

12-1-2014

Establishing the validity and reliability of the fairness of items tool

Nikole Anderson Hicks

Follow this and additional works at: <http://digscholarship.unco.edu/dissertations>

Recommended Citation

Hicks, Nikole Anderson, "Establishing the validity and reliability of the fairness of items tool" (2014). *Dissertations*. Paper 154.

This Text is brought to you for free and open access by the Student Research at Scholarship & Creative Works @ Digital UNC. It has been accepted for inclusion in Dissertations by an authorized administrator of Scholarship & Creative Works @ Digital UNC. For more information, please contact Jane.Monson@unco.edu.

© 2014

NIKOLE ANDERSON HICKS

ALL RIGHTS RESERVED

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

ESTABLISHING THE VALIDITY AND RELIABILITY
OF THE FAIRNESS OF ITEMS TOOL

A Dissertation Completed in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Nikole Anderson Hicks

College of Health and Human Services
School of Nursing
Nursing Education

December 2014

This Dissertation by: Nikole Anderson Hicks

Entitled: *Establishing the Validity and Reliability of the Fairness of Items Tool (FIT)*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Health and Human Services in School of Nursing, Program of Nursing Education

Accepted by the Doctoral Committee

Janice Hayes, PhD, Chair

Vicki Wilson, PhD, Committee Member

Faye Hummel, PhD, Committee Member

Lisa Rue, PhD, Faculty Representative

Date of Dissertation Defense November 13, 2014

Accepted by the Graduate School

Linda L. Black, Ed.D.
Dean of the Graduate School and International Admissions

ABSTRACT

Hicks, Nikole Anderson. *Establishing the Validity and Reliability of the Fairness of Items Tool*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2014.

This dissertation manuscript describes a research study to validate the discipline-specific Fairness of Items Tool (FIT) for its use by nurse educators in identifying bias in multiple-choice questions (MCQs) to improve the quality of examinations. Multiple-choice (MC) examinations are a common assessment method used in programs of nursing, and conclusions based on these assessments have high stakes consequences. Faculty members therefore have an obligation to ensure that tests are valid and reliable assessments of student learning. For an examination to be a fair, valid, and reliable, it must contain well-written test items. Constructing and revising test items is difficult and time consuming, and nursing faculty members lack adequate preparation and sufficient time for examination construction and analysis. Published guidelines are available to assist faculty in creating examination items; however, assessments and textbook item banks contain violations of these guidelines, resulting in the administration of assessments containing flawed test items. Developing clear and concise guidelines for nursing faculty to use in developing unbiased test items is one strategy that may improve the quality of nursing assessments, thereby improving the quality of the decisions made based on these assessments.

Development and validation of the FIT was a three-phase process grounded in two theoretical frameworks adapted for this research study: the Revised Framework for Quality Assessment and the Conceptual Model for Test Development. In the first phase, the tool was developed by the primary investigator through an extensive review of published higher education and nursing literature related to item-writing rules, examination bias, and cultural bias. This dissertation study comprised phases two and three, using systematic methods to establish the validity and reliability of the FIT. In phase two, content validity and face validity were established through review by a panel of item-writing experts. In phase three, multiple measures were used to establish reliability and construct validity through testing of the FIT by nursing faculty ($N = 488$) to evaluate sample MCQs. The results of this research study support the hypothesis that the FIT is a valid and reliable tool for identifying bias in MC examination items as one component of a systematic process for test development.

Nurse educators can use the Fairness of Items Tool (FIT) as a guide for writing MCQs and revising textbook test bank items to improve the quality of examinations. The FIT also provides a means to facilitate systematic research to validate item-writing guidelines and testing procedures and to improve the quality of MC test items. Improving the quality of nursing examinations has the potential to improve student success and better prepare graduates for licensure and certification examinations, indirectly increasing the quality, quantity, and diversity of nurses joining the workforce.

Key words: multiple-choice test items, assessment, item-writing guidelines, item bias, test development

ACKNOWLEDGEMENTS

Thank you to Cengage Learning, Health Education Systems, Inc. / Elsevier, and Taylor & Frances for granting permission to use content for the theoretical frameworks in this research study.

Thank you to the members of the expert panel for lending your time and expertise. I appreciate you going before me and inspiring my own passion for this important work.

Thank you to Dr. Jun Ying for providing statistical consultation for this research study and to Christopher Mastin for your technical expertise with the surveys.

Thank you to the members of my dissertation committee for your guidance and support through my entire PhD journey.

I appreciate all of the friends and colleagues who have mentored and encouraged me and the students who inspire me.

To my family, thank you for your love and support and for tolerating all of the inconveniences that were part of this process. I am so blessed! Thank you to my parents for igniting my love of learning. To my children and grandchildren, I hope that you will also love learning and that you find and pursue your passion in life. Finally, I am thankful that God is the source of my strength and that through Him all things are possible.

TABLE OF CONTENTS

CHAPTER

I.	INTRODUCTION.....	1
	Background of the Study	
	Problem Statement	
	Purpose Statement	
	Research Question and Hypotheses	
	Definitions of Key Terms	
	Significance of the Study	
II.	REVIEW OF LITERATURE.....	18
	Keywords, Databases, and Resources	
	Theoretical Literature	
	Empirical Literature	
III.	METHODOLOGY.....	58
	Research Design	
	Phase One – Development of the Fairness of Items Tool	
	Phase Two – Validating the FIT through Expert Review	
	Phase Three – Validating the FIT with Nursing Faculty	
	Protection of Human Subjects	
IV.	ANALYSIS OF RESULTS.....	78
	Phase Two – Validating the FIT through Expert Review	
	Phase Three – Validating the FIT with Nursing Faculty	
	Additional Findings	
	Summary of the Findings	
V.	CONCLUSIONS AND RECOMMENDATIONS.....	108
	Discussion of Findings	
	Limitations	
	Importance for the Nursing Education	
	Recommendations for Further Research	

REFERENCES.....	125
APPENDIX A: Fairness of Items Tool.....	142
APPENDIX B: Permission to Use Content from <i>Educational Psychologist</i>	144
APPENDIX C: Permission to Use Content from HESI/Elsevier.....	146
APPENDIX D: Permission to Use Quinn’s (2000) <i>Cardinal Criteria</i>	148
APPENDIX E: Expert Panel Survey.....	150
APPENDIX F: Expert Panel Survey – Revised FIT.....	158
APPENDIX G: Results of Expert Panel Survey 1 & 2.....	165
APPENDIX H: Themes from Expert Panel Review 1	168
APPENDIX I: Expert Panel Decision Rubric	170
APPENDIX J: Revised Fairness of Items Tool (FITr).....	172
APPENDIX K: Participant Announcement Email	174
APPENDIX L: Participant Interest Form	176
APPENDIX M: FIT Survey – Demographics	178
APPENDIX N: Sample FIT Survey – Comprehensive.....	182
APPENDIX O: MC Test Item Selection for FIT Surveys.....	207
APPENDIX P: Participant Invitation Email	216
APPENDIX Q: Institutional Review Board Letters of Approval	218
APPENDIX R: Rationale for Revisions to Fairness of Items Tool	223
APPENDIX S: Demographic Characteristics.....	226
APPENDIX T: Descriptive Statistics	230
APPENDIX U: Comparisons of Means for Guideline and Dimension	251

APPENDIX V: Results of Tests of Independence of Scores	257
APPENDIX W: Agreement Indices	265

LIST OF TABLES

1. Descriptive Statistics for Data Analysis.....	71
2. Validity from Expert Panel Review 1 and 2.....	80
3. Completion Data for FIT Surveys.....	89
4. Known Groups Comparison: Differences of Means of Test Item Scores.....	99
5. Interpretation of Raw Agreement Indices.....	101
6. Comparison of Cronbach's Alpha Coefficients.....	104
S.1. General Demographic Characteristics of Sample Population.....	227
S.2. Demographic Characteristics of Sample – Education and Experience.....	228
S.3. Demographic Characteristics of Sample – Faculty Status and Expertise.....	229
T.1. Descriptive Statistics – Test Item B-1	231
T.2. Descriptive Statistics – Test Items B-2, B-3.....	232
T.3. Descriptive Statistics – Test Items B-5, B-10.....	233
T.4. Descriptive Statistics – Test Item B-11.....	234
T.5. Descriptive Statistics – Test Items B-12, B-14.....	235
T.6. Descriptive Statistics – Test Item B-13.....	236
T.7. Descriptive Statistics – Test Items B-15, B-16, B-20.....	237
T.8. Descriptive Statistics – Test Item B-18.....	238
T.9. Descriptive Statistics – Test Items B-21, B-22, B-23.....	239
T.10. Descriptive Statistics – Test Items B-25, B-27, B-28.....	240

T.11. Descriptive Statistics – Test Items B-29, B-30.....	241
T.12. Descriptive Statistics – Test Items B-31, B-32, B-33.....	242
T.13. Descriptive Statistics – Test Items B-34, B-36.....	243
T.14. Descriptive Statistics – Test Item B-35.....	244
T.15. Descriptive Statistics – Test Item F-10.....	245
T.16. Descriptive Statistics – Test Item F-8.....	246
T.17. Descriptive Statistics by Dimension for the Items on the Comprehensive Survey.....	247
T.18. Descriptive Statistics by Dimension for the Items on the Stem Survey.....	247
T.19. Descriptive Statistics by Dimension for the Items on the Options Survey.....	248
T.20. Descriptive Statistics by Dimension for the Items on the Linguistic-Structural Survey.....	249
T.21. Descriptive Statistics by Dimension for the Items on the Structural Survey.....	250
U.1. Known Groups Comparison: Evaluate the Stem (ES).....	252
U.2. Known Groups Comparison: Evaluate the Options (EO).....	253
U.3. Known Groups Comparison: Linguistic-Structural Bias (LS).....	254
U.4. Known Groups Comparison: Cultural Bias (C).....	255
U.5. Known Groups Comparison: Dimensions Scores.....	256
V.1. Independence of Cultural Bias and Demographic Variables.....	258
V.2. Independence of Linguistic-Structural Bias and Demographic Variables.....	259
V.3. Independence of Bias in the Options and Demographic Variables.....	260
V.4. Independence of Bias in the Stem and Demographic Variables.....	261
V.5. Independence of Comprehensive Bias and Demographic Variables – B1, B11.....	262
V.6. Independence of Comprehensive Bias and Demographic Variables – B13, B18....	263

V.7.Independence of Comprehensive Bias and Demographic Variables – B35, F10....	264
W.1. Agreement Indices – Guideline Level – Bias in the Stem (ES).....	266
W.2. Agreement Indices – Guideline Level – Bias in the Options (EO).....	267
W.3. Agreement Indices – Guideline Level – Linguistic-Structural Bias (L-S).....	268
W.4. Agreement Indices – Guideline Level – Cultural Bias (C).....	269

LIST OF FIGURES

1. Conceptual Framework for Test Development.....	32

CHAPTER I

INTRODUCTION

Multiple-choice examinations are a common assessment method used in programs of nursing. Multiple-choice questions (MCQs) are efficient, objective, easy to grade, can be used to test a broad sampling of the curriculum, and facilitate timely feedback and self-assessment (Brady, 2005). A single well-constructed test item may take an hour to write (Clifton & Schriener, 2010; Morrison & Free, 2001), however, and nursing faculty members often lack sufficient time for examination construction and analysis. Discipline-specific education in nursing means that few faculty members have formal preparation in assessment methods such as item construction (Tarrant, Knierem, Hayes, & Ware, 2006; Tarrant & Ware, 2008, Zungolo, 2008). Published guidelines are available to assist faculty in creating examination items that promote and measure critical thinking and to increase the validity and reliability of tests that measure student mastery of course concepts. Multiple reports demonstrate, however, that assessments and textbook item banks contain violations of these guidelines.

Flawed test items can affect student performance on MCQs, making the questions either easier or more difficult to answer (Downing, 2005) and resulting in distorted test results and lowered test reliability (Camilli & Shepard, 1994). Testing bias occurs when test results contain measurement error because of factors unrelated to the purpose of the exam. Sources of measurement error include the student, the environment, scoring

factors, and the test itself (Gaberson, 1996). When an examination is biased, students perform differently based on variables that are unrelated to their knowledge and abilities. A biased test item contains construct-irrelevant variances, such as item-writing flaws, which may be confusing to students and can affect performance on the item. For the purposes of evaluating quality, a reliable test will produce an accurate test score, as free from measurement error as possible (Chenevey, 1988; Demetrulias & McCubbin, 1982). A valid test measures what it is intended to measure – the test score is meaningful and contributes to accurate interpretation (Chenevey, 1988; McDonald, 2014). A test item is fair when it is free of bias, and students of equal ability are equally likely to answer it correctly (Klisch, 1994).

