appropriateness

Qualitative and quantitative research provides evidence for practice and explanations for events and situations. Quantitative research examines data and evaluates the statistics and numbers involved to assess and provide explanations. Qualitative research considers experience or perceptions, and how a situation is experienced by an individual compared to another individual in the same situation. Quantitative research measures specific categorizations while a qualitative method processes sensory impressions and subjective interpretations (Jacobsen, 2017).

Quantitative methods are systematic plans that objectively review data that can be applied to other situations (Boswell & Cannon, 2014). These studies focus on why, where, who, what, when, and how questions about a situation. There are three categories of quantitative research designs. An experimental design has two groups randomly assigned to determine if there is a difference with the one variable that is changed. Quasi-experimental designs are similar to the experimental design without the random assignment or control group (Boswell & Cannon, 2014). Quantitative research designs are appropriate for studies testing theories, for conducting a study that will generate data for evidence-based practice, and for gathering the views, perceptions, and meaning of how people live and interact to the environment around them (Yin, 2015). Deductive reasoning is used with quantitative designs. Quantitative research can be used to test theories (Jacobsen, 2017). Non-experimental designs are correlation, secondary analysis,

meta-analysis, descriptive, and time dimensional because the design cannot exclude extraneous variables and environmental differences (Boswell & Cannon, 2014).

Two commonly used quantitative designs are descriptive and correlational. Descriptive designs focus on characteristics in one sample population and can be used to examine differences in groups, an issue over time, cross-sectional in populations, or relationships between two or more things. This type of study can use documented data that has occurred in the past, however, studies that are current, or prospective studies, are considered more robust with researchers controlling or explaining outcomes (Boswell & Cannon, 2014). Correlational designs can determine relationships, or lack of, between variables. Boswell and Cannon reported this design is the most widely used form of descriptive designs.

Qualitative methods answer questions that explore motivations, perceptions, expectations, understanding of experiences, beliefs, attitudes, and behaviors (Ritholz, Beverly, & Weinger, 2011). Hoe and Hoare (2013) described qualitative techniques as exploring phenomena in relation to how people assign meaning to them. Qualitative research uses inductive reasoning, which evaluates feelings, thoughts, and attitudes about a specific phenomenon. Moustakas (2011) described qualitative research as a way to gain insight into the dynamics of various experiences and perceptions of feelings and thoughts.

The three major types of qualitative research are phenomenology, ethnography studies, and grounded theory. Phenomenology explores the lived experience and seeks to discover feelings and individual perceptions of an event (Jacobsen, 2017). Ethnography research studies cultures and historical research to learn more about the past (Boswell &

Cannon, 2014). Grounded theory uses a process of inductive reasoning to develop theories to explain behaviors (Jacobsen, 2017).

Qualitative data are collected by compiling comprehensive descriptions through interviews, direct observation, and artifacts such as journals (Yin, 2015). The evaluation criteria for qualitative research are based on the concepts of credibility, confirmability, transferability, and dependability (Boswell & Cannon, 2014). Qualitative designs are appropriate when the purpose of a study is to develop themes and to discover, define and capture thoughts, feelings and perceptions at a given time and place. Inductive reasoning is used with qualitative designs. This type of research study is used to build theory or to look at issues that are poorly understood (Boswell & Cannon, 2014).

A quantitative approach was used to obtain information for this study measuring the use best practice in test construction, test item analysis, and test revisions. A survey approach was used to collect data on which components of test construction, test item analysis, and test revisions by nursing instructors report employing. Demographic data such as nurse educator's education, certifications, age, length of time spent in the educational setting, type of program teaching and location of educational facility was obtained. Study data allowed for more analysis of the incidence, distribution and potential relationships and the use of best practice developing, analyzing and revision of multiple choice questions. A quantitative method fulfills the needs of the research questions with the data collected providing statistical data for correlations between data points. The correlational design is appropriate because quantitative research is used for studies testing theories or conducting a study that will generate data for evidence-based practice (Yin,

2015). The conclusions based on the findings can determine if there is a relationship between the variables which is an additional advantage of this design.

Research Question and Hypotheses

A correlational design provides information about the relationships among nursing instructors' use of best practice in classroom test construction, item analysis, and revision and factors in instructor demographic and teaching background. Because there are no clear common guidelines in nursing education to guide test development, item analysis, and test revisions, grading practices in nursing are inconsistent, which can lead to grade inflation (Tarrant & Ware, 2012). There is a gap in the knowledge regarding nursing instructors' decisions making regarding the grading of multiple choice questions (Bristol et al., 2018; Killingsworth, 2013). This study will help identify the different practices currently used by nursing instructors. Other data identified will provide information regarding practices used when testing students at different educational levels. In addition, the educational level of nursing instructors will be reviewed to determine if education background or certification achievement influences the decisions of nursing instructors when evaluating the multiple choice questions.

A research question identifies what is being studied and what information the researcher is searching for. Doody and Bailey (2016) reported research questions arise from theoretical knowledge, previous research, or a practical need for study. Questions can arise from a gap in current literature. Questions can identify the population, dependent variables and design of the research study (Doody & Bailey, 2016). These questions guide the approach, what is being studied, research instrument and ways to

analyze data (Doody & Bailey, 2016). In this study, grading practices of nursing instructors are being reviewed for the use of best practices.

1) What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions?

Null hypothesis: There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

Alternative Hypothesis: There is a relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

2) What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

Null hypothesis: There is no relationship between nursing instructors using best practices in test analysis, type of nursing instructors' education, and emphasis in education.

Alternative Hypothesis: There is a relationship between nursing instructors using best practices in test analysis, type of nursing instructors' education, and emphasis in education.

Population and Sample

The population for the study was registered nurses teaching nursing students in undergraduate registered nursing programs in the United States. The Bureau of Labor Statistics (2016) reported the number of nursing instructors teaching at colleges, universities, and professional schools is 49,370. The data obtained from a sample of participants selected from the larger population can be examined and inferences can be made about the entire population (Hayat, 2013). Hayat (2013) reported determining the appropriate size for a sample of the population is an important consideration. This sample

size was determined by using a power analysis to obtain the statistical number needed in a study for determining a relationship (Boswell & Cannon, 2014).

Power analysis is a mathematical equation developed with classical hypothesis-testing framework to determine an appropriate sample size based on a specified statistical power, confidence level, and effect size (Hayat, 2013). The probability of making a correct inference using the data is the statistical power (Malone, Nicholl & Coyne, 2016). The confidence level refers to the probability of rejecting a null hypothesis and is mathematically shown when the p value is less than the α value (Hayat, 2013). The effect size is the amount of the difference size in groups and is measured as a standard deviation (Malone, Nicholl & Coyne, 2016). The margin of error is another value reviewed and is the maximum difference determined to be allowed between the population and sample (Hayat, 2013). Power analysis is a mathematical inquiry that has a few underlying assumptions that should be considered. The assumptions include that the study sample is assumed to be a simple random sample, power calculations are based on a subjective decision based on the willingness to accept a mistake, and the statistical power is a subjective number placed on correctly detecting an effect (Hayat, 2013).

The sample size was calculated for the study using the power analysis formula $S = (z2 (d(1 - d))/ e2) / 1 + (z2 (d(1 - d)) / e2)$ where S represents the sample size, P is the population size, z refers to the confidence level, e signifies the margin of error, and d denotes the standard deviation (Hayat, 2013). The population of 49,370 nursing instructors was determined using a report by the Bureau of Labor Statistics (2016) on occupational employment statistics of nursing instructors and teachers.

The study used a margin of error of 5 meaning a type I error could occur 5% of time, which is a common choice for researchers (Hayat, 2013). A narrow confidence level of 95% was used as this to more likely to represent the real population value (Hayat, 2013). A z-score of 1.96 was determined using the table below:

Table 2
Confidence level

Desired confidence level	Z-score
80%	1.28
85%	1.44
90%	1.65
95%	1.96
99%	2.58

The ideal sample size for the study was determined to be 382 participants. Typically, 15-20% of potential participants will respond to take a survey, therefore, surveys were sent out to more potential participants. Calculating 20% of those receiving surveys will respond, 0.2 response x 382 needed participants = 1925 surveys to send out. Emails were sent out to deans and directors of programs compiled from individual state's board of nursing web sites and The American Association of Colleges of Nursing.

The participants in the sample will have certain characteristics to be in the group. Inclusion criteria are the set of characteristics participants must have to be included in the sample (Boswell & Cannon, 2014). For the study, the inclusion criteria included nursing instructors teaching the same theory course more than two times, using multiple choice questions in assessments of student learning, and working in a nursing program in the United States. Exclusion criteria are characteristics that would cause elimination from a research sample (Boswell & Cannon, 2014). The exclusion criteria in the study are teaching less than a year, not having not taught a theory class at least twice, not being

involved in the development or evaluation of multiple choice questions, or not teaching in one of the 50 states of the United States.

Informed Consent and Confidentiality

The study was submitted and approved by the Institutional Review Board at the University of Phoenix. Research was conducted using SurveyMonkey©, which gathered data without linking it to any individual personal information. Nursing instructors agreeing to participate in the research first completed the informed consent form before taking the survey. Once the participant read the informed consent and signed the form, the survey screen was available. Any participant who did not want to participate was able to exit the consent form and was not included in the study.