Nurse educators have an obligation to ensure that assessments are fair, valid, and reliable measures of learning for all students (Stuart, 2013). This dissertation manuscript describes a research study to test an intervention to improve the quality of multiple-choice (MC) examinations in programs of nursing. The purpose of this chapter is to discuss the background of assessment practices in nursing education. This discussion includes an overview of the purposes of assessment, nursing workforce issues, assessment practices and quality, and item-writing guidelines. An introduction to the purpose and significance of the research study will also be presented.

Background of the Study

Assessment plays an increasingly important role in evaluating student performance, satisfying quality issues, and addressing the needs of stakeholders (McCoubrie, 2004). The purpose of any assessment is to provide data from which conclusions are drawn, the most obvious of which is whether students have achieved the

desired learning outcomes (King, 1978; Stuart, 2013). If assessments are biased, faculty evaluations of student competency will be distorted (Brady, 2005).

Information from assessment tests determines students' grades and informs decisions about progression through a program of study. Assessments provide a basis for reporting student progress, motivating student learning, diagnosing learning difficulties of individuals and groups of students, and identifying areas of weakness within the course and curriculum (Case & Swanson, 2002; McDonald, 2014). An assessment test ultimately determines whether an applicant for licensure has the requisite knowledge to practice nursing.

Assessments also serve as valuable tools for learning (Bailey, Mossey, Moroso, Cloutier, & Love, 2012; Morrison & Free, 2001). Tests can communicate to students what material is important and provide students with information about areas in need of remediation and further study (Case & Swanson, 2002). The fact that assessment performance determines grades and progression is also a powerful motivator for students who have a desire for success. Results from assessments provide students with information about how their performance compares to other students. Assessment tests administered throughout the nursing program provide an opportunity for students to practice and prepare for success on the licensure examination.

Conclusions based on assessment have high stakes consequences for multiple stakeholders – students, course facilitators, schools of nursing, communities, accrediting bodies, and licensing agencies (Clifton & Schriener, 2010). For students, assessments determine success or failure in nursing. For faculty, assessments affect evaluations of teaching effectiveness and promotion decisions. For schools of nursing, assessments

provide a measure of program and graduate quality, upon which accreditation decisions are based. It is essential that these conclusions are based on unbiased measures that “fairly evaluate students’ achievement” (Demetrulias & McCubbin, 1982, p. 61).

Nursing Workforce Issues

Assessments influence decisions that ultimately affect the quality and quantity of graduates entering the nursing workforce. The recent Institute of Medicine (IOM) (2011) report calls for improvement in care delivery through developing a larger, more diverse, and highly educated nursing workforce. These goals will not be achieved without sound assessment practices.

Nursing shortage. The nursing shortage is expected to reach 500,000 registered nurses (RNs) by 2025 (Buerhaus, Staiger, & Auerbach, 2008). Workforce analysts with the Bureau of Labor Statistics (2014) project a need for an increase in the RN workforce of 19% through the year 2022. The supply of new nurses is not keeping up with the demand that is fueled by the needs of the aging population, despite increasing program enrollments and graduations (Joynt & Kimball, 2008). Significant numbers of qualified applicants are denied admission to basic RN programs each year, with almost two-thirds of qualified applicants rejected from admission to baccalaureate (BSN) programs in 2012 (National League for Nursing, 2013).

Given this demand for spots in nursing education programs, institutions have adopted highly selective admission policies (National League for Nursing, 2014a). The limited number of slots also increases the pressure to ensure that students progress through their education programs and achieve licensure successfully (Joynt & Kimball, 2008). Assessing the ability of students to meet identified learning outcomes and

preparing them to pass the licensure examination are established goals within programs of nursing. High quality assessment practices are an essential requirement for successfully meeting these goals.

Cultural diversity. Increasing the diversity of the registered nurse population is a high priority (American Association of Colleges of Nursing, 1997; American Association of Colleges of Nursing, 2014; Ayoola, 2013; Sitzman, 2007). A goal is that the number of culturally and ethnically diverse students entering and completing nursing programs reflects the diversity represented in the communities they serve (American Association of Colleges of Nursing, 2014; Evans & Greenberg, 2006). Minority students in nursing programs have high attrition rates, ranging from 15% to 85% (Taxis, 2002). According to the National League for Nursing (NLN) (2013), the percentage of racial-ethnic minorities and males graduating from prelicensure RN programs has not significantly increased in recent years. In order to improve graduation rates, nursing programs must increase recruitment and retention of a diverse student body (American Association of Colleges of Nursing, 2014; Andrews, 2003).

An approach that has received very little attention is the examination of how educators “assess [the] qualifications” of a diverse student body (American Association of Colleges of Nursing, 1997, ¶ 6). The lack of success for minority students has been attributed to multiple factors, including issues of bias in assessment practices (Bosher, 2009; Klisch, 1994). Nursing’s commitment to diversity requires that educators examine how they teach and evaluate students, including the use and construction of multiple-choice (MC) tests (Bosher, 2003). Improving the fairness of assessments for diverse groups of students will result in improved testing practices for all students.

Assessments in Nursing Education

Multiple-choice (MC) tests are the most frequently used assessment tool worldwide (Al-Faris, Alorainy, Abdel-Hameed, & Al-Rukban, 2010) and are the primary method used to evaluate competence in nursing programs and on the National Council Licensure Examinations (NCLEX) (Clifton & Schriner, 2010; Considine, Bottie, & Thomas, 2005; Giddens, 2009; Wendt & Harmes, 2009a). The MC format is a type of selected response item in which examinees are required to choose the correct answer to a question from a list of possible answers. MC tests are efficiently administered to large numbers of students, quickly and easily scored, and objectively graded (Brady, 2005; Clifton & Schriner, 2010; Tarrant & Ware, 2008). Well-constructed MCQs can measure a broad sampling of learning outcomes and higher order cognitive abilities (Downing, 2006; Haladyna, 2004; Hansen & Dexter, 1997). Quality test items discriminate between high and low performers and can test a student's ability to apply nursing concepts to "clinically-oriented situations" (Morrison & Free, 2001, p. 16). When properly written, MC tests can facilitate the preparation of nursing students for the licensure exam (Clifton & Schriner, 2010).

The National Council of State Boards of Nursing (NCSBN) requires examinees to answer a minimum of 75 questions on the NCLEX, most of which are MCQs at the application or above cognitive level (Clifton & Schriner, 2010; National Council of State Boards of Nursing, 2014a; Wendt, 2008). NCLEX pass rates are a common outcome measure of nursing program quality, and nurse educators have a responsibility to prepare students to pass. The NCSBN (2014b) reported an 83.33% first-time pass rate for RN candidates in 2014, compared with a 38.48% repeat pass rate, which translates to over

58,000 potential registered nurses who were unable to achieve licensure and contribute to the nursing workforce. Nursing students expend a great deal of time and resources through the duration of the nursing program, and those who successfully complete a nursing degree should be qualified to achieve licensure. The NCLEX measures minimal level of competence to practice nursing, and schools of nursing, at the very least, should prepare students at a minimal level of competence (Morrison, 2005). Valid and reliable assessment practices are an essential component of this process.

The NLN survey of assessment and grading practices in schools of nursing identified NCLEX pass rates as the most significant factor influencing faculty decisions about what strategies to use to assess learning outcomes (Oermann, Saewert, Charasika, & Yarbrough, 2009). Nurse educators commonly believe that MC tests prepare students for the licensure exam, and teacher-made tests are used widely for this purpose (Walloch, 2006). While other types of assessments – papers, group projects, and case studies – are used more frequently for assessment, tests weigh more heavily in student course grades (Oermann et al., 2009), often accounting for 75% to 100% of the grade in theory courses (DePew, 2001). With such high-stakes decisions based on MC tests, it is imperative that these tests contain quality MCQs that are valid and reliable means of assessing student learning.

Quality of Nursing Assessments

Good MCQs are difficult to write, and poorly constructed items are often misinterpreted and fail to assess what is intended (Case & Donahue, 2008; Downing, 2006; Farley, 1989). The development of valid and reliable classroom tests challenges even the most experienced educators (Demetrulias & McCubbin, 1982). The majority of

college faculty have minimal formal education or training in teaching and testing strategies (DiBattista & Kurzawa, 2011; Petress, 2007), and nurse educators are no different. The focus on preparation for advanced practice nurses in graduate education means that nurse educators often have little educational preparation for assessment and item writing (Masters et al., 2001; Morrison, Nibert, & Flick, 2006; Zungolo, 2008). As a result, many tests administered in nursing programs are poorly constructed (Clifton & Schrinier, 2010; Cross, 2000; Masters et al, 2001; Tarrant et al., 2006).

The most common source of MCQs is course faculty, followed by textbook test banks, with many educators using a combination of both (DePew, 2001). Textbook test banks provide a readily available source of questions upon which faculty rely because of time pressures and lack of item-writing skills (Clifton & Schrinier, 2010; Lampe & Tsaouse, 2010). Studies conducted in multiple disciplines, however, demonstrate that textbook test banks are full of flawed test items that result in biased examinations (Ellsworth, Dunnell, & Duell, 1990; Hansen & Dexter, 1997; Masters et al., 2001).

The security of these test banks presents an additional source of bias in student assessment. Cross (2000) examined 130 nursing examinations from 66 programs in 31 states and found the same items appearing on exams from widely varied locations, suggesting a common source of test items, but also calling into question the security of these items. An Internet search conducted by this author yielded multiple sites that sell test banks from current and previous editions of common nursing textbooks. Burns (2009) searched auction sites for nursing test banks and found 12 pages with over 70 listings for sale. Obviously, students can easily purchase access to these test banks, and prior knowledge of test questions contributes to measurement error. When faculty use test

bank items, editing is crucial if the items are to contribute to a valid and reliable test (Burns, 2009; Ellsworth et al., 1990; Masters et al., 2001).

Classroom assessment consumes large amounts of instructor time, effort, and resources (Downing, 2005). Preparing a reliable and valid test is challenging and time consuming. MCQs are difficult to write, and a good test item takes an hour or more to develop (Clifton & Schriener, 2010; Farley, 1989; Morrison & Free, 2001; Piasentin, 2010). Item writing is only one component of the test development process. To ensure high-quality assessments, adequate time needs to be devoted to item writing, peer review, and revision prior to administration of the test (Haladyna & Downing, 1985; Haladyna & Downing, 1989a; Morrison et al., 2006; Tarrant et al., 2006; Tarrant & Ware, 2012; Vacc, Loesch, & Lubik, 2001; Weaver, 1982). Following administration, test items should continue to be revised according to item performance analysis statistics (Morrison et al., 2006; Tarrant & Ware, 2012; Vacc et al., 2001). While faculty often spend considerable time preparing course materials and planning class sessions, insufficient time is allotted in faculty workloads for test preparation and review (Tarrant et al., 2006). Consequently, time pressures force faculty to develop examinations hastily, sometimes even the night before a test is administered. Pre- and post-administration review and revision is often ignored, and poor quality test items remain in the test (Clifton & Schriener, 2010; Tarrant et al., 2006).

Item-Writing Guidelines

Guidelines for preparing effective test items are well documented in textbooks (Billings & Halstead, 2009; Case & Swanson, 2002; Downing & Haladyna, 2006; Gronlund, 1998; Haladyna, 2004; McDonald, 2014; Morrison et al., 2006; Osterlind,

1998; Quinn, 2000; Quinn & Hughes, 2007) and journal articles (Aiken, 1987; Al-Faris et al., 2010; Boland, Lester, & Williams, 2010; Bosher & Bowles, 2008; Brady, 2005; Breitbach, 2010; Campbell, 2011; Case & Donahue, 2008; Chenevey, 1988; Considine et al., 2005; Ellsworth et al., 1990; Farley, 1989; Gaberson, 1996; Hansen & Dexter, 1997; Hicks, 2011; King, 1978; Morrison & Free, 2001; Stanton, 1983; Tarrant & Ware, 2012; Vacc et al., 2001; Weaver, 1982). While much of the published literature is based upon personal experience, common sense, and values, many of these guidelines are developed through reviews of published item-writing literature and substantiated through empirical research.

Frey, Peterson, Edwards, Pedrotti, and Peyton's (2005) review of literature yielded a list of 40 item-writing rules. Tarrant et al. (2006) developed nursing discipline-specific guidelines containing 32 item-writing rules, which was reduced to a 19-guideline tool for use in their research (Tarrant et al., 2006; Tarrant & Ware, 2008). Other researchers in nursing education have developed guidelines containing from 20 to 52 item-writing rules (Bosher, 2003; Klisch, 1994; Masters et al., 2001; Van Ort & Hazzard, 1985).

Researchers have also attempted to organize guidelines into taxonomies for use in evaluating multiple-choice test items, some of which have been validated through systematic research. The most comprehensive work on taxonomy development has been conducted by Haladyna and Downing (1985; 1989a; 1989b). The original taxonomy of 43 MC item-writing rules was validated and reduced 31 guidelines (Haladyna, Downing, & Rodriguez, 2002). Moreno, Martínez, & Muñiz (2006) further revised this taxonomy to a set of 15 guidelines.

There is much agreement and overlap between these guidelines and taxonomies, but the development of multiple item-writing taxonomies contributes to confusion about their validity and utility. In fact, there are so many item-writing guidelines in textbooks and journal articles that faculty simply cannot make sense of all of them. Inevitably, misunderstandings and poor practices develop and flourish (Burton, 2005), and violations of item-writing guidelines in nursing examinations are common and persistent (Bosher, 2003; Clifton & Schriener, 2010; Cross, 2000; Masters et al., 2001; Tarrant et al., 2006; Tarrant & Ware, 2008).

Problem Statement

Faculty have an ethical obligation to ensure that tests are valid and reliable assessments of student learning so that students do not fail tests and/or courses because of poorly written test items. Guidelines to assist faculty in developing quality MC test items are abundant, but faculty have limited time to sift through the information to determine which guidelines are most relevant for assessment in nursing education. Graduates of nursing programs are expected to be capable of applying nursing concepts in complex clinical situations (Morrison & Free, 2001). MC tests that are designed to assess student outcomes related to providing safe and competent nursing care must therefore be written according to discipline-specific guidelines (Morrison & Free, 2001). There is a need for a tool to provide discipline-specific guidelines for nursing faculty to improve the quality of test items. An effective tool will provide a clear and concise description of the most relevant guidelines in an easy-to-use format that facilitates writing and revising fair, valid, and reliable MC test items within a nurse educator's full workload. Use of such a tool in nursing education has the potential to improve MC assessments, better prepare

students for success on the licensure examination, and enhance the quantity and diversity of the nursing workforce.

Purpose Statement

The purpose of this dissertation study was to test an intervention to improve the quality of nursing examinations – specifically, to evaluate the Fairness of Items Tool (FIT) (Appendix A) for its use in the identification of bias in MC questions. This study examined the question: *Is the Fairness of Items Tool (FIT) a valid and reliable tool for identification of bias in multiple-choice examination items by nurse educators?* The study had as its aim to establish the validity and reliability of the FIT through expert review and comparison of faculty scores on MC test items.

Research Question and Hypotheses

- Q: Is the Fairness of Items Tool (FIT) a valid and reliable tool for identification of bias in multiple-choice examination items by nurse educators?
- H1: The Fairness of Items Tool (FIT) is a valid tool for identification of bias in multiple-choice examination by nurse educators.
- H2: The Fairness of Items Tool (FIT) is a reliable tool for identification of bias in multiple-choice examination by nurse educators.

Definitions of Key Terms

Multiple-Choice Question (MCQ)

An examination item intended to obtain objective information about examinees' cognitive behaviors and requiring the selection of one correct response from a set of response choices (Haladyna, 2004). The conventional format for MCQs includes a stem, one correct answer, and distracters.

Stem. Provides a stimulus for the response and presents the problem or question to be answered.

Options. All of the possible answer choices.

Correct answer. There is only one correct answer for each MCQ.

Distracters. Incorrect answers that may be plausible to those who have not mastered the knowledge that the item is designed to measure yet are clearly incorrect to those who possess the knowledge required (Haladyna, 2004). A good distracter is one that is selected by those who perform poorly and ignored by those who perform well (McDonald, 2014).

Construct-Irrelevant Variance (CIV)

The introduction of extraneous variables into an assessment that are irrelevant to the construct being measured and which can increase or decrease test scores for some or all examinees (Downing, 2002a). The presence of CIV prevents proper interpretation of test scores, reducing the validity of the assessment, and corrupting the decisions made on the basis of test scores (Abedi, 2006; Downing, 2002a; Tarrant & Ware, 2008).

Differential Item Functioning (DIF)

The presence of “some characteristic of an item that results in differential performance” (Hambleton & Rodgers, 2005, ¶3) for “different subgroups of examinees” (Downing, 2002b, p. 238).

Flawed Item

A poorly crafted test item that contains one or more item-writing flaws.

Item Bias

Any test item that is potentially unfair to examinees. A test item is biased if it might disadvantage some students more than others based on variables that are irrelevant to the construct being tested. A test item is free of bias if students of equal ability are

equally likely to answer it correctly. Item bias results from flawed test items and may be linguistic, structural, or cultural in nature or related to irrelevant difficulty, testwise cues, or formatting errors.

Testwise cues. Irrelevant and unintended clues to the correct answer that enable testwise students to select the correct response without having the required ability (McDonald, 2014).

Irrelevant difficulty. Flaws of irrelevant difficulty in test items make questions difficult to understand for “reasons unrelated to the content or focus of assessment” (Bosher, 2003, p. 27).

Linguistic bias. “Unnecessary complexity” in the wording of the stem or options producing test items that are not easily understood (Bosher, 2003, p. 26).

Structural bias. Long, unclear, or incorrect grammatical components that are confusing to all students but present more difficulty for nonnative speakers of English (Klisch, 1994).

Cultural bias. The use of culturally specific information in a test item that is not “equally available to all cultural groups” (Bosher, 2003, p. 26).

Item-Writing Flaws

Errors in construction that violate one or more item-writing guidelines (Downing, 2002b) and introduce construct-irrelevant variance into the test item.

Multilogical Thinking

“Thinking that requires knowledge of more than one fact to logically and systematically apply concepts to a clinical problem” (Morrison et al., 2006, p. 151).

Taxonomy

A classification of item-writing guidelines grouped into major content categories and intended to “completely represent the array of advice on preparing MC items” (Haladyna & Downing, 1989, p. 38).

Test Bias

Any construct-irrelevant source of variance that results in systematically higher or lower test scores for groups of examinees (*Standards for Educational and Psychological Testing*, 1999). Test bias in a nursing exam refers to the difference in a group’s mean performance based on “non-nursing elements in the exam” (McDonald, 2014, p. 22).

Test Fairness

A judgment of a test’s authenticity, reflecting faith in the test’s reliability and validity, as well as quality of construction and appropriate standard setting (McCoubrie, 2004). A fair test is defensible. Anything that lowers the validity and reliability of a test for a group of test takers reduces the fairness of the test (Zieky, 2006).

Test Reliability

The reproducibility of a set of test scores obtained from a particular group, on a particular day, under particular circumstances (McDonald, 2014). Reliability is a property of the test scores, not a characteristic of the test itself (Downing, 2004; McDonald, 2014). The primary requirement for test reliability is well-constructed test items (McDonald, 2014).

Test Validity

The “reasonableness and meaningfulness of the inferences drawn from assessments” (Downing, 2002b, p. 240). Validity is influenced by the extent or degree to

which a test measures what it is supposed to measure, which provides evidence to justify the inferences made on the basis of the test results (McDonald, 2014; Stuart, 2013).

Validity is a property of the test scores, not a characteristic of the test itself (Downing, 2004; McDonald, 2014). Any factors that interfere with the interpretation of assessment scores threaten validity (Downing & Haladyna, 2004).

Testwiseness

An “examinee characteristic” (Haladyna & Downing, 1985, p. 20) in which students understand how to select the correct options based on the structure or wording of the questions.

Significance of the Study

There is a need for development of a valid and reliable tool for use by nurse educators in evaluating and revising MC test items. The existing taxonomies lack evidence of their validity and reliability in identifying item bias. Previous research has been conducted with descriptive methods and inconsistent use of taxonomies. This research study contributes to the body of knowledge by establishing the validity and reliability of a tool that will be used by nurse educators to develop high quality assessments of student learning. The Fairness of Items Tool (FIT) provides a means to facilitate systematic research to validate item-writing guidelines, testing procedures, and the actual quality of test items. Improving the quality of MC tests will better prepare nursing students for success on the licensure examination, increasing the quantity of nurses eligible to join the workforce. Assurance of fair testing in nursing programs also has the potential to enhance the diversity of the nursing workforce by removing a barrier to student success.

Summary

Multiple-choice examinations are a common assessment method used in programs of nursing, and conclusions based on these assessments have high stakes consequences. Faculty members therefore have an obligation to ensure that tests are valid and reliable assessments of student learning. For an examination to be a fair, valid, and reliable, it must contain well-written test items. Writing well-constructed test items is difficult and time consuming, and nursing faculty members lack adequate preparation and sufficient time for examination construction and analysis. Published guidelines are available to assist faculty in creating examination items; however, assessments and textbook item banks contain violations of these guidelines, resulting in flawed assessments containing flawed test items. Developing a clear and concise guideline for nursing faculty to use in developing unbiased test items is one strategy that may improve the quality of nursing assessments, thereby improving the quality of the decisions made based on these assessments. This chapter presented a discussion of the background of assessment practices in nursing education, providing an overview of the purposes of assessment, nursing workforce issues, nursing assessment practices and quality, and item-writing guidelines. The following chapter provides a thorough discussion of the theoretical and empirical literature related to this dissertation research study.

CHAPTER II

REVIEW OF LITERATURE

Multiple-choice (MC) tests are the most frequently used assessment tool worldwide (Al-Faris, Alorainy, Abdel-Hameed, & Al-Rukban, 2010) and are the primary method used to evaluate competence in nursing programs and on the National Council Licensure Examinations (NCLEX) (Clifton & Schriener, 2010; Considine, Bottie, & Thomas, 2005; Giddens, 2009; National Council of State Boards of Nursing, 2014a). Conclusions based on assessment have high stakes consequences, and it is essential that these conclusions are based on unbiased measures that “fairly evaluate students’ achievement” (Demetrulias & McCubbin, 1982, p. 61). When an examination is biased, students perform differently based on variables that are unrelated to their knowledge and abilities. A biased test item contains construct-irrelevant variances, such as item-writing flaws, that may be confusing to students and can affect performance on the item. A test item is fair when it is free of bias, and students of equal ability are equally likely to answer it correctly (Klisch, 1994).

Good multiple-choice questions (MCQs) are difficult to write, with a single well-constructed test item taking take an hour or more to develop (Morrison & Free, 2001; Piasentin, 2010), and poorly constructed items are often misinterpreted and fail to assess what is intended (Case & Donahue, 2008; Downing, 2006; Farley, 1989). The

development of valid and reliable classroom tests challenges even the most experienced educators (Demetrulias & McCubbin, 1982). The majority of college faculty have minimal formal education or training in teaching and testing strategies (DiBattista & Kurzawa, 2011; Petress, 2007), and nurse educators are no different. The focus on preparation for advanced practice nurses in graduate education means that nurse educators often have little educational preparation for assessment and item writing (Masters et al., 2001; Morrison, Nibert, & Flick, 2006; Zungolo, 2008). Published guidelines are available to assist faculty in creating examination items that promote and measure critical thinking and to increase the validity and reliability of tests that measure student mastery of course concepts. Multiple reports demonstrate that assessments and textbook item banks contain violations of these guidelines, and, as a result, many tests administered in nursing programs are poorly constructed (Clifton & Schriener, 2010; Cross, 2000; Masters et al, 2001; Tarrant, Knierim, Hayes, & Ware, 2006).

Nurse educators have an obligation to ensure that assessments are fair, valid, and reliable measures of learning for all students. Graduates of nursing programs are expected to be capable of applying nursing concepts in complex clinical situations (Morrison & Free, 2001). MC tests designed to assess student outcomes related to providing safe and competent nursing care must therefore be written according to discipline-specific guidelines (Morrison & Free, 2001). This dissertation manuscript describes a research study to validate the discipline-specific Fairness of Items Tool (FIT) for its use in identifying bias in MCQs to improve the quality of examinations in programs of nursing.