Participants in research studies need to understand their rights as a participant. Informed consent occurs when a study participant understands the purpose, procedures, risks, benefits, alternative procedures, and limits of confidentiality (Boswell & Cannon, 2014). Potential participants were provided information about the study and had an opportunity to ask questions and decide if they wanted to participate in the study. An informed consent form was the first part of the survey. Any participant that did not consent was not able to take the survey.

Anonymity in a research study occurs when information that could identify a single person is encoded, not collected, or removed to protect privacy and provide participants with protection from being identified (St. John et al., 2016). Confidentiality in a research study is the protection of a participant's personal information that is given to the researcher (Jacobsen, 2017). Anonymity protects the person from being identified during and after the study process, while confidentiality is the protection of the

information about participants given to the researcher. During this study, surveys were sent out to 1935 participants, as 382 responses were needed. The survey requested personal information on workplace, geographic area, work status, degree obtained, ethnicity, and academic qualifications. This information was kept confidential to prevent participants from being identified. The data that could identify an individual was stored in the survey program data. The data that was downloaded was kept in a locked file in a locked office. The surveys were administered using SurveyMonkey© program. SurveyMonkey© provides tools that help set up surveys and allows researchers to collect data while keeping participant information anonymous. A feature of SurveyMonkey© is the Secure Sockets Layer used to maintain privacy by encrypting information collected (SurveyMonkey©, 2017).

Participants were able to withdrawal from the research study up until the survey was submitted. During the survey, participants could decide not to continue participating and could close out of the survey. Incomplete surveys were not used in data analysis. Incomplete surveys were defined as surveys without responses to all questions. Once surveys were completed and submitted, the participant was unable to withdraw from the research study. The data were submitted anonymously and the participant was not identified, therefore, to withdraw after submitting was not possible. The inability to withdraw after completing all survey questions was highlighted in the informed consent.

Participant information, such as name, address and social security numbers were not requested during the survey. The information on age, education, and type of degree program in which the individual is teaching was be gathered. The information was kept in

the Survey Monkey© program and was only disclosed in data formation. The information was not linked, nor could it be linked, to any individual participant.

Data were kept in the survey monkey program© and was only disclosed in numerical data. No personal information could be linked to individual participants. Data printouts were kept in a locked file that is located in a locked office room. The data will be kept for three years. After three years, the data will be shredded. Data stored on SurveyMonkey© is physically located in the United States in data centers with physical security of continual monitoring using cameras, visitor logs, and entry requirements (SurveyMonkey, 2017).

Instrumentation

A survey instrument is a tool consisting of a series of questions used to gather information from participants (Jacobsen, 2017). The Best Practices in Test Development Instrument (Killingsworth, 2013) was used for the study. Permission was obtained from tool developer, Erin Killingsworth, to use Best Practices in Test Development Instrument. This tool was used by researchers to ask questions about the components of test construction, test item analysis, and test revision. Each component is scored on a scale from 1 to 5 with 1 being not at all and 5 being all the time. There are 12 components that can be used during test construction. The score ranges from 12 to 60 with the higher the score, the greater use of best practices. The components for test instruction are course objectives, class or unit objectives, major content topics, specific content topics, test blueprint, the NCLEX-RN test plan, peer review of test items, higher cognitive levels according to Bloom's taxonomy, clinical context for test items, plausible distractors in multiple-choice test items, even distribution of correct answer in multiple-choice options,

and use various test item types (Killingsworth, 2013). Test item analysis has 6 informational items that instructors can obtain after a test is administered. The possible score ranges from 6, meaning not used at all, to 30 indicating greater use of best practices. Test items components include number of students who answered each question incorrectly, number of students who answered each question correctly, question's ability to discriminate between the high and low scoring students, frequency of distractor choices with each test question, discrimination between the high and low scoring students choosing distractors, and the central tendency of the student grades on the exam. Test revision includes 10 actions instructors can perform when performing revisions. This score is from 10 to 50 with the higher scores indicating greater use of best practices. Using item analysis data, comparing item analysis data for test questions, using distractor discrimination to revise test items, using difficulty level of test items to revise test items, assessing for linguistic/cultural bias in test items, assessing for changes in domain content, assessing for outdated language used in test items, changing test items to ensure test security, changing test items to reflect emphasis in classroom content, and changing test items to ensure sufficient sampling of content are components of test revision. The higher the score, the more components of test construction, test item analysis, and test revision best practices are used by the participant.

The last set of 14 questions on the instrument identify demographic and teaching background information including the type of nursing programs worked in, education received, geographic location, full or part-time status, the amount of course work taken in the education field, age, gender, participation in a professional test development program, and if participant is a certified nurse educator.

Validity and Reliability

Reliability occurs when an instrument consistently measures the same thing (Boswell & Cannon, 2014). Jacobsen (2017) described reliability occurring when measurements are repeated and the same results are achieved. A pilot study was conducted to assess reliability of the study instrument (Killingsworth, 2013). Thirty-four BSN nursing faculty members from six nursing programs participated in the pilot study. Best practices in test construction, test item analysis, and test revision, and scoring methods using Likert scale and dichotomization method with the results of 0.73-0.77, which is adequate (Killingsworth, 2013). Although the Likert scale was deemed acceptable, it was used as a larger sample reliability would be higher. (Killingsworth, 2013). The participants provided feedback on the survey. The pilot study feedback confirmed a completion time of 20 minutes (Killingsworth, 2013).

An instrument is valid if it measures what it claims it measures (Boswell & Cannon, 2014). During research, internal and external validity are reviewed to confirm an appropriate instrument is being used. The internal validity refers to whether the effects observed in a study are due to the manipulation of the independent variable and not some other factor. The Best Practices in Test Development tool was developed with the guidelines recommended by Tarrant and Ware in 2012 (Killingsworth, 2013). The tool developer reported the Cronbach's alpha was greater or equal to .70 on test construction and test analysis, with the test revision scoring .61 (Killingsworth, 2013). Interval and ratio level study variables were normally distributed for test construction. Test analysis and test revision had non normal distributions and after transforming variables, near normal distributions were obtained (Killingsworth, 2013). Interrater agreement of .86

with .85 content validity index for relevance, and clarity was .80 with the portions developed for surveying participants about demographics, credentialing, teaching experience, and the nursing program (Killingsworth, 2003). Killingsworth reported studies provide evidence of validity for the questions used from the Ethical Climate Questionnaire (2003). Evaluation of Learning Advisory Council tool's validity was evaluated by 15 faculty members to confirm validity (Killingsworth, 2003). Cronbach's alpha is a test used to test internal reliability in questionnaires (Jacobsen, 2017). Killingsworth (2013) reported an adequate internal consistency of the instrument.

A research study can be generalized to a similar population. One way to achieve this is to use random sampling of subjects. A power analysis was performed to determine the minimal number of participants needed to achieve a good sample of the population. The sampling was sent out to different geographical areas to have a random sample from different areas. The validity of the survey tool used in this study were reviewed by 15 nursing instructors to confirm validity (Killingsworth, 2013).

Data Collection

Recruitment for surveys consisted of informed consent and survey link to be sent to nursing institutions located in the United States. Names of nursing programs with full accreditation were obtained from the board of nursing of each state. Information and a request for nursing instructors to participate was emailed to each nursing program's dean or director. Data were collected until the number of participants needed to achieve statistical analysis was achieved. After a goal of 400 completed surveys was reached, the survey was closed.

Data Analysis

Completed survey data were taken from SurveyMonkey and transferred into SPSS for data analysis. Descriptive statistics, scatterplots, and multiple linear regressions were used to address the research questions. Statistical analysis provided information used to determine if there could be a relationship between two or more variables. The relationship could be a positive or negative. Data can be assessed to determine if one variable relationship corresponds with other variables. These relationships can help determine whether best practices in test development, item analysis, and test revisions are being implemented by nursing instructors.

Analyzing data for correlational relationship requires a two-step process. First, there must be a linear relationship between the variables (Laerd Statistics, 2015). Using a scatter plot, the data can be entered onto a plot to observe the pattern. The pattern noted on the scatterplot should show minor variances in the data with no major outliers and should be homoscedastic. When data is dispersed and there is no order or correctness, this is called homoscedasticity (Laerd Statistics, 2015). The data points should be in a general diagonal direction that is not perfectly aligned. The data is then analyzed using a multilinear. This analysis reviews the variation in the dependent variable explained by the independent variables, can predict dependent variable values based on new independent variables values and determines the dependent variable change for a one portion change in the independent variables (Laerd Statistics, 2015). This value measures the strength of the association between the variables. The R^2 value represents the variance from the independent values. This value is then adjusted to correct positive bias which can be applied to the entire population (Laerd Statistics, 2015). The adjusted R2 percent is the

strength of the relationship and according to Cohen (1988) under 0.4 is no or small effect, 0.5 is a medium/ moderate effect and 0.8 represents a strong or large effect.

The data were collected using a self-reported survey. A Likert scale of 1 to 5 was used to identify the use of best practice in test construction, test analysis, and test revision. The data obtained were used to answer each research question. The data were analyzed using SPSS program to determine the response to the research questions.

Research question 1: What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions?

Data was entered in SPSS and multiple linear regression test was completed to determine if a positive correlation is identified. The data was plotted on a graph to show how the data relate to each other. A positive correlation suggests that there is a relationship between the two variables but does not identify the type of relationship (Jacobsen, 2017).

Research question 2: What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

Data was entered in SPSS and multiple linear regression test was completed to determine if a positive correlation is identified. The data was plotted on a graph to show how the data relate to each other. A positive correlation suggests that there is a relationship between the two variables but does not identify the type of relationship (Jacobsen, 2017).

What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

Table 3
Data analysis

Variable	Level of measurement	Statistical analysis
Research question 1		
best practices in test analysis	interval	multiple linear regression
best practice in test construction	interval	multiple linear regression
best practice in test revision	interval	multiple linear regression
Research question 2		
best practices in test analysis	interval	multiple linear regression
nursing instructors' education	interval	multiple linear regression
educational emphasis in education	interval	multiple linear regression

Summary

A quantitative correlational design was used to determine the relationships of grading practices, use of evidence-based practice, educational levels of nursing instructors, type of educational institution, and use of item analysis. Nursing instructors must evaluate undergraduate students to determine if the student has mastered the material and can apply that information. With the study, grading practices related to nursing instructors' methods for developing evidence-based exam items were examined to explore nursing instructors' use of evidence-based practices when developing evaluations and grading multiple choice questions. The results of the study identified current practices and can help educators formulate plans for consistent ethical grading practices in the future to prevent future grade inflation. Information on how the study was conducted was provided. The development of Best Practices in Test Development Instrument (Killingsworth, 2013) was reviewed along with how the tool was validated. Chapter 4 will present the research findings.

Chapter 4

Analysis and Results

This correlational study examined the use of best practice in classroom test construction, item analysis, and revision in nursing programs in the United States. The specific problem of this study addressed how best practice in classroom test construction, item analysis, and revision is used by nursing instructors. The findings of this study identified relationships of best practices currently used by nursing instructors. Other data identified provided information regarding practices used when testing students at different educational levels. In addition, the educational level of nursing instructors was reviewed to determine if education background or certification achievement influenced the decisions of nursing instructors when evaluating the multiple choice questions.

Chapter 4 examines the research performed using the Practices in Test Development Instrument. Best practices were analyzed to determine the relationship between test construction and test revision along with faculty demographics. A review of the sample and hypothesis testing will be provided. Demographics of the participants will be described. The sample size for the study will be reviewed. The hypotheses of the research study will be presented and the data collected will be analyzed. The chapter will conclude with the findings from the research study.

Research Questions/Hypotheses

1) What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions?

Null hypothesis: There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

Alternative Hypothesis: There is a relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

2) What is the relationship between nursing instructors using best practices in test analysis, type of nursing instructors' education, and emphasis in education?

Null hypothesis: There is no relationship between nursing instructors using best practices in test analysis, type of nursing instructors' education, and emphasis in education.

Alternative Hypothesis: There is a relationship between nursing instructors using best practices in test analysis, type of nursing instructors' education, and emphasis in education.

Data Collection

Nursing instructors examine students using multiple choice questions to determine student mastery of course material. In order to explore best practices in test construction, item analysis and revision of multiple choice items, study participants included nursing instructors teaching undergraduate nursing theory for more than two courses and using multiple choice assessments. One thousand eight hundred fifty-six survey invitations were sent to deans and directors of nursing programs. Twenty-five emails were identified as undeliverable and 16 responses were returned with the request to opt out of the study. Four hundred and nine nursing instructors answered the survey.

A sample of the population of nursing instructors was used to gather data in order to make inferences of the practices used by the entire population. An analysis of the population was performed to determine the appropriate sample size. The sample size is important since a sample size to big is more complex and costly, while a small group could cause data analysis to miss association between variables. Random sampling was

employed to allow inclusion of the entire population being studied which allows for unbiased data and a better representation of the population. The power analysis determined 382 participants were needed for the survey. A total of 409 nursing instructors completed the survey.

Demographics

More females than males responded to the survey. This is indicative of the majority of nursing educators being female. Participants ranged in age from 26 to 79 years with an average age of 54 years. The majority of respondents identified themselves as white. This corresponds with the data of this population. Ninety-five percent of respondents are full time instructors. Only a quarter of nursing faculty were certified in nursing education (see Appendix C).

Participants were from 46 out of 50 states. The highest number of participants were from Texas. California had the second largest number of participants with Louisiana having the third largest number (see Appendix D).

Data Analysis

The research study used Best Practices in Test Development Tool to collect data about the use of best practices in test construction, item analysis, and revision. The tool comprised of 28 items; 12 in test construction, six in item analysis, and ten in revision. A Likert scale from one to five was used; one was not at all, 2 was 25% of the time, 3 was 50% of the time, 4 was 75% of the time and 5 was all the time. Results revealed that best practices in test construction was used 75% of time with scores between 3.23 for peer review of test items to 4.76 for major content topics covered. Items identified as used less than 75% of the time include using a test blueprint, the NCLEX-RN® test plan, peer

review of test items, even distribution of correct answer in multiple-choice options, and various test item types. Class or unit objectives, major content topics, and specific content topics were the three highest used (see Table 4).

Table 4
Test construction

	Mean
Course objectives	4.23
Class or unit objectives	4.50
Major content topics	4.76
Specific content topics	4.61
A test blueprint or table of specifications	3.83
The NCLEX-RN® test plan	3.59
Peer review of test items	3.23
Higher cognitive levels according to Bloom's taxonomy	4.24
Clinical context for test items	4.04
Plausible distractors in multiple-choice test items	4.35
Even distribution of correct answer in multiple-choice options	3.81
Use various test item types	3.88
Mean of all items	4.09

Reviewing the individual responses to the items in best practices for test construction, the numbers of individuals reporting in the top three items had a percentage of 89.5% of respondents or higher. Those areas noted to be used less than 75% of the time were reviewed. Forty-five percent of participants reported using the NCLEX-RN® test plan review during test construction. The respondents using peer review of test items 50% of the time or less was 54.1% (see Table 5).

Table 5
Test construction, percentage reporting use

	Percentage of Time				
	0%	25%	50%	75%	100%
Course objectives	5.1	6.4	7.1	23.5	57.9
Class or unit objectives	2	4.2	4.4	21.3	68.2
Major content topics	1.5	1	1.5	12.2	83.9
Specific content topics	1.5	.5	5.4	20.8	71.9
A test blueprint or table of specifications	13	7.3	12.2	18.3	49.1
The NCLEX-RN® test plan	12.7	9	18.1	26.4	33.7
Peer review of test items	15.4	19.6	19.1	18.3	27.6
Higher cognitive levels according to Bloom's taxonomy	1	2.9	10.3	42.8	43.0
Clinical context for test items	2.2	3.9	14.9	45.5	33.5
Plausible distractors in multiple-choice test items	.5	2.7	8.8	36.9	51.1
Even distribution of correct answer in multiple-choice Options	11.7	7.1	12.2	26.2	42.8
Use various test item types	5.6	16.1	10.5	20.3	47.4

Best Practices in Item Analysis

Best practice in item analysis was used 75% of the time with the scores between 3.75 for item discrimination to 4.75 for number answering question correctly. Frequency of distractor choices with each test question, discrimination between the high and low scoring students choosing distractors, and central tendency of the student grades on the test were used less than 75% of the time. The highest scoring items included using the number of students who answered each question incorrectly, the number of students who answered each question correctly, and item analysis data to deciding to remove test questions before finalizing test scores (see Table 6).

Table 6
Test analysis

	Mean
The number of students who answered each question incorrectly.	4.75
The number of students who answered each question correctly	4.70
A question's ability to discriminate between the high and low scoring students	4.30
The frequency of distractor choices with each test question.	3.89
The discrimination between the high and low scoring students choosing distractors	3.75
The central tendency of the student grades on the test.	3.95
Use item analysis data when determining to keep or eliminate test questions before finalizing test scores.	4.68
Mean of all items	4.29

The individual responses to reviewing best practice in item analysis were examined. The number of nursing faculty responding to this survey self-reported using the number of students who answered a question wrong and the use of discrimination between high and low scoring students is concerning. Eighty-five percent of faculty reported they look 100% of the time at the number of students that answered a question incorrectly, but only 67% look at the between high and low scoring students' discrimination 100% of the time. When determining to keep or discard a question, decisions based on incorrect answers does not discriminate between the high and low scoring students. If 75% of students answer a question wrong, this does not mean the question is a bad question. If 25% of the students answering correctly are the high scoring students, this question could be a valid one. When analyzing test items, nursing instructors reporting looking at the discrimination between high and low scoring students choosing distractors and the central tendency of student grades all the time was only 51.1% and 48.9% respectively. These are important data points to consider when determining if a test question is valid (see Table 7).

Table 7
Test item analysis, percentage reporting use

	Percentage of Time				
	0%	25%	50%	75%	100%
The number of students who answered each question incorrectly.	1.7	2	1.7	9.3	85.3
The number of students who answered each question correctly (difficulty level or p-value).	2.4	1	2.9	11	82.6
A question's ability to discriminate between the high and low scoring students	7.6	3.2	8.3	13.9	67
The frequency of distractor choices with each test question.	11.7	5.9	14.9	16.4	51.1
The discrimination between the high and low scoring students choosing distractors	17.4	5.1	12	16.6	48.9
The central tendency of the student grades on the test.	11.7	5.1	13	16.9	53.3

Best Practices in Test Revision

Best practice in test revision was used 75% of the time with the scores between 3.38 for assessing for linguistic/cultural bias in test items to 4.68 for using item analysis data when determining to keep or eliminate test questions. Comparing item analysis data for test questions used from one term to another and changing test items to reflect emphasis in classroom content were identified as being used 75% of the time. Distractor discrimination to revise test items, assessing for linguistic/cultural bias in test items, and assessing for changes in domain content based upon new research data were used less than 75% of the time (see Table 8).