A significant body of literature on the development of valid and reliable MC assessments provides a basis for this proposed research study. This chapter explains the

search process in reviewing that literature and examines the theoretical literature that frames the proposed research study. The body of literature related to MC item writing is then discussed, followed by a thorough analysis of the empirical research related to the topic.

Keywords, Databases, and Resources

The literature search included the key words: multiple-choice, multiple-choice question, bias, nursing education, item writing, multiple-choice examination, examination, higher education, testing, linguistics, and test item. The primary electronic databases used in the search were: Cumulative Index to Nursing and Applied Health Literature (CINAHL), EBSCOHost, Education Full Text, Education Research Complete, Academic Search Premier, MasterFile Premier, PsychINFO, ProQuest Nursing and Allied Health Source, ProQuest Dissertations, and Educational Resources Information Center (ERIC). References frequently cited in relevant articles were reviewed, as were works by the key expert authors who developed the theoretical frameworks and taxonomies of item-writing guidelines. Other resources included the University of Cincinnati Health Sciences Library and the University of Northern Colorado Michener Library. The review was limited to English-language articles, although it was international in scope.

Theoretical Literature

There is much controversy in the nursing and education literature related to whether MC tests are an effective testing method for a practice discipline. The best practice is to use a variety of assessment methods. While it is beyond the scope of this research study to evaluate other assessment methods, the fact remains that MCQs are

widely used in educational assessment. MC tests are the most frequently used assessment tool worldwide (Al-Faris et al., 2010) and are the primary method used to evaluate competence in nursing programs and on the National Council Licensure Examinations (NCLEX) (Clifton & Schriener, 2010; Considine et al., 2005; Giddens, 2009; National Council of State Boards of Nursing, 2014a; Wendt, 2008). Regardless of the method used, it is important that assessment tests provide quality data from which inferences about a student's performance can be drawn.

Cardinal Criteria for Assessment

Quinn's (2000) Cardinal Criteria of Assessment provides a theoretical framework upon which to evaluate assessments. According to Quinn, every effective assessment must meet the following criteria:

Validity – the extent to which a test measures what it is designed to measure;

Reliability – the consistency with which a test measures what it is designed to measure;

Discrimination – the ability of a test to distinguish between the more knowledgeable and the less knowledgeable students;

Practicality/utility – whether the test is practical for its purposes (p. 204-205).

A discussion of these terms and analysis of MCQs in relationship to these criteria follows.

Practicality/usability. The first measure of a test's effectiveness is whether it is practical for its purpose in relation to time, cost, and ease of use (Quinn & Hughes, 2007). MCQs are efficient, affording the ability to assess large numbers of students at one time (Brady, 2005; Clifton & Schriener, 2010; Tarrant & Ware, 2008). MC tests allow

students to answer a large number of questions in a short time period, so a wide range of content areas can be included on a single examination (Breitbach, 2010; Brady, 2005; Clifton & Schriener, 2010; Tarrant & Ware, 2008). MCQs can be used to measure a range of learning outcomes and multiple levels of cognitive domains (Chenevey, 1988; Downing, 2006; Hansen & Dexter, 1997; King, 1978). MC tests also provide ease of scoring and can be easily pretested, stored, used, and reused with the use of computerized item-banking systems (Brietbach, 2010; Haladyna & Downing, 1989a; Weaver, 1982). The primary limitation of MCQs is that well-constructed questions are time-consuming and difficult to develop (Brady, 2005; Clifton & Schriener, 2010; Tarrant & Ware, 2008). Despite this limitation, MCQs are popular specifically because of their ease of use and practicality (Boland, Lester, & Williams, 2010).

Reliability. The second requirement is reliability, meaning the consistency with which a test measures what it is designed to measure (Quinn & Hughes, 2007). Reliability is a property of the test scores and reflects the reproducibility of the scores with repeated administration (Chenevey, 1988; Considine et al., 2005; Downing, 2004; McDonald, 2014). Reliability is measured using correlation coefficients, most commonly the Kuder-Richardson formula (KR-20) (Demetrulias & McCubbin, 1982). The higher the reliability coefficient (KR-20), the more likely the scores will be consistent if the test is administered again (Downing, 2004). Reliability is influenced by variations in the test items, test administration, and examinee that affect the test taker's performance (Demetrulias & McCubbin, 1982). These variations include susceptibility to guessing, errors in test administration or scoring, and individual characteristics, such as anxiety and fatigue (Demetrulias & McCubbin, 1982; Layton, 1986). When assessment data are used

to make high stakes decisions, the reliability of the data must be high in order to provide sufficient evidence upon which to base those decisions (Downing, 2004).

MCQs have greater reliability than other testing formats (Aiken, 1987). An advantage of MCQs is the objective scoring process that provides a high degree of reliability and enforces consistent standards for all examinees (Case & Donahue, 2008; Considine et al., 2005). Computerized test item analysis programs are readily available and widely used, enabling reliability coefficients to be easily calculated. Computerized item banking provides a means for revising MC test items based on analysis of data and storing the revised item and its data over repeated administrations. These practices improve the reliability of individual test items (Downing, 2004; Morrison, 2005).

Test reliability is improved by including “sufficiently large numbers” (Downing, 2004, p. 1010) of high quality test items. The number of options within each question is also significant. Reliability increases with the number of test options from two to five (Haladyna & Downing, 1985). These factors can significantly impact the efficiency of the test administration – too many questions with too many options will negatively affect a test’s practicality. Because MC tests allow students to answer a large number of questions in a short time period, sufficient numbers of questions can be included for reliability without negatively affecting practicality. MCQs can include an unlimited number of options; however, multiple studies have demonstrated that test efficiency and reliability are best balanced by using three or four options (Haladyna & Downing, 1985).

High quality test items are needed for a reliable assessment, and test items must be clearly understood and free from construction errors (McDonald, 2014). The primary limitation of MCQs is that well-constructed items are difficult to write, and construction

errors are common (Brady, 2005). Despite this limitation, MCQs lend themselves to objective scrutiny through a peer review and pretesting process. Improving the quality of individual test items through item-writing procedures, obtaining pretest reliability data, and using post-administration analysis data (KR-20) to guide revision are the best means for improving the reliability of assessment data (Downing, 2004). These processes are time consuming; however, they can be reasonably accomplished with MC test items.

Validity. An effective assessment measures what it is designed to measure (Quinn & Hughes, 2007). A valid test provides data from which meaningful inferences can be drawn (Chenevey, 1988). A test must be reliable to be valid (McDonald, 2014). Unless assessment scores are reliable and reproducible, interpreting the meaning of the scores is difficult (Downing, 2003). A reliable test is not always valid, however (McDonald, 2014). For example, a math test may have reproducible scores (reliability); however, if the scores are used to make inferences about a student's language mastery, rather than math ability, the test is not valid. A poorly constructed examination that is not reflective of course outcomes is not valid for the purposes of making performance and progression decisions or evaluating the curriculum. When high stakes decisions are based on assessment data, the need for strong validity evidence increases (Downing, 2003). Important components of validity evidence include content and construct validity.

Content validity concerns whether MCQs are measuring relevant and important information, and the complete test is representative of the learning outcomes (Chenevey, 1988). Content validity is best demonstrated through adequate planning of the assessment and development of a test blueprint (Chenevey, 1988). Construct validity is related to whether the questions measure the domain of knowledge being examined and is

established through analysis of item response statistics – point biserial, item difficulty, and evaluation of distracter effectiveness (Considine et al., 2005).

High quality test items are a critical source of both content and construct validity (Downing, 2003). Poorly written questions contain construct-irrelevant variances (CIV) that may be confusing to students and can affect performance on the item. The presence of CIV prevents proper interpretation of test scores and reduces the validity of the assessment (Abedi, 2006; Downing, 2002a; Tarrant & Ware, 2008). MCQs must use clear language at an appropriate reading level to ensure that the question is measuring understanding of a nursing concept, rather than language mastery (Paxton, 2000). MCQs must not contain flaws that cue testwise students to select the correct response without having the requisite knowledge (Downing, 2002b; Tarrant & Ware, 2008). Another source of CIV is cheating during the examination or through availability of test items prior to the examination, both of which have the potential to artificially elevate test scores and reduce test validity (Downing, 2002b; Downing & Haladyna, 2004). The difficulty in constructing high quality MC items is, again, a primary limitation of this format. The MC format is well suited for assessing cognitive knowledge; however, this format is not effective for the assessment of psychomotor and affective knowledge (Downing, 2006). Developing an effective assessment program must therefore incorporate multiple testing formats, one of which includes high quality MCQs.

Despite their limitations, MC formats have many advantages over other assessment formats and provide strong validity evidence (Downing, 2006) when the test items are well constructed. MC tests can efficiently test a thorough sample of the domain of knowledge at higher cognitive levels (i.e., application, synthesis, and evaluation)

(Aiken, 1987; Downing, 2006) and are considered superior to other formats for this purpose (Haladyna & Downing, 1989a). MCQs are easily analyzed with computer software to produce item-response statistics that enable test writers to evaluate construct validity data and improve the quality of individual items (Aiken, 1987). Use of a test blueprint to demonstrate correlation between learning outcomes, program standards, and test items is an important component of test construction that must be incorporated to ensure test validity (Layton, 1986; Morrison et al., 2006). These processes can reasonably be accomplished with MC tests.

Discrimination. The final requirement for effective assessments is whether the test is able to discriminate between the more knowledgeable and the less knowledgeable test-takers (Quinn, 2000). Highly discriminating test items are desirable and tend to produce high score reliability (Downing, 2005). If a test makes no discrimination between students, then it has no purpose (Quinn & Hughes, 2007). Discrimination is established through analysis of item-response statistics – item difficulty and the point biserial. Item difficulty is calculated according to the proportion of students answering the item correctly (Downing, 2005). The point biserial is a discrimination index determined by comparing the proportion of students in the top and bottom 27% of the grade distribution who selected the correct response (Weaver, 1982).

Well-constructed test items are essential for discrimination. Poorly written or easily guessed questions have low discrimination, because the correlation between selecting the correct response and the ability or knowledge of the examinee is difficult to determine – correct responses may be selected simply by chance or good guessing (Campbell, 2011). The more effective the distracters an item contains, the more

discriminating the item; therefore, an important component of item analysis is reviewing the frequency of distracter selection and improving poor performing distracters (McDonald, 2014). If properly constructed, MCQs can accurately discriminate between high- and low-performing examinees (Boland et al., 2010; Tarrant & Ware, 2008).

Well-written MCQs meet Quinn's (2000) Cardinal Criteria for Assessment as one component of a comprehensive testing program. MCQs are efficient, objective, easy to grade, and can be used to test a broad sampling of the curriculum at higher cognitive levels (Brady, 2005). When test items are constructed according to discipline-specific guidelines and incorporate a process of planning and analysis, the criteria of practicality, reliability, validity, and discrimination can reasonably be accomplished with MC assessments.

Bias in Testing

An essential criteria that is absent from Quinn's (2000) framework is that an effective assessment must be free of bias. Testing bias occurs when test results contain error because of factors unrelated to the purpose of the exam. Sources of error include the student, the environment, scoring factors, and the test itself (Gaberson, 1996). When an examination is biased, students perform differently based on variables that are unrelated to their knowledge and abilities. Bias may advantage or disadvantage, resulting in artificial inflation or deflation of test scores (Scheuneman, 1984). A test item is fair when it is free of bias, and students of equal ability are equally likely to answer it correctly (Klisch, 1994). The reliability, validity, and discrimination of assessments are reduced when biased test items are present.

Bias is traditionally conceptualized as a factor relating to performance that differs because of membership in a group (Camilli & Shepard, 1994; Hambleton & Rodgers, 2005). This definition provides a framework upon which large-scale testing programs analyze test data according to demographics to determine if an item functions differently for different ethnic, gender, cultural, or religious groups. The National Council of State Boards of Nursing (NCSBN) employs such a differential item functioning (DIF) process to evaluate all test items on the NCLEX (O'Neill, Marks, & Liu, 2006; Wendt & Worcester, 2000; Woo & Dragan, 2012). This process involves statistical analysis and comparison of majority (White and female) and non-majority groups and evaluation of the items and statistics by a diverse panel of experts (Wendt & Worcester, 2000; Woo & Dragan, 2012). Other standardized testing programs employ similar procedures to ensure that tests are as free from bias as possible.