Table 8
Test revision

	Mean
Use item analysis data when determining to keep or eliminate test questions before finalizing test scores.	4.68
The central tendency of the student grades on the test.	3.95
Compare item analysis data for test questions used repeatedly from one term to another.	4.06
Use distractor discrimination to revise test items.	3.63
Use difficulty level of test items to revise test items.	3.91
Assess for linguistic/cultural bias in test items.	3.38
Assess for changes in domain content based upon new research data.	3.45
Assess for outdated language used in test items.	3.90
Change test items to ensure test security.	3.95
Change test items to reflect emphasis in classroom content.	4.28
Change test items to ensure sufficient sampling of content.	4.15
Mean of all items	4.33

Test revision is not being conducted by nursing instructors all of the time with linguistic/cultural bias changes that occur in domain content based upon new research data. Only 28.9% of participants reported reviewing material for updates based on new research and 30.1% reviewing tests for language and cultural biases all the time (see Table 9). With the fast-paced research in healthcare and the changing demographics of students, these two items are important to test reliability and validity.

Table 9
Test revision, percentage reporting use

	Percentage of time				
	0%	25%	50%	75%	100%
Use item analysis data when determining to keep or eliminate test questions before finalizing test scores.	1	1.5	4.4	14.7	78.5
Compare item analysis data for test questions used repeatedly from one term to another.	5.9	7.1	12.7	24	50.4
Use distractor discrimination to revise test items.	13	6.8	16.6	31.8	31.8
Use difficulty level of test items to revise test items.	4.6	4.9	21	33.7	35.7
Assess for linguistic/cultural bias in test items.	15.4	13.7	18.1	22.7	30.1
Assess for changes in domain content based upon new research data.	14.2	10	20.8	26.2	28.9
Assess for outdated language used in test items.	7.6	9	14.9	22.7	45.7
Change test items to ensure test security.	2.4	8.6	21	27.6	70.3
Change test items to reflect emphasis in classroom content.	1.5	3.4	13.2	29.8	52.1
Change test items to ensure sufficient sampling of content.	2	5.4	15.4	30.3	46.9

Results

The purpose of this correlational study was to examine the relationship between best practice and the reality of practice in classroom test construction, item analysis, and revision in nursing programs in the United States. There were two research questions for the study with the following hypotheses.

What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions?

$H1^0$: There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

$H1^a$ There is a relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions

What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

$H2^0$: There is no relationship between factors in nursing instructor demographics and educational background and nursing instructors use of best practice in test construction, test analysis and test revisions.

$H2^a$: There is a relationship between factors in nursing instructor demographics and educational background and nursing instructors use of best practice in test construction, test analysis and test revisions

Research questions were tested using correlational statistics to evaluate the data. Correlational research is performed to determine relationships between two or more variables within the same population using characteristics and analyzing associations. Relationships of direction and strength can be noted, but this does not determine the exact causation. A scatterplot was performed to establish a possible relationship in the data shown with a linear plot.

A multiple correlational analysis was performed to evaluate the relationship between the variables using the correlation coefficient R to measure the strength of that association. Once relationships are determined, then the hypotheses can be reviewed to determine which is indicated. If no relationship is determined, then the null hypothesis is supported. When the results are positive, the alternative hypothesis is accepted.

Research Question 1

What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions? A scatter plot revealed

a linear relationship with a general diagonal pattern, minor variances, and no major outliers (see Figure 2).

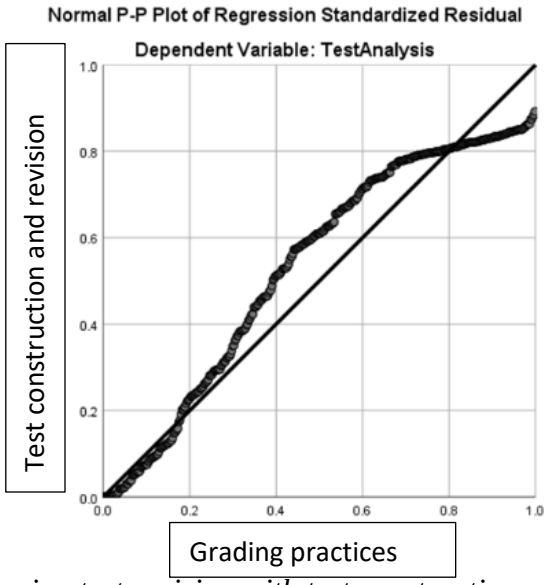

Figure 1. *Linear regression test revision with test construction*

Table 10

Multiple linear regression test revision with test construction

Model	R	R Square	Adjusted R Square	Std. Error Estimate	p value
	.107	.012	.007	5.284	0.05

Predictors: (Constant), Test Revision, Test Construction

The data were further analyzed with a multiple linear regression. The effect size is 0.007 showing little to no effect on the dependent variable (see Table 6). Although a relationship was noted, the effect level was minimal. There is no correlation due to a low variance between grading practices the use of best practices in test construction, test analysis and test revisions. The findings reveal that a nursing instructors' grading practices are not dependent on use of best practice in test construction, test analysis and test revisions. The null hypothesis was supported: there is no relationship between factors

in nursing instructor demographics and educational background and nursing instructors use of best practice in test construction, test analysis and test revisions.

Research Question 2

What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

A scatter plot revealed a linear relationship. A relationship was noted with the general diagonal pattern shown along with minor variances and no major outliers (see Figure 3).

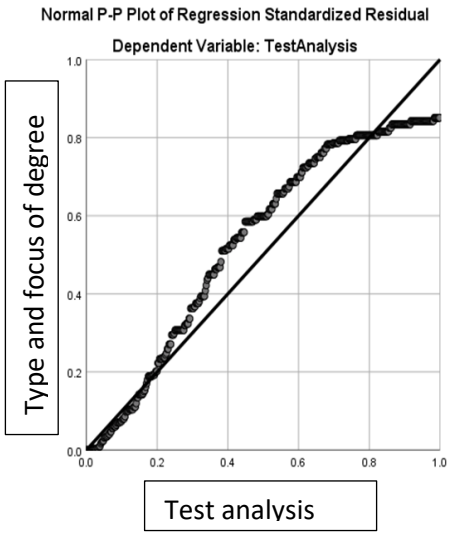

Figure 2 *Linear regression test revision with test analysis*

Table 11

Multiple linear regression test revision with test analysis

Model	R	R Square	Adjusted R Square	Std. Error Estimate	p value
	.075	.006	.001	5.300	0.05

Predictors: (Constant), high degree, academic degree with an emphasis education
Dependent Variable: Test Analysis

The data were further analyzed with a multiple linear regression. The effect size is 0.001 showing no substantial effect on the dependent variable (see Table 11). Although a relationship was noted, the effect level was minimal. There is no significant correlation due to a low variance between nursing instructors' use of best practices in test analysis and the educational preparation of the educator. The analysis indicated that there was no relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator. There is no relationship between factors in nursing instructor demographics and educational background and nursing instructors use of best practice in test construction, test analysis and test revisions.

Chapter Summary

The results of the research study using the Best Practices in Test Development Instrument were presented. The sample size for the study and demographic data were reviewed. Correlational relationships between best practices in item analysis were not found with best practices in test construction and test revisions. No correlation was noted with best practices in test analysis and the educational preparation of the educator. Chapter 5 will review the findings and recommendations presented by this research. The research methodology will be reviewed. Research questions and corresponding hypothesis will be discussed. Implications of the study will be presented along with recommendations for education, practice and further research.

Chapter 5

Conclusions and Recommendations

There are no clear common guidelines in nursing education to guide nursing instructors when making decisions in test construction, item analysis, and test item revisions. This chapter will review the quantitative research methodology. The research questions and hypotheses will be presented together with the implications of the research findings. Recommendations will be made for practitioners and further research studies will be discussed.

Research Purpose, Question, and Hypotheses

The purpose of the correlational study was to examine the relationship between best practice and the reality of practice in classroom test construction, item analysis, and revision in nursing programs in the United States.

1) What is the relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions?

Null hypothesis: There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

2) What is the relationship between nursing instructors using best practices in test analysis, and the educational preparation of the educator?

Null hypothesis: There is no relationship between nursing instructors using best practices in test analysis, type of nursing instructors' education, and emphasis in education.

Discussion of Findings

The research study used Best Practices in Test Development Tool (Killingsworth, 2013) to collect data about the use of best practices in test construction, item analysis,

and revision. A Likert scale from one to five was used; one was not at all, 2 was 25% of the time, 3 was 50% of the time, 4 was 75% of the time and 5 was all the time. Results revealed that best practices were used 75% of time with scores between 3.23 and 4.76 for test construction, 3.75 to 4.75 for item analysis, and 3.63 to 4.68 for test revision.

Research Question 1

Correlational relationships between best practices in item analysis were not found with best practices in test construction and test revisions. The survey data showed a statistical relationship in the use of the three areas of best practices by nursing instructors. Data revealed that all three areas of best practice were used 75% of the time by instructors. This study reveals a trend of more instructors using best practices in testing practices. The past studies by Killingsworth (2013) reported the use of best practices in test construction 75.7%, item analysis 78.5%, and test revision 70% of the time. The first study performed in 2009 by Oermann, Saewert, Ika, and Yarbrough reported test construction items used less than 75% of the time. Although more nursing instructors report using best practices, the strength of the relationship in the data is not statistically significance to a relationship.