Schools of nursing do not have the resources that are available to the NCSBN and other standardized testing services for evaluation of test items based on a definition of bias as a function of group membership. While nursing programs do conduct analysis of aggregate student data, the primary concern is in facilitating the education and success for individuals and ensuring that course and program assessments are therefore reliable, valid, and fair measurements of individual student learning. A definition of bias that only concerns the performance of students based on group membership is not adequate for the purposes of assessment in nursing education programs.

Scheuneman (1984) conceptualized bias as affecting individuals rather than only “members of particular subgroups” (p. 221), which is a more meaningful depiction of bias for the purposes of nursing assessments. According to Scheuneman, bias is a

multidimensional concept consisting of elements of the test content, test items, reading level, test environment, administration procedures, and the examinee. The following model represents the relationship of bias and the test score:

$$X = \theta + \beta + \varepsilon$$

In this equation, an examinee's observed score (X) is equal to the true score (θ) plus the bias quantity (β) plus any measurement error (ε) that occurs in the administration of the test. According to this model, there is a link between bias in items and bias in test scores. The true score is a value that represents an individual's true knowledge or ability being measured on the assessment (Nibert, 2003), while the observed score is the score achieved on the examination. Measurement error is considered to be random error that is always present and which cannot be controlled. The bias quantity represents systematic error within tests and test items, which can be controlled. Bias can advantage or disadvantage, resulting in an increased or decreased observed score, respectively. With any given assessment, the desired outcome is one in which the observed score accurately reflects the individual's knowledge and ability. Reducing the bias that is a characteristic of the assessment will result in an observed score that is representative of the true score.

According to Scheuneman's (1984) conceptualization of bias as affecting individuals, a biased item is any test item that is potentially unfair to examinees. A test item is biased if it might disadvantage some students more than others based on variables that are irrelevant to the construct being tested. A test item is free of bias if students of equal ability are equally likely to answer it correctly. Item bias results from flawed test items and may be linguistic, structural, or cultural in nature or related to irrelevant difficulty, testwise cues, or formatting errors. The goal of test development, then, is to

ensure that tests and test items are unbiased so that an individual student's earned score is a true measure of learning.

Revised Framework for Quality Assessment

Combining Quinn's (2000) Cardinal Criteria of Assessment and Scheuneman's (1984) conceptualization of bias provides a more complete theoretical framework upon which to evaluate assessments. According to this revised framework, every effective assessment must meet the following criteria:

Valid – the test measures what it is designed to measure, leading to meaningful inferences from the scores;

Reliable – the test consistently measures what it is designed to measure;

Discriminating – the test distinguishes between the more knowledgeable and the less knowledgeable students;

Practical – the test is useful and practical for its purposes;

Unbiased – the test is fair to examinees and contains items that students of equal ability are equally likely to answer correctly.

Well-written MCQs are designed to fulfill these criteria as one component of a comprehensive testing program. In order to construct well-written MCQs, discipline-specific guidelines need to be used and a process of planning and analysis incorporated. The focus of this dissertation study was to evaluate a discipline-specific guideline for nursing faculty to use in developing MC test items that will meet the requirements for quality assessments.

Framework for Test Development

The process of item writing and evaluation used in this research proposal is based on The Conceptual Model for Test Development (see Figure 1). The conceptual model for item writing and test construction was originally developed by Nibert (2003), modified by Morrison et al. (2006), and adapted for this dissertation research study. The model identifies a clear process for constructing high quality test items within the domain of nursing. Application of this model provides nursing faculty with discipline-specific guidelines and a systematic process by which to develop reliable, valid, discriminating, and unbiased assessments of student learning. The Conceptual Model for Test Development depicts item writing and test development as a three-phase process: Exam Creation, Test Item Writing, and Exam Evaluation. This model provides a foundation for the development of high quality MC test items. Each phase of the process informs the other stages. The identification and revision of biased items, for which the Fairness of Items Tool (FIT) is designed, is an integral component of the last two phases of the model.

Exam creation phase. In the first phase of the process, the examination is purposefully planned, beginning with an understanding of the purpose of the assessment and the constructs being tested. The purpose of any nursing examination is to assess whether students have achieved the desired learning outcomes and to “predict the entry-level performance” (Morrison et al., 2006, p. 13). The content to be tested is determined through examination of the guiding constructs – course objectives, program outcomes, and professional standards – and informed by the clinical knowledge of expert nursing

faculty (Morrison et al., 2006). The test blueprint demonstrates the organization of the assessment and its relationship to the guiding constructs.

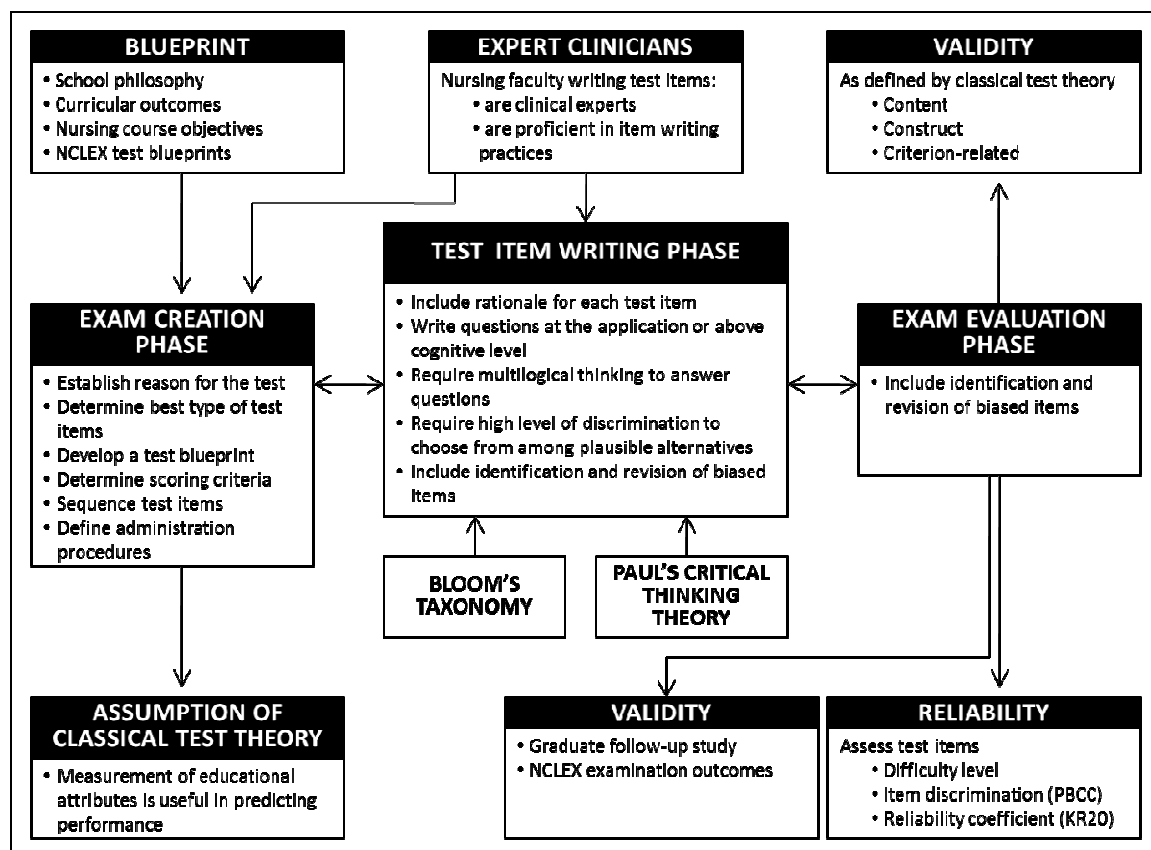


Figure 1. Conceptual Model for Test Development. Adapted with permission from the HESI Model for Developing Critical Thinking Test Items in Morrison, S., Nibert, A., & Flick, J. (2006). *Critical Thinking and Test Item Writing* (2nd ed.), p. 12. Houston, TX: Health Education Systems, Inc.

Item writing phase. During the item-writing phase, test items are written and/or edited to develop high quality items that evaluate achievement of learning outcomes within the domain of nursing. Item writing is informed by the knowledge of faculty with clinical and item-writing expertise, the test blueprint, classic test theory, Bloom's taxonomy, and critical thinking theory. Note that the arrow between phases is bi-directional, indicating that the process is not linear – rather, each phase in the item and

test development process informs previous and subsequent phases, resulting in developing, evaluating, revising, and “tweaking” the products in each phase (Morrison et al., 2006, p. 27).

Morrison et al. (2006) identify four criteria used in developing objective nursing test items: (1) include rationale for each test item; (2) write questions at the application or above cognitive level; (3) require multilogical thinking to answer questions; and (4) require a high level of discrimination to choose from among plausible alternatives (p. 12). Identifying and revising biased test items is included in the adapted model as the fifth criterion to emphasize its importance in the item-writing phase for developing reliable, valid, discriminating, and unbiased assessments of student learning.

Exam evaluation phase. Test items and examinations are evaluated through analysis of pre- and post-administration reliability and validity data. Pre-administration evaluation includes peer review of items and tests and piloting and/or pretesting items. Post-administration evaluation includes analysis of item-response statistics – item difficulty, point biserial (discrimination index), and distracter effectiveness – and the reliability coefficient (KR-20). When biased test items are identified, they are revised, and the process of evaluation continues. In addition, exam evaluation involves analyzing broader implications of nursing assessment, such as NCLEX pass rates and success of graduates.

Summary of Theoretical Literature

Three themes have emerged from the analysis of theoretical literature: (1) high quality MCQs are necessary for reliable, valid, discriminating, and unbiased assessments of student learning; (2) item quality is improved through item-writing procedures,

obtaining pretest reliability data, and using post-administration analysis data to guide revision; and (3) test quality is improved through adequate planning of assessments and development of a test blueprint. The Conceptual Model for Test Development provides a process by which these effective, high quality MC test items can be developed. The Fairness of Items Tool (FIT) is designed to assist faculty in identifying bias in MC test items and is therefore an integral component of the item-writing and exam evaluation phases of test development.

Empirical Literature

An exhaustive search of the literature from 1978-2014 resulted in a collection of 48 references containing guidelines for developing MCQs. This collection includes 20 journal articles from education and assessment literature, including the educational specialties of accounting, law, marketing, medicine, psychiatry, pharmacy, and teacher education; 18 journal articles from nursing education; and 10 books with nursing and medical education and educational assessment foci. Item-writing guidelines are presented in paragraph form, tables, and checklists of various lengths containing from 8 to 30 item-writing rules. The intent of these publications is to provide discipline-specific guidelines for faculty development of MC test items. Because these guidelines are primarily the product of experience and not based on systematic reviews of literature, they are not included in this review of empirical literature. It is important to note, however, that there is such a proliferation of item-writing guidelines in textbooks and journal articles that educators are unable to make sense of all of them; therefore, misunderstandings and poor item-writing practices inevitably occur.

There is also a significant body of empirical literature devoted to item-writing guidelines. The literature search yielded 34 research studies that developed or used item-writing guidelines, 12 of which were nursing discipline-specific.

Taxonomy Development

The most comprehensive work on taxonomy development has been conducted by Haladyna and Downing (1985; 1989a; 1989b) whose original taxonomy contained 43 MC item-writing rules (Haladyna & Downing, 1985) and was revised to 31 in its final version (referred to hereafter as Revised Taxonomy) (Haladyna, Downing, & Rodriguez, 2002). The researchers systematically analyzed the literature to achieve a taxonomy that completely represents the general array of advice on preparing MC items (Haladyna & Downing, 1989a). The taxonomy was validated through examination of the empirical support for each item-writing rule through multiple iterations in 1985, 1989, and 2002. When there was little empirical data to support a rule, validity was established by consensus of the literature. The strength of the Revised Taxonomy is its simplicity, enabling understanding of each guideline with minimal explanation. The Revised Taxonomy is complete in the sense that it provides general guidance for item development; however, it is not discipline-specific and therefore does not represent the complete array of advice necessary for a practice discipline such as nursing.