The use of best practice in test construction and revisions does not mean the nursing instructors use best practices with item analysis. Item analysis is important in determining the quality of a test question. Nursing instructors have the opportunity to use the analysis to improve assessments and identify areas of weakness. The use of item analysis is also used to measure the reliability of the test items used in student assessments. There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions. The first research

question analysis showed the null hypothesis was supported. There is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions.

Research Question 2

The second research question was about a relationship between factors in nursing instructors use of best practices in test analysis and the educational preparation of the educator. The research analysis does not support a statistically significance relationship between factors in nursing instructor demographics, educational background and the use of best practice in test construction, test analysis and test revisions. The null hypothesis was supported. There is no relationship between nursing instructors use of best practices in test analysis, type of nursing instructors' education, and emphasis in education.

This research supports previous studies by Oermann and colleagues (2019) and Killington (2013) on use of best practices. The current research shows a lack of education on developing and grading nursing examinations (Booth et al., 2016; Docherty & Dieckmann, 2015; Salminen et al., 2013). Inconsistencies in grading practices was also prevalent in the current nursing literature (O'Flynn-Magee & Clauson, 2013; Pazargadi, Ashktorab, & Khosravi, 2012; Salminen et al., 2016). In a study by Reynolds (2015) half the faculty surveyed reported not having education on grading. Lack of education on grading practices could cause these findings.

The overall limitation on this research was the potential of participants to answer how they think they should be using best practices or how participants believe they use best practices. A research study that includes exam questions and test item analysis for participants could provide data not influenced by self-reported limitations. Disseminating

more information on education and testing practices could be helpful in understanding factors in these results.

Limitations

During the research, the surveys were sent out to directors and deans of registered nursing programs to distribute to the nursing faculty. Upon review of this process, a potential limitation was noted. The distribution of surveys could be skewed depending on who was given the survey. This could explain the low numbers of part time faculty respondents. Academic institutions are imposing requirements for research performed. Some institutions require advanced notice of potential research studies and require the institution's permission to participate in studies from outside sources. This precluded many potential respondents from participating in the research.

The predetermined sample size for the survey was 382 participants. Four hundred and nine nursing instructors responded. Even though the sample size was reached, this may not be a representative sample. Nursing instructors who received the survey had a choice of participating. Those who responded could have had a higher interest in the use of best practices, be more knowledgeable about best practices with the use of multiple choice questions, or could be interested but unknowledgeable about best practices in test construction, test analysis and test revisions. These factors could interfere with the accuracy the results. Survey answers were self-reported. Participants might rate their use of best practices at a different level than what is actually used. The descriptions of best practices were provided. These can possibly be interpreted differently by various nursing instructors. This would alter the result of the data submitted.

The limitations identified can be considered in the next research endeavor. Recommendations to leaders and practitioners will address this area of need. Education in grading practices is recommended for all nursing educators. Education on grading practices should include writing test items and not using test banks. Research shows that test bank questions should not be used due to low quality of questions (Booth et al., 2016). Mentoring new faculty, particularly in grading practices, is needed. Administrators and lead instructors should be mindful of new faculty needs for assistance in this area.

Recommendations for Leaders and Practitioners

The research results can provide insight into the use of best practices and the prevalence of the inconsistencies in test construction, item analysis, and revision by nursing instructors, specifically focusing on assessments using multiple choice questions. A review of current literature has shown a gap in knowledge on the use of evidence-based practice and the development and evaluation of nursing multiple choice questions. There were two research studies on nursing instructor grading practice conducted in the past. Reviewing the results confirmed the inconsistencies in self-reported use of best practices. Nursing educators and educational administrators should be aware of these discrepancies and implement in-services and policies to reflect the need for consistent grading practices and use of best practices in MCQ test item construction, test item analysis and test item revisions.

There are variances in the use of best practices and educational background. The current trend is for nursing educational institutions to hire nurse practitioners and clinical experts to teach in their programs. This study and previous studies by Booth and associates (2016), Cooley and De Gagne (2016), and Schoening (2013) revealed the need

for nursing instructors trained in educational pedagogy. More specialized nurse clinicians are becoming nursing instructors, but lack knowledge and preparation for the role of nurse educator (Cooley & De Gagne, 2016). Booth and associates (2016) reported clinical expertise does not include knowledge of evidence-based research and practice, teaching methods, or curriculum design and development. Without this knowledge, development of appropriate MCQs is difficult. Schoening (2013) found 63% of nursing instructors had no formal preparation for teaching. Effective mentoring of new nursing instructors along with workshops on MCQ item construction, analysis and revisions would be beneficial.

A review of MCQs used in a nursing school and textbooks revealed an alarming rate of poor-quality items (Booth et al., 2016). Nursing institutions and deans/ directors of nursing programs need to be aware of current research on education of their students. Those in nursing academia need to have an educational background with test development instruction. Policies are needed to ensure best practices are being used, including areas reported to be underused such as peer review of test items and item analysis after test after administration. The level of difficulty and use of application and analysis questions should be stressed. The development of guidelines for use of best practices and resources to enhance and support the use of best practices in MCQ item construction, analysis and revisions.

Recommendations for Future Research

Further research could be conducted exploring the use of best practices in test construction, item analysis, and revision by nursing instructors. The discrepancies in self-reported use of best practices and the reports of poor performing MCQs need to be

investigated. There are several methods that could be used to explore this discrepancy. A research study providing the subjects with the same test items and item analysis could provide information on the consistency between instructors. The findings of this interrater reliability study could provide insight into how much instructors agree on individual assessment findings while confirming the self-reported use of best practices in analysis and revisions in nursing assessments.

Further research can focus on efficacy of faulty development programs regarding the use of best practices for grading practices and use of best practice in MCQ item construction, analysis, and revisions. Educational programs are needed to assist nursing instructors to maintain current in educational pedagogies and testing strategies. Faculty development programs are needed to assist nursing instructors to learn, understand and use best practices in testing. A quantitative research study could be performed using a survey to determine if faculty have a development program that addresses the use of best practice for examinations.

Research on policies and procedures on grading is necessary. A lack of clear guidelines for testing practices has been identified in the literature (Hicks, 2011; Killingsworth, Kimble, & Sudia, 2015; Oermann & Gaberson, 2014). Clear guidelines and recommendations for testing can enhance resources available to nursing instructors. Development of these can assist nursing instructional institutions to effectively assess student learning.

Further research is needed to explore the effect of grade inflation in nursing programs. Nursing instructors use different methods for how they grade with a MCQ that is identified as poor quality. Many of these methods inflate student grades. The studies

reviewed provided valuable information on how the different techniques can change the scores of an exam. In the research, no method was identified as a best practice for grading multiple choice questions. A mixed method study could explore the various ways nursing instructors deal with poor quality exam questions

This research supports the findings in the literature. Literature review revealed grading practices of nursing instructors are influenced by a lack of clear guidelines for assessment practices and a lack of educational pedagogies knowledge (Booth et al., 2016; O'Flynn-Magee & Clauson, 2013; Pazargadi, Ashktorab, & Khosravi, 2012). Grade inflation and discrepancies in grading have been identified. Multiple choice questions are used by nursing programs to assess student learning. Properly written MCQs can evaluate higher level thinking and evaluate multiple concepts at a time. Item analysis can be used to evaluate the effectiveness the MCQs. Nursing instructors need to understand and utilize these techniques to develop appropriate student assessments. Grading strategies need to be understood and used according best practices to prevent unintentional grade inflation (Caruth & Caruth, 2013; King-Jones & Mitchell, 2012; O'Halloran & Gordon, 2014); Paskausky & Simonelli, 2014); Smith & Fleisher, 2011).

Summary

Chapter 5 reviewed the research findings. The research explored the use of best practices. The first question revealed there is no relationship between nursing instructors' grading practices and use of best practice in test construction, test analysis and test revisions. The second question revealed no relationship nursing instructors using best practices in test analysis, and the educational preparation of the educator. Implications for practitioners were offered. The implications of the research findings were presented.

Further research into the process of developing nursing assessments is needed. The need for nursing instructor education in curriculum and assessments is needed for nurses entering academia from a clinical based practice.

References

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189-209. doi:http://dx.doi.org/10.1037/edu0000282

Almala, A. H. (2005). A constructivist conceptual framework for a quality e-learning environment. *Distance Learning, 2*(5), 9-12. Retrieved from https://search-proquest-com.contentproxy.phoenix.edu/docview/230696773?accountid=35812

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2011). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, S. (2019). Best (but oft forgotten) practices: Sample size planning for powerful studies. *American Journal of Clinical Nutrition, 110,* 280–295. https://doi.org/10.1093/ajcn/nqz058

Bada, S. (2015). Constructivism Learning Theory: A paradigm for teaching and learning. *IOSR Journal of Research & Method in Education, 5*(6), 66-70 doi: 10.9790/7388-05616670

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 Science, 30*(1), 3-6. doi: http://dx.doi.org/10.12669/pjms.301.4458

Bailey, P., Mossey, S., Moroso, S., Cloutier, J., & Love, A. (2012) Implications of multiple-choice testing in nursing education, *Nurse Education Today, 32*(6), e40-e44. doi: 10.1016/j.nedt.2011.09.011.

Barnard, C. J. (2013). The marking of multiple choice questions. *South African Journal of Science and Technology, 32*(1), 402-409. doi: 10.4102 /satnt.v32i1.402

Bauer, B., Holzer, M., Kopp, V., & Fischer, M. (2011). Pick-N multiple choice-exams: A comparison of scoring algorithms. *Advances in Health Sciences Education, 16*, 211–221. doi:10.1007/s10459-010-9256-1

Begum, T. (2012). A guideline on developing effective multiple choice questions and construction of single best answer format. *Journal of Bangladesh College of Physicians & Surgeons, 30*(3), 159. http://dx.doi.org/10.3329/jbcps.v30i3.12466

Billings, D. M., & Halstead, J. A. (2016). *Teaching in nursing: A guide for faculty* (5th ed.). St. Louis, MO: Elsevier.

Boone, E., Meiners, S., Frankland, L., Laursen, J., & Colombo, R. (2019). A comparison between fixed and random sampling of a low density spotted bass population in a large river. *Journal of Freshwater Ecology, 34*(1), 533-540. doi: 10.1080/02705060.2019.1631222

Booth, T. L., Emerson, C. J., Hackney, M. G., & Souter, S. (2016). Preparation of academic nurse educators. *Nurse Education in Practice, 19*, 54-57. http://dx.doi.org/10.1016/j.nepr.2016.04.006

Boswell, C., & Cannon, S. (2014). *Introduction to nursing research* (3rd ed.). Burlington, MA: Jones & Bartlett Publishers.

Bowen, R. E. S., Grant, W. J., & Schenarts, K. D. (2015). The sum is greater than its parts: Clinical evaluations and grade inflation in the surgery clerkship. *American Journal of Surgery, 209*(4), 760–764. doi:10.1016/j.amjsurg.2014.10.023

Bristol, T. J., Nelson, J. W., Sherrill, K. J., & Wangerin, V. S. (2018). Current state of

test development, administration and analysis: A study of Faculty Practices. *Nurse Educator, 43*(2), 68-72. doi: 10.1097/NNE.0000000000000425.

Bureau of Labor Statistics. (2016). *Occupational employment and wages, May 2016: Nursing instructors and teachers, postsecondary*. Retrieved from https://www.bls.gov/oes/current/oes251072.htm#nat

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education, 40*(2), 218-231. doi:10.1080/02602938.2014.902192

Caruth, D. D., & Caruth, G. D. (2013). Grade inflation: An issue for higher education? *Turkish Online Journal of Distance Education, 14*(1), 102-110.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: NY: Psychology Press.

Cooley, S. S., & De Gagne, J. C. (2016). Transformative experience: Developing competence in novice nursing faculty. *Journal of Nursing Education, 55*(2), 96-100. doi:http://dx.doi.org/10.3928/01484834-20160114-07

Curtis, E. A., Comiskey, C., & Dempsey, O. (2016). Importance and use of correlational research. *Nurse Researcher, 23*(6), 20–25. https://doi.org/10.7748/nr.2016.e1382

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment.* Princeton, NJ: Educational Testing Service.

Diedenhofen, B., & Musch, J. (2015). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment*. http://dx.doi.org/10.1027/1015-5759/a000295

Docherty, A., & Dieckmann, N. (2015). Is there evidence of failing to fail in our schools of nursing? *Nursing Education Perspectives, 36*(4), 226-231. Retrieved from http://search.proquest.com/docview/1700288167?accountid=458

Doody, O., & Bailey, M. E. (2016). Setting a research question, aim and objective. *Nurse Researcher, 23*(4), 19. http://dx.doi.org/10.7748/nr.23.4.19.s5

Ferrara, S. (2014). Formative assessment and test security: The revised standards are mostly fine; Our practices are not. *Educational Measurement: Issues & Practice, 33*(4), 25-28. doi:10.1111/emip.12050

Fowler, M. D., & Davis, A. J. (2013). Ethical issues occurring within nursing education. *Nursing Ethics, 20*(2), 126-41. http://dx.doi.org/10.1177/0969733012474290

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine, 39*(1), 17-20. http://dx.doi.org/10.4103/0970-0218.126347

Goldmark J. (1923). Nursing and nursing education in the United States. Report of the Committee for the Study of Nursing Education. *Journal of the American Medical Association, 80*(21), 1538. doi:10.1001/jama.1923.02640480042033

Graue, M. E. (1993). Integrating theory and practice through instructional assessment. *Educational Assessment, 1*, 293-309.

Günüsen, N., Serçekus, P., & Edeer, A. (2014). A comparison of problem-based and traditional education on nursing students' locus of control and problem-solving skills. *International Journal of Nursing Knowledge, 25*(2), 110-115.

Hayat, M. (2013). Understanding sample size determination in nursing research. *Western Journal of Nursing Research, 35*(7), 943 – 956. doi:10.1177/0193945913482052

Hicks, N. A. (2011). Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. *Nurse Educator, 36*, 266-270.

Hoe, J., & Hoare, Z. (2013). Understanding quantitative research: Part 1. *Nursing Standard, 27*, 52-57. Retrieved from http://search.proquest.com/docview /1242111533?accountid=458

Hughes, L., Mitchell, M., & Johnson, A. (2016). Failure to fail' in nursing - A catch phrase or a real issue? A systematic integrative literature review. *Nurse Education in Practice, 20*, 54-63. doi: 10.1016/j.nepr.2016.06.009

Hunt, D. D. (2018). *The new nurse educator, second edition: Mastering academe*. New York, NY: Springer Publishing Company.

Hunter, J. L., & Krantz, S. (2010). Constructivism in cultural competence education. *Journal of Nursing Education, 49*(4), 207-14. Retrieved from https://search-proquest-com.contentproxy.phoenix.edu/docview/193898259?accountid=134061

International Test Commission. (2014). ITC Guidelines on quality Control in scoring, test analysis, and reporting of test scores. *International Journal of Testing, 14*(3), 195-217. doi:10.1080/15305058.2014.918040

Jacobsen, K. (2017). *Introduction to health research methods: A practical guide* (2nd ed.). Burlington, MA: Jones & Bartlett Publishers.

Kaur, M., Singla, S., & Mahajan, R. (2016). Item analysis of in use multiple choice questions in pharmacology. *International Journal of Applied & Basic Medical Research, 6*(3), 170-173. doi:10.4103/2229-516X.186965

Keating, S. B. (2014). C*urriculum development and evaluation in nursing*, (3rd ed.). New York: Springer Publishing Company.

Khoshaim, H., & Rashid, S. (2016). Assessment of the assessment tool: Analysis of items in a non-MCQ mathematics exam. *International Journal of Instruction, 9*(1), 119-131. doi:10.12973/iji.2016.9110a

Kickman, A., Neubert, S., & Reich, K. (Eds.). (2009). *John Dewey between pragmatism and constructivism.* New York, NY: Fordham University Press.

Killingsworth, E. (2013). *Nursing faculty decision making about best practices in test construction, item analysis, and revision* (Order No. 3577368). Available from ProQuest Dissertations & Theses Global. (1467588588). Retrieved from https://search.proquest.com/docview/1467588588?accountid=458

Killingsworth, E., Kimble, L. P., & Sudia, T. (2015). What goes into a decision? How nursing faculty decide which best practices to use for classroom testing. *Nursing Education Perspectives, 36*(4), 220-225. doi:10.5480/14-1492

King-Jones, M., & Mitchell, A. (2012). Grade inflation: A problem in nursing? *Creative Nursing, 18*(2), 74- 77. doi:10.1891/1078-4535.18.2.74

Kumandas, H., & Kutlu, O. (2015). High stakes tests. *Journal of Educational Sciences Research, 5*(2), 63-75. http://dx.doi.org/10.12973/jesr. 2015.52.4

Laerd Statistics (2015). Simple linear regression using SPSS Statistics. *Statistical tutorials and software guides.* Retrieved from https://statistics.laerd.com/

Locke, L., Spirduso, W., & Silverman, S. (2014). *Proposals that work* (6th ed.). Los Angeles, CA: Sage.

Madara, B., Resha, C., Krol, M. D., Lacey, K., Martin, E. F., O'Sullivan, C., & Smith, J. W. (2017). Nursing students' access to test banks: Are your tests secure*? Journal of Nursing Education, 56*(5), 292-294. http://dx.doi.org.contentproxy.phoenix.edu/10.3928/01484834-20170421-07

Maher, M. H. K., Barzegar, M., & Ghasempour, M. (2016). The relationship between negative stem and taxonomy of multiple-choice questions in residency pre-board and board exams. *Research and Development in Medical Education, 5*(1), 32-35. doi: 10.15171/rdme.2016.007

Malone, H.E., Nicholl H., & Coyne, I. (2016). Fundamentals of estimating sample size. *Nurse Researcher, 5*, 21-25. doi: 10.7748/nr.23.5.21.s5.

McDonald, M. (2014). *The nurse educator's guide to assessing learning outcomes* (3rd ed.). Burlington, MA: Jones and Bartlett Learning.

Moustakas, C. (2011). *Phenomenological research methods.* Thousand Oaks, CA: Sage Publications.