Spanish researchers Moreno, Martínez, & Muñiz (2004) condensed the Revised Taxonomy into a set of 12 guidelines, citing a need for a more concise list containing fewer guidelines and eliminating overlap (Moreno et al., 2006). This taxonomy was then reformatted to a set of 15 guidelines for developing MC items (referred to hereafter as New Guidelines) after a validation study with a sample of 29 experts and 51 users in

Spain (Moreno et al., 2006). While the Revised Taxonomy (Haladyna et al., 2002) contains clear, concise guidelines, the New Guidelines cannot stand alone. The item-writing rules are too consolidated, resulting in vague generalities that are not easily understood without referencing more detailed explanations. Some of the research on the New Guidelines was published in Spanish and was therefore not accessible for this literature review. It is likely that the lack of clarity in the New Guidelines is related to issues with translation, limiting its usefulness in the English language.

The work of Haladyna and Downing (1985; 1989a; 1989b) and Haladyna et al. (2002) focused on the development of a general taxonomy for MC item writing and represents the seminal work on taxonomy development. Several researchers have published guidelines for other purposes, with many citing the work of Haladyna, Downing, and colleagues. Discipline-specific taxonomies have been developed in accounting and teacher education research. Researchers in nursing education have developed sets of guidelines containing from 20 to 52 item-writing rules established through reviews of nursing and assessment literature (Bosher, 2003; Klisch, 1994; Masters et al., 2001; Van Ort & Hazzard, 1985).

Discipline-specific guidelines. Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) developed a list of 40 item-writing rules for teacher education, 30 of which apply to the development of MCQs. These rules were established through review of 20 educational assessment textbooks and “standard reference works” (p. 359) and represent the most common item-writing rules from the reviewed literature. Nineteen of the 30 MCQ guidelines included in the taxonomy were consistent with those in the Haladyna et al. (2002) Revised Taxonomy. Frey et al.’s work contributes additional guidelines

applicable to other types of objective testing, such as matching, true-false, and completion items. Frey et al. confirm that the work of Haladyna, Downing, and colleagues was “exhaustive” (p. 358), and their literature review yielded only more recent empirical studies providing additional support for the previously validated rules in the Revised Taxonomy.

Ellsworth, Dunnell, and Duell (1990) developed a research instrument for the purpose of comparing the MC item-writing guidelines published in textbooks with the test bank items included with those text books. A review of 42 undergraduate educational measurement and psychology textbooks, 32 of which contained clear guidelines, yielded a list of 37 different guidelines for writing MC items (Ellsworth et al., 1990). This list was subjected to a selection process that considered recommendation by 50% or more of the texts, that did not require excessive analysis of textbook content, and included all guidelines associated with testwiseness (Ellsworth et al., 1990). The resulting research instrument contained 12 of the most-cited guidelines and was used to evaluate 1,080 randomly selected MC items from the textbook test banks.

Hansen and Dexter (1997) consulted “several books and articles” (p. 94), including those by Ellsworth et al. (1990) and Haladyna and Downing (1989) to develop a list of 17 guidelines for use by accounting and business faculty in the development of MC test items. The list was piloted as a research instrument with 40 MCQs selected from textbook test banks and was ultimately used to evaluate the quality of 440 MCQs in auditing test banks (Hansen & Dexter, 1997). Additional discipline-specific lists have also been developed through review of literature for medicine (Al-Faris et al., 2010; Boland, Lester, & William, 2010) and academic psychiatry (Begum, 2012).

Development in nursing. The earliest work on nursing guidelines was reported by Van Ort and Hazzard (1985), who developed a guide for faculty at the University of Arizona College of Nursing. The researchers reported consulting two published references – a National League for Nursing (NLN) publication and a nursing journal article – as well as “materials from test construction workshops” to identify criteria for inclusion in the checklist (p. 15). Criterion were categorized, ranked in order of importance, and limited to no more than 12 in any category. The resulting checklist contained 35 guidelines for evaluation of test items that were used to make course-specific improvements in the quality of examination items (Van Ort & Hazzard, 1985).

Cultural diversity. Discipline-specific guidelines have been developed in nursing as a result of interest in improving outcomes for diverse student populations, specifically non-native speakers of English – also known as English as additional language (EAL) (Lampe & Tsaouse, 2010). Klisch (1994) investigated sources of bias in nursing examinations for EAL students by conducting telephone interviews with testing experts at the National Council Licensing Examinations (NCLEX) and faculty in nursing, cultural diversity, education, psychology, and sociology. Klisch’s research led to the development of 20 guidelines for reducing item bias in nursing examinations. These guidelines were then used to evaluate the quality of test items and develop a faculty development program in the author’s nursing program.

Bosher (2003) investigated test item flaws that might have a negative impact on EAL students. Bosher consulted three assessment texts, including the work of Haladyna, to develop a partial list of test item flaws. The resulting 52 subcategories of item flaws emerged from a review of 19 exams with 673 MC test items from one nursing program.

The most common test item flaws were assembled into a list of 25 Criteria for Test Questions, the purpose of which was to improve the quality of test items constructed within the author's nursing program (Bosher, 2003).

Research instruments. Nursing taxonomies have also been developed for specific research purposes. Masters et al. (2001) evaluated the quality of 2,913 nursing test bank questions with a research instrument developed through review of literature that included Ellsworth et al. (1990) and four nursing journal articles. The research instrument consisted of 30 guidelines, including two guidelines that emerged during the study (Masters et al., 2001). No systematic means for selecting the references or guidelines were described, except that "all published guidelines that made sense educationally and did not require close textbook examination were used" (Masters et al., 2001, p. 27).

Tarrant et al. (2006) developed a taxonomy of 32 item-writing guidelines identified through review of the "most cited sources for MCQ construction" (p. 356), the majority of which were contributed by Haladyna, Downing, and their colleagues. The research instrument was used to evaluate a random sub-sample of 250 MCQs from examinations in one English-speaking nursing department in Hong Kong (Tarrant et al., 2006). Violations of 19 item-writing guidelines were found, and these 19 guidelines were then used to evaluate the quality of 2,770 MCQs (Tarrant et al., 2006).

Naeem, van der Vleuten, and Alfaris (2012) developed a checklist of 21 item-writing guidelines to evaluate the quality of MC test items submitted by faculty in medicine and nursing in Saudi Arabia. The guidelines were identified through review of the literature and included the work of Haladyna, Downing, and colleagues and Frey et al. (2005). The research instrument was used to evaluate the quality of MC test items

submitted by 51 faculty members to demonstrate the effectiveness of a faculty development program in item writing.

Analysis. While there is much agreement and overlap between the taxonomies and guidelines, the development of multiple item-writing taxonomies contributes to confusion about their validity and utility. These guidelines do not meet the need for a discipline-specific taxonomy for use by nursing faculty in the development and evaluation of MCQs. Other than the extensive work in the development of the Revised Taxonomy (Haladyna et al., 2002), there has been very little evaluation of the evidence supporting item-writing guidelines, much of which is based on common practice and frequency of recommendation, rather than empirical data (Frey et al., 2005; Haladyna et al, 2002; Haladyna & Downing, 1989b).

Validity and reliability. Procedures to establish the validity of the item-writing taxonomies were described in most of the studies, the most common of which was the frequency with which a guideline or rule appeared in the reviewed literature (Frey et al., 2005; Hansen & Dexter, 1997). Frey et al. (2005) also ranked the relative importance of the guidelines by their frequency in the reviewed literature and included only guidelines that were mentioned more than once. Tarrant et al. (2006) validated guidelines by examining the frequency that item-writing violations appeared in a sample of MC test items in nursing.

Haladyna and Downing (1989b) and Haladyna et al. (2002) incorporated the most rigorous methods for establishing validity. To validate the first taxonomy (Haladyna & Downing, 1989a), Haladyna and Downing (1989b) examined the levels of supporting evidence for each rule and documented the frequency with which each rule was studied,

rated the effect of using or not using the rule, and finally reached researcher consensus about the validity of each rule. When validating the Revised Taxonomy, Haladyna et al. considered two sources of evidence for each guideline – expert consensus of textbook authors and empirical research from 19 research studies. One problem these researchers encountered is that published research reports contained inconsistent data, making aggregation difficult (Haladyna & Downing, 1989b; Haladyna et al., 2002). While all of the guidelines in the Revised Taxonomy (Haladyna et al., 2002) were validated by some level of evidence, only four specific rules were supported without contradiction by empirical research, and two of these were supported by evidence from only one study. However, the majority of the guidelines in the Revised Taxonomy (24 of the 31 item-writing rules) were supported by unanimous author endorsements (Haladyna et al., 2002).

The validity evidence for the majority of the MC item-writing guidelines is limited to expert consensus, rather than empirical research. In lieu of empirical evidence to support these guidelines, Haladyna et al. (2002) and Frey et al. (2005) conclude that expert consensus provides a solid base for a theoretical approach for item development. Haladyna et al. advocate for continued systematic research studies, especially using experimental methods, with corresponding revisions to the taxonomy as warranted.

Since the publication of the Revised Taxonomy (Haladyna et al., 2002), 11 published research studies have tested one or more specific item-writing rules, the majority of which have used non-experimental methods (Ascalon, Meyers, Davis, & Smits, 2007; Caldwell & Pate, 2013; Hays, Coventry, Wilcock, & Hartley, 2009; Martínez, Moreno, Martín, & Trigo, 2009; Odegard & Koen, 2007; Piasentin, 2010; Redmond, Hartigan-Rogers, & Cobbett, 2012; Rodriguez, 2005; Rogausch, Hofer, &

Krebs, 2010; Tarrant & Ware, 2010; Taylor, 2005). No published research studies have evaluated the validity and reliability of the taxonomies for identifying bias in MC test items. No published studies have been conducted to validate a discipline-specific taxonomy for nursing education.

Dimensions of Bias

For the purposes of this dissertation research study, item bias is any test item that is potentially unfair to examinees. A test item is biased if it might disadvantage some students more than others based on variables that are irrelevant to the construct being tested. A test item is free of bias if students of equal ability are equally likely to answer it correctly. Item bias results from the presence of poorly crafted test items that contain one or more errors in construction that violate item-writing guidelines.

There is agreement in the literature that item bias is a multidimensional concept. Haladyna et al.'s (2002) Revised Taxonomy contains eight guidelines addressing content concerns; five guidelines addressing formatting and style concerns; four guidelines for writing the stem; and 14 guidelines for writing the choices, one of which has six variations. Moreno et al. (2006) grouped their taxonomy according to foundations prior to item construction, general criteria for construction of the item and test, and construction of response options. Dimensions identified in Frey et al.'s (2005) taxonomy include content, clarity, guessing, efficiency, and testwiseness. Boshier (2003) identified four categories of biased test items: testwise flaws, irrelevant difficulty, linguistic/structural bias, and cultural bias. Structural and cultural biases were identified by Klisch (1994). Dimensions common to all of the reviewed taxonomies include testwise cues, flaws of

irrelevant difficulty, linguistic/structural bias, and general/content issues. Cultural bias is an additional dimension noted in the nursing literature.

Prevalence of Flawed Items

The taxonomies and guidelines have been used in non-experimental descriptive studies to identify the frequency of item-writing violations in MC assessments for multiple disciplines. High levels of flawed and biased test items have been identified in textbook item banks and examinations.

Analysis of test item banks. Ellsworth et al. (1990) compared the MC item-writing guidelines published in textbooks with their test bank items in order to evaluate the test items being modeled to future teachers. A research instrument containing 12 guidelines was developed from a review of 32 educational measurement and educational psychology textbooks (Ellsworth et al., 1990). A sample of 60 test items was randomly selected from each of 18 educational psychology test item banks for a total of 1,080 MC test items (Ellsworth et al., 1990, p. 290). Over 60% ($n = 653$) of the items contained violations of at least one guideline (Ellsworth et al., 1990, p. 291). Averaging across all of the textbooks, less than 40% ($n = 23.72$) of the items from each test bank were developed according to the guidelines, and nearly one-third of the items ($n = 18.67$) in each test bank contained grammatical errors (Ellsworth et al., 1990, p. 291). The quality of test banks varied across the different textbooks, and the “best” test bank contained 23 out of 60 items (36.67%) with at least one guideline violation (Ellsworth et al., 1990, p. 291). There was no analysis of whether these test items were used for examination purposes without revision, but the researchers speculated that many test bank items were administered unedited (Ellsworth et al., 1990).