Namdeo, S. K., & Sahoo, B. (2016). Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. *International Journal of Research in Medical Sciences, 4*(5), 1716-1719. doi:10.18203/2320-6012.ijrms20161256

National Advisory Council on Nurse Education and Practice. (2010). Addressing new challenges facing nursing education: Solutions for a transforming healthcare environment, Eighth annual report. Retrieved from https://www.hrsa.gov/advisorycommittees/bhpradvisory/nacnep/.../eighthreport.pdf

National Council of State Boards of Nursing. (2016). *2016 NCLEX-RN® detailed test plan*. Retrieved from https://www.ncsbn.org/2016_RN_Test_Plan_Candidate.pdf

National League for Nursing. (2012). *NLN Fair Testing Guidelines*. Retrieved from http://www.nln.org/docs/default-source/default-document-library/fairtestingguidelines.pdf?sfvrsn=2

National League for Nursing. (2016). *NLN research priorities in nursing education 2016 – 2019.* Retrieved from http://www.nln.org/docs/default-source/professional-development-programs/nln-research-priorities-in-nursing-education-single-pages.pdf?sfvrsn=2

Nickerson, R., Butler, S., & Carlin, M. (2015). Knowledge assessment: Squeezing information from multiple-choice testing. *Journal of Experimental Psychology: Applied, 21*(2), 167–177. doi:10.1037/xap0000041

Oermann, M. H., & Gaberson, K. B. (2014). *Evaluation and testing in nursing education* (4th ed.). New York, NY: Springer.

Oermann, M. H., Saewert, K. J., Charaika, M., & Yarbrough, S. S. (2009). Assessment and grading practices in schools of nursing: National survey findings part 1. *Nursing Education Perspectives, 30*(5), 274-278.

O'Flynn-Magee, K., & Clauson, M. (2013). Uncovering nurse educators' beliefs and values about grading academic papers: Guidelines for best practices. *Journal of Nursing Education, 52*(9), 492-499. doi:10.3928/01484834-20130819-01

O'Halloran, K., & Gordon, M. (2014). A synergistic approach to turning the tide of grade inflation. *Higher Education, 68*(6), 1005-1023. doi:10.1007/s10734-014-9758-5

Paskausky, A. L., & Simonelli, M. C. (2014). Measuring grade inflation: A clinical grade discrepancy score. *Nurse Education in Practice, 14*(4), 374-9. http://dx.doi.org/10.1016/j.nepr.2014.01.011

Pazargadi, M., Ashktorab, T., & Khosravi, S. (2012). Nursing students' experiences on the evaluating role of their clinical educators: A qualitative study. *Asian Journal of Nursing Education and Research, 2*(3), 149-153.

Phelps, S., McDonough, S., Parker, R., & Finks, S. (2013). Throwing a Curveball: The Impact of Instructor Decisions Regarding Poorly Performing Test Questions. *American Journal of Pharmaceutical Education, 77*(9), 204. doi: 10.5688/ajpe779204

Privitera, G. (2017). *Research methods for the behavioral sciences* (2nd ed.). Thousand Oaks, CA: Sage Publishing, Inc.

Quinn, G., & Novotny, J. (2012). *Nuts and bolts approach to teaching nursing* (4th ed.). New York, NY: Springer Publishing Company.

Reynolds, D. (2015). Variability of passing grades in undergraduate nursing education programs in New York State. *Nursing Education Perspectives, 36*(4), 232-236.

Ritholz, M., Beverly, E., & Weinger, K. (2011). Digging deeper: the role of qualitative research in behavioral diabetes. *Current Diabetes Reports, 11*(6), 494-502. doi:10.1007/s11892-011-0226-7

Roa, M., Shipman, D., Hooten, J., & Carter, M. (2011). The costs of NCLEX-RN failure. *Nurse Education Today, 31*, 373-377. doi: 10.1016/j.nedt.2010.07.009

Roberts, C. (2010). *The dissertation journey* (2nd ed.). Thousand Oaks, CA: Sage

Romero, C., Zafra, A., Luna, J. M., & Ventura, S. (2013). Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Systems, 30*(2), 162-172. doi:10.1111/j.1468-0394.2012.00627.x

Roux, G. & Halstead, J. (2009). *Issues and trends in nursing: Essential knowledge for today and tomorrow*. Sudbury, MA: Jones and Bartlett Publishers.

Sagendorf, K. (2013). Writing and evaluating effective multiple choice tests. *Center for Excellence in Teaching and Learning*. Retrieved from http://www.oakland.edu/Assets/Oakland/cetl/files-and-documents/TeachingTips/TTBookVKOct15.pdf

St. John, F. et al. (2016). Research ethics: Assuring anonymity at the individual level may not be sufficient to protect research participants from harm. *Biological Conservation, 196*, 208-209. https://doi.org/10.1016/j.biocon.2016.01.025.

Salminen. L., Metsämäki, R., Numminen, O., & Leino-Kilpi , H. (2013). Nurse educators and professional ethics—Ethical principles and their implementation from nurse educators' perspectives. *Nurse Education Today, 33*, 133–137. doi:10.1016/j.nedt.2011.11.013

Salminen, L., Rinne, J., Stolt, M., & Leino-Kilpi, H. (2017). Fairness and respect in nurse educators' work- nursing students' perceptions. *Nurse Education in Practice, 23*, p.61-66. http://dx.doi.org/10.1016/j.nepr.2017.02.008

Salminen, L., Stolt, M., Metsämäki, R., Rinne, J., Kasen, A., & Leino-Kilpi, H. (2016). Ethical principles in the work of nurse educator: A cross-sectional study. *Nurse Education Today, 36*, 18-22. http://dx.doi.org/10.1016/j.nedt.2015.07.001

Schoening, A. (2013). From bedside to classroom: The Nurse Educator Transition Model. *Nursing Education Perspectives*, *34*(3), 167- 172.

Şentürk, C. & Zeybek, G. (2019). Teaching-learning conceptions and pedagogical competence perceptions of teachers: A correlational research. *Istraživanja u Pedagogiji, 1*, 65. https://doi.org/10.17810/2015.92

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher,29*(7), 4-14. doi:10.3102/0013189X029007004

Siddiqui, N. I., Bhavsar, V. H., Bhavsar, A. V., & Bose, S. (2016). Contemplation on marking scheme for Type X multiple choice questions, and an illustration of a practically applicable scheme. *Indian Journal of Pharmacology, 48*(2), 114-121. doi:10.4103/0253-7613.178836

Singh, A & Masuku, M. (2014). Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of Economics, Commerce, and Management, 11*(11), 1-22.

Smith, D., & Fleisher, S. (2011). The implications of grade inflation: Faculty integrity versus the pressure to succeed. *Journal of Research in Innovative Teaching, 4*(1), 32-38.

Sowbel, L. R. (2011). Field note gatekeeping in field performance: Is grade inflation a given? Journal of Social Work Education, 47(2), 367–377. doi 10.5175/JSWE.2011.201000006

Sullivan, D. (2014). A concept analysis of ‘‘High Stakes Testing.’’ *Nurse Educator, 39*(2), 72-76. doi:10.1097/NNE.0000000000000021

SurveyMonkey. (2017). *Data collection & privacy best practices*. Retrieved from https://help.surveymonkey.com/articles/en_US/kb/Collecting-secure-data-and-privacy-best-practices

SurveyMonkey. (2017). *Security statement*. Retrieved from https://www.surveymonkey.com/mp/policy/security/

Tagher, C., & Robinson, E. (2016). Critical aspects of stress in a high-stakes testing environment: A phenomenographical approach. *Journal of Nursing Education, 55*(3), 160-163. doi:10.3928/01484834-20160216-07

Talebi, G. A., Ghaffari, R., Eskandarzadeh, E., & Oskouei, A. E. (2013). Item analysis an effective tool for assessing exam quality, designing appropriate exam and determining weakness in teaching. *Research and Development in Medical Education, 2*(2), 69-72. Retrieved from http://search.proquest.com/docview/1509097368?accountid=458

Tarrant, M., & Ware, J. (2012). A framework for improving the quality of multiple-choice assessments. *Nurse Educator, 37*(3), 98-104. doi:10.1097/NNE.0b013e31825041d0

Theisen, J. L., & Sandau, K. E. (2013). Competency of new graduate nurses: A review of their weaknesses and strategies for success. *The Journal of Continuing Education in Nursing, 44*(9), 406-414. http://dx.doi.org/10.3928/00220124-20130617-38

Tobbell, D. A. (2014). "Coming to grips with the nursing question": The politics of nursing education reform in 1960s America. *Nursing History Review, 22*, 37-60. Retrieved from http://search.proquest.com/docview/1433806660?accountid=458

Vasiliki, G., Filippos, T. F., Christina, R., & Serafim, N. (2015). Software-assisted identification and improvement of suboptimal multiple choice questions for medical student examination. *Health Science Journal, 9*(2), 1-6. Retrieved from https://search.proquest.com/docview/1681853244?accountid=458

Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2017). *When to use what research design*. New York, N.Y.; The Guilford Press.

Watts, L., & Winters, R. (2017). Examining grade inflation and considerations for radiologic sciences: A literature review. *Journal of Medical Imaging and Radiation Sciences, 48*, 95-102. doi:10.1016/j.jmir.2016.08.002

Yildiz, M., Icli, G., & Gegez, A. (2013). Perceived academic code of ethics: A research on Turkish academics. *Procedia - Social and Behavioral Sciences, 99*, 282 – 29. doi:10.1016/j.sbspro.2013.10.496

Yin, R. (2015). *Qualitative research from start to finish* (2nd ed.). New York, NY: The Guilford Press.