Hansen and Dexter (1997) evaluated the quality of 440 MCQs randomly selected from auditing test banks ($n = 400$) and prior certified public accountant (CPA) examinations ($n = 40$) using 17 guidelines developed for use by accounting and business faculty. At least one violation was found in 75% ($n = 299$) of the test bank questions and 30% ($n = 12$) of the CPA examination questions (Hansen & Dexter, 1997, p. 96).

Masters et al. (2001) evaluated the quality of 2,913 nursing test items randomly selected from 17 test banks. Each question was evaluated for cognitive level and consistency with 30 guidelines (Masters et al., 2001). There were 2,233 guideline violations recorded, and some questions contained multiple violations (Masters et al., 2001, p. 28). The most common violation was inadequate spacing, which occurred in 33% ($n = 960$) of the test items and was contained within 4 of the 17 test banks, with 73 to 263 items in each test bank affected (Masters et al., 2001, p. 28). The researchers noted that “most” test banks contained some items in violation of the guidelines, but these violations tended to follow a pattern in which limited types of violations tended to be “pervasive” when present within a test bank (Masters et al., 2001, p. 29). Individual guideline violations occurred in a range of 1 to 14 test banks, so quite a bit of variation was evident. This study also found only 28.3% of the test bank items written at the application or above cognitive level (Masters et al., 2001, p. 27).

Statistical analysis in the test bank studies was limited and did not include significance levels. The findings suggest that poor quality test items are being modeled to future teachers in each of these disciplines. Faculty should not rely on test banks to contain high quality test items, and the items must be thoroughly reviewed and revised, using discipline-specific guidelines, before they are administered.

Analysis of examinations. Research evaluating the quality of previously administered examinations was consistent with the analysis of test bank items. Bosher (2003) analyzed 19 nursing course exams totaling 673 MCQs in her investigation of linguistic and cultural bias. Each MCQ was systematically analyzed for 52 subcategories of item flaws, and those within each category that occurred at least 10 times were identified in the report. Examples of each of the 28 types of commonly occurring flaws were presented and recommendations for correction discussed (Bosher, 2003). Flaws of irrelevant difficulty comprised 61% of the errors (807 occurrences), linguistic/structural bias comprised 35% of the errors (145 occurrences), testwise flaws comprised 3% of the errors (31 occurrences), and cultural bias occurred in less than 1% of the errors (26 occurrences) (Bosher, 2003, p. 33). While this research report did not present a statistical analysis of these findings, it is apparent that multiple questions contained more than one error. In addition, it is likely that additional flaws were present but did not reach enough significance (occurring at least 10 times) to be reported.

Cross (2000) conducted a nation-wide analysis of the quality of teacher-made tests in nursing education in her dissertation research. A total of 110 examinations from 61 programs in 29 states were included with MCQs comprising 91.9% of the test items (Cross, 2000). Quality was defined in this research study by seven indicators of appearance and format: the presence of directions to the test-taker, indication of point values for test items, neatness and legibility, consecutive numbering of pages and items, and the presence of typographical or usage errors. Descriptive data are included for the analysis of quality, and format and mechanics errors were common (Cross, 2000). In Cross' analysis, only 20.9% of the examinations were free of typographical or usage

errors, problems with item sequencing were “common”, and 19.1% of examinations had problems with page sequencing, including missing pages (p. 49). While these quality indicators are not indicative of validity and bias in the individual test items, inattention to overall formatting issues introduces construct-irrelevant difficulty and lowers the validity of the test scores, regardless of item quality.

Tarrant et al. (2006) conducted a descriptive study in an English-speaking baccalaureate program in Hong Kong over a five-year period. The quality of MCQs used in nursing assessments was evaluated using the 19 most frequently occurring guidelines identified through review of literature and a random sub-sample of MCQs (Tarrant et al., 2006). Of the 2,770 MCQs analyzed, 46.2% ($n = 1,280$) contained at least one item-writing flaw, and 12% ($n = 341$) contained multiple flaws (Tarrant et al., 2006, p. 357). MCQs written at lower cognitive levels (recall/comprehension) were significantly more likely to contain flaws than items written at higher cognitive levels ($p < .001$) (Tarrant et al., 2006, p. 358).

Nedeau-Cayo, Laughlin, Rus, and Hall (2013) replicated Tarrant et al.’s (2006) study in the United States using a sample of MCQs from tests in a midsize acute care hospital’s elearning system with similar results. Of the 2,491 MCQs analyzed, 49.9% ($n = 1,243$) contained one item-writing flaw, and 34.9% ($n = 862$) contained more than one flaw (Nedeau-Cayo et al., 2013). Almost 94% of the items ($n = 2,332$) were written at lower cognitive levels (recall/comprehension), and these items were also more likely to contain flaws (94.4% of items, $p = .0008$).

Downing (2002a; 2005) found a similarly high prevalence of flawed test items in medical school examinations using Haladyna et al.’s (2002) Revised Taxonomy. In the

first study, one examination contained errors in 11 of the 33 MCQs (33%) (Downing, 2002a). A second larger study analyzed four examinations totaling 219 MCQs (Downing, 2005). Flawed test questions comprised 36% to 65% of the test items on each of the four tests (Downing, 2005, p. 133), and there were a total of 100 (46%) flawed items (p. 137).

Jozefowicz et al. (2002) included the presence of item flaws as one component of a quality rating in their analysis of medical school examinations. Nine examinations from three different medical schools provided a sample of 555 questions, including all item types. Items were rated independently on a 5-point scale by expert test developers who were blinded to the question writers and study hypothesis. Item flaws were assessed according to Haladyna & Downing's (1989a; 1989b) original taxonomy (Jozefowicz et al., 2002). The mean quality assessment score (QAS) for all questions was 2.39 ± 1.21 ; School A ($n = 222$) had a mean QAS of 1.94 ± 0.90 , School B ($n = 180$) had a mean QAS of 3.26 ± 1.28 , and School C ($n = 153$) had a mean QAS of 2.03 ± 0.94 (Jozefowicz et al., 2002, p. 157). Not only was the overall quality of the test items low, but there was significant ($p < .001$) variation in the quality of test items between the schools (Jozefowicz et al., 2002, p. 158). This research suggests that standard adoption of item-writing guidelines may improve the consistency of examination quality across programs.

Further evidence. Research studies conducted with MCQs also provide evidence of the prevalence of flawed and biased test items in nursing examinations. Schroeder (2007) used a quasi-experimental design in a dissertation study to determine if student training in test taking and the use of MCQs written at higher cognitive levels would result in improved scores on examinations. The research found no significant improvement in test scores with the higher level MCQs. The research report contained no analysis of the

quality of the MCQs used in the study. Four example MCQs were included in the dissertation report with a reference indicating that they were taken from a textbook test bank. These sample MCQs contained multiple item-writing flaws with the cognitive levels incorrectly identified.

Kelly's (1998) dissertation study used a qualitative design to compare the ability of MC and constructed response test items to demonstrate critical thinking. Both test formats enabled students to demonstrate critical thinking, and there was no correlation between the formats related to student performance (Kelly, 1998). No example questions are included in the report; however, neither is there an analysis of the quality of any of the test items used in the study. It is logical that the quality of the research findings may be affected by the quality of the test items, yet none of the dissertation studies reviewed in this section addressed issues of item quality (Cross, 2000; Kelly, 1998; Schroeder, 2007). Having a discipline-specific tool available for use in evaluating the quality of MCQs will also benefit researchers and improve the quality of research conducted with MCQs, as well as the reliability and validity of the findings. The lack of attention to item quality evident in these studies also calls into question the researchers' knowledge of item-writing principles and whether this is an indication of an overall lack of knowledge among nurse educators.

Impact of Flawed Items

Flawed test items interfere with accurate and meaningful interpretation of test scores and negatively affect students' passing rates. A significant number of students fail high-stakes examinations because of their performance on flawed test items. Previous research has demonstrated that very low and high achieving students received lower

examination scores overall when flawed items were present, while borderline students tended to have improved test scores, presumably because of guessing and testwiseness. Progression decisions rely heavily on examination scores, and it is therefore important to ensure that examinations are valid and reliable assessments of student mastery.

Downing (2002a; 2005) analyzed the effects of flawed test questions on item and test difficulty and grading decisions based on tests containing flawed questions. In both studies, standard items did not contain violations of the Revised Taxonomy (Haladyna et al., 2002) and flawed items contained one or more violation. Test items were classified as standard or flawed, and statistical analysis was conducted on each group of questions for each test independently and combined (Downing, 2002a; Downing, 2005). Flawed items were more difficult than standard items in four of the five examinations studied, averaging 6.6 percentage points more difficult than the standard items (Downing, 2002a; Downing, 2005). Analysis of the pass-fail decisions in the second study revealed that, of the 749 students taking the examinations, 102 (14%) passed the standard items but failed the flawed items, while 30 students (4%) passed the flawed items and failed the standard items ($p < .0001$) (Downing, 2005, p. 141). Downing (2005) concluded that as high as 10% to 15% of students were incorrectly failed due to flawed test items.

Tarrant and Ware (2008) conducted a similar study to examine the impact of item-writing flaws in 10 high stakes examinations in an English-speaking nursing school in Hong Kong. In this study, a total scale was computed to reflect the test as it was administered (with flawed items), and a standard scale was computed for a “hypothetical test” that contained no flawed items (Tarrant & Ware, 2008, p. 200). Mean item difficulty and mean discrimination were calculated for each scale with varying results – flawed

items ranged from 10 percentage points more difficult to 8 percentage points less difficult than standard items, and flawed items were less discriminating on 7 of the 10 examinations (Tarrant & Ware, 2008). These findings suggest a complex relationship between flawed items and student achievement and are reflective of the variation in numbers and types of flawed items in the individual examinations. Out of 824 examinees, 90.9% ($n = 749$) passed the standard scale, compared with 94.5% ($n = 779$) passing the total scale with the flawed items (Tarrant & Ware, 2008, p. 201). On both scales, 90.2% ($n = 743$) passed and 4.7% ($n = 39$) failed, with a passing standard of 50%; however, 36 additional examinees would have failed if flawed items had been removed (Tarrant & Ware, 2008, p. 201). The proportion of high-achieving students with scores at or above 80% was higher on the standard scale: 21% ($n = 173$) of examinees on the standard scale versus 14.6% ($n = 120$) on the total scale (Tarrant & Ware, 2008, p. 201). On both scales, 11.7% ($n = 96$) of the examinees scored 80% or above, and 76.1% ($n = 627$) scored less than 80%; however, 77 additional examinees would have scored 80% or greater if the flawed items had been removed from the examinations (Tarrant & Ware, 2008, p. 202). These findings suggest that borderline students benefitted from the flawed items, while high achieving students were negatively affected (Tarrant & Ware, 2008).

The differences in the findings between these studies may be related to a difference in analysis, with Downing (2002a; 2005) using a standard and flawed scale and Tarrant and Ware (2008) using standard and total scales. These studies were conducted with non-experimental, descriptive methods, which limit the conclusions that can be drawn from the findings; however, other research studies provide evidence that poorly constructed test items introduce construct-irrelevant variances that prevent proper

interpretation of test scores. Caldwell and Pate (2013) conducted a quasi-experimental study to determine the effect of three item-writing flaws on item statistics and student performance in a pharmacy program. This study was designed to further examine select item-writing guidelines from Haladyna, Downing, and Rodriguez' (2002) taxonomy that were not strongly endorsed in the literature: (1) Word the stem positively; avoid negatives such as *not* or *except*; (2) Develop as many effective choices as you can, but research suggests 3 is an adequate number; (3) None-of-the-above (NOTA) should be used carefully (Caldwell & Pate, 2013, p. 2). For this study, pairs of MC test items were developed for each guideline and added to the end of a course examination. The standard items were written according to the guidelines, and the nonstandard items violated the guidelines. There was some randomization as the two versions of the test were distributed alternatively, but students were allowed to select their seats. Results demonstrated that students were more successful answering the standard items (71% compared with 47% for the nonstandard items), and there was no difference in item discrimination ($p = .22$). For this study, the presence of flawed items increased item difficulty but did not improve item discrimination, both of which prevent proper interpretation of scores.