Zaidi, N. L., Grob, K. L., Monrad, S. M., Kurtz, J. B., Tai, A., Ahmed, A. Z., Gruppen, L. D., & Santen, S. A. (2018). Pushing critical thinking skills with multiple-choice questions: Does Bloom's Taxonomy work. *Academic Medicine, 93*(6), 856-859. doi: 10.1097/ACM.0000000000002087.

Appendix A

Best Practices in Test Development

Test Construction

The following are components of test construction that faculty can use when constructing a test. Please indicate how often you use each component when developing test items for tests within the *identified nursing course*. Please answer on a scale of 1-5 (1= not at all to 5= all the time).

	1	2	3	4	5
Course objectives					
Class or unit objectives					
Major content topics					
Specific content topics					
A test blueprint or table of specifications					
The NCLEX-RN test plan					
Peer review of test items					
Higher cognitive levels according to Bloom's taxonomy (i.e., application, analysis, evaluation)					
Clinical context for test items					
Plausible distractors in multiple-choice test items					
Even distribution of correct answer in multiple-choice options					
Use various test item types (i.e., multiple-choice, choose all that apply, fill in the blank, etc.)					

Test Item Analysis

The following is a list of different information faculty can obtain about test items after a test is administered. Please indicate how often you use this information after test administration for tests within the *identified nursing course*. Please answer on a scale of 1-5 (1= not at all to 5= all the time).

	1	2	3	4	5
The number of students who answered each question incorrectly.					
The number of students who answered each question correctly (difficulty level or p-value).					
A question's ability to discriminate between the high and low scoring students (discrimination index or point biserial coefficient).					
The frequency of distractor choices with each test question.					
The discrimination between the high and low scoring students choosing distractors (distractor discrimination).					
The central tendency (mean, standard deviation) of the student grades on the test.					

Test Revision

The following are actions faculty can perform when revising classroom tests. Please indicate how often you perform these actions during test revision for tests within the *identified nursing course*. Please answer on a scale of 1-5 (1= not at all to 5= all the time).

	1	2	3	4	5
Use item analysis data when determining to keep or eliminate test questions before finalizing test scores.					
Compare item analysis data for test questions used repeatedly from one term to another.					
Use distractor discrimination to revise test items.					
Use difficulty level of test items to revise test items.					
Assess for linguistic/cultural bias in test items.					
Assess for changes in domain content based upon new research data.					
Assess for outdated language used in test items.					
Change test items to ensure test security.					
Change test items to reflect emphasis in classroom content.					
Change test items to ensure sufficient sampling of content.					

Instrument Scoring Directions- Best Practices in Test Development

Test Construction

Components of Test Construction

Possible scores: 12 to 60 (higher score indicates greater use of best practices in test construction)

Test Item Analysis

Components Used in Test Item Analysis

Possible Scores: 6 to 30 (higher score indicates greater use of best practices in test item analysis)

Test Revision

Components Used in Test Revision

Possible scores: 10 to 50 (higher scores indicates greater use of best practices in test revision)

Demographic and Teaching Background

In this section, the questions are intended to collect information about you, your teaching experience, and the nursing program you work in. Please select the option that best describes you and your nursing program.					
What type of institution do you work in?	Private faith based	Private liberal arts	Public doctoral institution	Public master's level institution	Public baccalaureate college

What type of nursing program do you teach in? (mark all that apply)	LPN/ LVN	LPN/ LVN to ASN	LPN/ LVN to BSN	Generic BSN	RN to BSN	Second degree BSN	Accelerated degree BSN

What U.S. state is the nursing program in?	
Are you a full time faculty member?	Yes / No
How many full time years have you been teaching nursing? (give numerical value)	

What is your highest degree completed?	ADN	BSN	MS in nursing	MS in other field	DNP	Ed D	PhD in nursing or DSN	PhD in other field

What role did your education prepare you for? (Mark all that apply)	RN	MSN educator	NP	CNM	CRNA	CNS	CNL

Do you hold an academic degree with an emphasis (major or minor) in education?	Yes / No
What is your age in years? (give numerical value)	
What is your gender?	Male / Female

What is your race/ethnicity? (mark all that apply)	Asian	Native Hawaiian	Other Pacific Islander	Black or African American	American Indian or Alaska Native	White	Hispanic Latino	Not Hispanic Latino	More than one race

Please indicate the amount of course work you have had in test development.	No course work in test development	Part of one course devoted to test development	One course in test development	More than one course in test development
Have you ever participated in a professional development program focusing on test development?	Yes	No	Can not remember	
Do you hold certification as a nurse educator (i.e., the CNE credential)?	Yes	No		

Appendix B

INFORMED CONSENT: PARTICIPANTS 18 YEARS OF AGE AND OLDER

Dear Participant,

My name is Diane Droutman and I am a student at the University of Phoenix working on a PhD degree. I am doing a research study entitled Grading Multiple Choice Questions and The Use of Evidence-Based and Ethical Practices by Nursing Faculty. The purpose of the research study is to develop insight into grading practices and the prevalence of the inconsistencies in grading practices of nursing faculty, specifically focusing on formative assessments using multiple choice questions.

Your participation will involve responding to an online survey, which will take approximately 20 minutes to complete. Any surveys not completely filled out will not be used in the study. You can decide to be a part of this study or not. Once you start, you can withdraw from the study at any time without any penalty or loss of benefits. The results of the research study may be published but your identity will remain confidential and your name will not be made known to any outside party.

In this research, there are no foreseeable risks to you.

Although there may be no direct benefit to you, a possible benefit from your being part of this study is learning information about grading practices and the prevalence of the inconsistencies in grading practices of nursing faculty.

If you have any questions about the research study, please call. For questions about your rights as a study participant, or any concerns or complaints, please contact the University of Phoenix Institutional Review Board via email at IRB@phoenix.edu.

As a participant in this study, you should understand the following:

1. You may decide not to be part of this study or you may want to withdraw from the study at any time. If you want to withdraw, you can do so without any problems. You can withdrawal at any time when completing online survey up until you hit the submit button.
2. Your identity will be kept confidential. Participant information, such as name, address and social security numbers will not be asked during survey. The information on age, and type of degree individual is teaching will be gathered. The information is not linked, nor can it be linked, to any individual participant.
3. Diane Droutman, the researcher, has fully explained the nature of the research study and has answered all of your questions and concerns.

4. Data will be kept secure. The information is kept in the survey money program and is only disclosed in data formation. The data will be kept for three years, and then destroyed by shredding any printed material.
5. The results of this study may be published.

"By signing this form, you agree that you understand the nature of the study, the possible risks to you as a participant, and how your identity will be kept confidential. When you sign this form, this means that you are 18 years old or older and that you give your permission to volunteer as a participant in the study that is described here."

(☐) I accept the above terms. (☐) I do not accept the above terms. (CHECK ONE)

Signature of the research participant _____ Date _____

Signature of the researcher _____ Date _____

Appendices C

Demographics

	Frequency	Percent
Gender		
Female	383	93.6
Male	26	6.4
Age		
26 - 35	22	5.4
36 - 45	52	12.7
46 - 55	131	32.0
56 - 65	170	41.6
Over 65	34	8.3
Race/ethnicity		
Pacific Islander	1	0.2
American Indian/Alaska Native	6	1.5
Asian	7	1.7
Black	18	4.4
Hispanic/ Latino	6	1.5
White	371	90.7
Full and part time instructors		
Full time	389	95.1
Part time	20	4.9
Education with emphasis in education		
Yes	253	61.9
No	156	38.1
Highest degree obtained		
BSN	2	0.5
MSN	101	24.7
MSN educator	130	31.8
NP	16	3.9
CNL	1	0.2
DNP	55	13.4
Ph.D.	104	25.4
Certified nurse educator		
Yes	102	24.9
No	307	75.1
Taken course in test development		
No course	68	16.6
Part of one course	51	12.5
One course	82	20.0
More than one course	208	50.9
Professional development program focusing on test development		
Yes	322	78.7
No	79	19.3
Do not know	8	2.0

Type of RN program		
Generic RN	353	86.3
LPN to RN	55	13.5
RN to BSN	63	15.4
Second degree RN	48	11.7
Accelerated degree RN	48	11.7
Types of institutions		
Private faith based	75	18.3
Private liberal arts	34	8.3
Public doctoral	30	7.3
Public master's level	17	4.2
Pubic baccalaureate college	80	19.6
Public community college	173	42.3

Appendix D

State Where Participants Live

	Frequency	Percent
AK	13	3.2
AL	1	0.2
AR	2	0.5
AZ	7	1.7
CA	32	7.8
CO	1	0.2
CT	7	1.7
Fl	1	0.2
FL	14	3.4
GA	2	0.5
IA	1	0.2
ID	1	0.2
IL	7	1.7
IN	7	1.7
KS	4	1.0
KY	3	0.7
LA	24	5.9
MA	7	1.7
MD	10	2.4
ME	3	0.7
MI	11	2.7
MN	3	0.7
MO	8	2.0
MS	14	3.4
MT	2	0.5
NC	14	3.4
ND	8	2.0
NH	3	0.7
NJ	22	5.4
NM	7	1.7
NV	6	1.5
NY	21	5.1
OH	17	4.2
OK	11	2.7
OR	1	0.2
PA	18	4.4
RI	9	2.2
SC	4	1.0
SD	3	0.7
TN	2	0.5
TX	41	10.0

VA	11	2.7
WA	11	2.7
WI	11	2.7
WV	3	0.7
WY	1	0.2