Improving Item Writing

Multiple authors have recommended implementation of a systematic process for item writing similar to the Conceptual Model for Test Development (Tarrant et al., 2006; Tarrant & Ware, 2012). There is some research to support that the quality of MC items can be improved through faculty training in principles of item writing and the use of pre-established guidelines. Implementing a pre-test peer review process has also resulted in improved test item quality.

Faculty training. In Jozefowicz et al.'s (2002) comparison of the quality of in-house medical school examinations, questions written by trained item writers ($n = 92$) had a mean quality score (QAS) of 4.24 ± 0.85 , compared with a mean QAS of 2.03 ± 0.90 for questions ($n = 463$) written by writers without training ($p < .001$) (Jozefowicz et al., 2002, p. 157). The significantly higher QAS achieved by School B was attributed to the fact that 44% of the questions submitted were written by a trained item writer (Jozefowicz et al., 2002, p. 157). All of the item writers in this study were trained through the National Board of Medical Examiners. These findings confirm those of Hansen & Dexter's (1997) comparison of CPA exam questions and undergraduate accounting and marketing test bank items. Item writers for the CPA examination receive training, and CPA exam questions undergo a review process prior to being used on the CPA examination (Hansen & Dexter, 1997). These findings suggest that faculty training in item writing has significant potential for improving the quality of test items, and providing training according to consistent discipline-specific guidelines is also beneficial.

Naeem et al. (2012) provide further confirmation of the immediate value of faculty development to improve the quality of MC test items. Items submitted by 51 faculty in medicine and nursing were evaluated according to an objective checklist at three points during a structured faculty development workshop: pretest, midtest, and posttest. The items evaluated pretest were submitted by faculty as their "best effort" and evaluated prior to faculty development (Naeem et al., 2012, p. 371). Test items were revised twice during the workshop: once based on facilitator feedback (midtest) and a second after peer review (posttest). Results of the study demonstrated a significant improvement in test items from pretest to posttest ($p < .0005$) with large effect sizes. In

addition to demonstrating the benefit of faculty development, the researchers concluded that “items written by faculty without faculty development are generally lacking in quality” (Naeem et al., 2012, p. 369).

Research also demonstrates that this immediate benefit of faculty development has a long-term impact on the quality of MC test items. Khan, Danish, Awan, and Anwar (2013) investigated the presence of flawed items in 2009, 2010, and 2011 on medical college examinations in Pakistan using guidelines from the National Board of Medical Examiners. Faculty received training in item development in 2009, and evaluation of the test items demonstrated significant improvement from year one to year three. A total of 4,550 MCQs were evaluated during the three-year period, and the presence of item-writing flaws in each year was 67%, 36%, and 21% respectively. These findings suggest that, in addition to education about principle of item writing, having time for practice and repetition is necessary for long-term improvements in test item quality.

Pre-established guidelines. The discussion of inter-rater reliability in several research reports provides evidence that the use of clearly written guidelines facilitates faculty agreement on the quality of test items. In Ellsworth et al.’s (1990) study, inter-rater agreement for the use of the 12-guideline matrix was evaluated by reviewing a random sample of 60 test items. Agreement on the use of the guidelines occurred in 96% of the possible 720 entries (Ellsworth et al., 1990, p. 290). Hansen and Dexter (1997) achieved similar results with 97% agreement with 17 criteria and a sample of 80 items ($N = 1,360$). The nursing faculty reviewers in Masters et al. (2001) also had 97% agreement on a sample of 15 test items, after evaluating two “practice” examinations with a combined total of 55 test items (p. 27). Downing (2005) reported that three judges

independently classified test items with “few disagreements” about item classification using the Revised Taxonomy (Haladyna et al., 2002).

Peer review process. The above findings also suggest that faculty peers can successfully analyze the quality of test items using pre-established guidelines. This strategy for improving the quality of test items has been suggested by multiple authors and is consistent with the Item Evaluation Phase of the Conceptual Model for Test Development.

Wallach, Crespo, Hotzman, Galbraith, and Swanson (2006) evaluated the outcomes of a medical school quality improvement project in which pre-established guidelines for item writing were implemented along with committee review of all examinations prior to administration. Test items were randomly selected from examinations administered during the year prior to project implementation (2000-2001) ($n = 250$), following the project implementation (2001-2002) ($n = 270$), and during the second year after project implementation (2002-2003) ($n = 250$) (Wallach et al., 2006). Items were randomized, blinded for year, and rated by three item review experts from the National Board of Medical Examiners using the 5-point scale that was used by Jozefowicz et al. (2002). The mean quality score (QAS) for 2000-2001 was 2.51 ± 1.27 ; test items from 2001-2002 that were written according to the established guidelines and reviewed prior to administration received a QAS of 3.16 ± 1.33 ; and test items from 2002-2003 received a QAS of 3.59 ± 1.15 (Wallach et al., 2006, p. 64). These scores showed significant continuous improvement following implementation of the quality process ($p < .0001$).

Malau-Aduli and Zimitat (2012) similarly conducted an analysis of MC test item quality in a medical school in Australia after the education and implementation of a peer review process for test item development. All items ($N = 866$) for all examinees ($N = 989$) for tests administered in 2008 (prior to implementation of the peer review process) and 2009 through 2010 (after the peer review process) were included in the analysis. Item analysis statistics calculated by the university exam scoring software were examined. Overall, tests administered after the peer review process began contained fewer knowledge-level items (65% in 2008; 30% to 31% in 2009 and 2010) and had increased reliability ($\alpha = .61$ to $.75$ in 2008; $\alpha = .72$ to $.81$ in 2009 and 2010), item difficulty ($M = .17$ to $.25$ in 2008; $M = .24$ to $.29$ in 2009-2010), and improved effectiveness of distractors (44% in 2008; 54% to 57% in 2009 and 2010) ($p < .001$) (Malau-Aduli & Zimitat, 2012). The results from this research study demonstrate that sustained improvement in the quality of MCQs can be achieved through the peer review process (Malau-Aduli & Zimitat, 2012).

Wallach et al. (2006) reported that the item-review process facilitated better communication among the faculty and served as an educational process for faculty members that likely contributed to the continued improvement in item quality. Flynn and Reese (1988) and Van Ort and Hazzard (1985) also discussed these impressions in their case study reports of the implementation of pre-established guidelines and peer review of items prior to administration. Research has also demonstrated that the use of pre-established guidelines and a peer review process to improve items benefits students directly.

Bosher and Bowles (2008) studied the effect of linguistic modification on examinees for nonnative speakers of English – also known as English as additional language (EAL). Linguistic modification is a process of simplifying the language of test items without altering key content area vocabulary and constructs (Bosher & Bowles, 2008). Linguistic complexity is one of the dimensions of bias addressed in item-writing guidelines that introduces construct-irrelevant variances. Sixty-seven test items were chosen for modification; guidelines were applied systematically, a peer review process was conducted, and 38 items were selected for analysis (Bosher & Bowles, 2008). The original and modified versions of the items were analyzed for readability and rated by five volunteer EAL students using a 4-point Likert scale (Bosher & Bowles, 2008). Overall, the readability scores improved for the modified versions, and 84% of the modified items were rated as more comprehensible than the original versions by at least three participants (Bosher & Bowles, 2008, p. 170). Qualitative analysis of the participant comments revealed reasons that the modified versions were easier to understand: use of shorter, simpler sentences; information stated directly; use of the question format; highlighting of key words, such as MOST, BEST, and FIRST; and use of more common words (Bosher & Bowles, 2008). Modified questions also required less amount of time for the participants to read and understand (Bosher & Bowles, 2008). Quantitative studies need to be conducted to validate these findings and analyze the effect of linguistic modification on student test scores, but these findings suggest that improving the quality of test items through a process of evaluation and committee review reduces construct-irrelevant difficulty.

Summary

It is evident that there is a need for development of a valid and reliable tool for use by nursing faculty in evaluating bias in MC test items. The existing taxonomies lack evidence of their validity and reliability in identifying item bias. Previous research has been conducted with non-experimental, descriptive methods and inconsistent use of taxonomies. Developing a clear and concise guideline for nursing faculty to use in developing unbiased test items is one strategy that may improve the quality of nursing assessments, thereby improving the quality of the decisions made based on these assessments. This dissertation study contributes to the body of knowledge by establishing the validity and reliability of a tool that can then be used for further research to validate item-writing guidelines, evaluate the impact of item bias on student success, and better prepare nurse educators to design valid and reliable assessments of student learning. This chapter presented a thorough review of the theoretical and empirical literature related to improving the quality of MC test items. The following chapter discusses the methodology for establishing the validity and reliability of the Fairness of Items Tool (FIT).

CHAPTER III

METHODOLOGY

The purpose of this dissertation study was to establish the validity and reliability of the Fairness of Items Tool (FIT) for its use by nursing faculty in the identification of bias in multiple-choice questions (MCQs). This study examined the question: *Is the Fairness of Items Tool (FIT) a valid and reliable tool for identification of bias in multiple-choice examination items by nurse educators?* This chapter discusses the research design for establishing the validity and reliability of the FIT. Details outlined in this chapter include a description of the population, sampling plan, and the procedure for data collection, data analysis, and protection of human subjects for each phase of the research study.

Research Design

Development and validation of the FIT (Appendix A) was a three-phase process. In the first phase, the tool was developed by the primary investigator through review of published higher education and nursing literature related to item-writing rules, examination bias, and cultural bias. This dissertation study comprised phases two and three, using systematic methods to establish the validity and reliability of the FIT. In phase two, content validity and face validity was established through review of the tool by a panel of item-writing experts. In phase three, reliability and construct validity was

established through testing of the tool by nursing faculty to evaluate sample multiple-choice (MC) test items.

Data collection was conducted electronically. The panel of experts was contacted by email and completed a web-based survey. Nursing faculty participants were contacted by email and completed an anonymous web-based survey in which they used the FIT to evaluate sample MC test items. Demographic information was collected from all participants. Threats to the participants were minimal and related to the time involved in completing the surveys. Every effort was employed in the survey design to minimize respondent burden.

Phase One – Development of the Fairness of Items Tool (FIT)

The FIT (Appendix A) was developed by the primary investigator through an exhaustive search of the literature from 1978-2011, which resulted in a collection of 69 references containing guidelines for developing MCQs. This collection included 18 journal articles from education and assessment literature, including the educational specialties of accounting, law, marketing, medicine, and teacher education; 16 journal articles from nursing education; 10 books with nursing and medical education and educational assessment foci; and 25 research studies that developed or used item-writing guidelines, 10 of which were nursing discipline-specific. The primary investigator also drew upon 15 years of experience in nursing education that included continuing education in the art and science of item writing, practice in developing MC examinations, and professional publication and presentations on the subject of test development.

As the literature was reviewed, several categories emerged, and item-writing rules were sorted according to these categories: testwise flaws; irrelevant difficulty, general, in

the stem, and in the options; linguistic/structural bias, composed of linguistic complexity, grammatical errors, lack of clarity or consistency in the wording, and formatting; cultural bias; and other. Within each category, rules were grouped according to similarity, and the source was included with each for later reference. Five broad dimensions emerged: bias in the stem, bias in the options, linguistic bias, structural bias, and cultural bias.

Guidelines were selected within each dimension for their representativeness of the construct and the consensus of the empirical and theoretical literature, as well as their applicability to nursing education.

The FIT is intended to serve a nursing discipline-specific taxonomy for use by educators in evaluating and revising MC test items. Quality MCQs in nursing must measure the ability to use multilogical thinking and apply nursing concepts to “clinically-oriented situations” (Morrison & Free, 2001, p. 16). Three guidelines (9, 10, and 25) from the Conceptual Model for Test Development were incorporated into the tool to directly address the need for valid, reliable, unbiased MC test items for the discipline of nursing. The resulting Fairness of Items Tool (FIT) (Appendix A) contained 41 item-writing guidelines according to five dimensions of bias: the stem (14 guidelines), the options (12 guidelines), linguistic bias (4 guidelines), structural bias (4 guidelines), and cultural bias (8 guidelines).

Phase Two – Validating the FIT through Expert Review

Phase two of the development of the FIT comprised the first step in its validation. In this phase, the tool was evaluated by a panel of experts in item construction and analysis. Prior to the expert review, the literature from 2011-2014 was evaluated to determine if recent research necessitated revision of the FIT. Results of this evaluation

were incorporated into the literature review in the previous chapter; all of the published literature confirmed previous research findings, and no revisions were made in the FIT.

Sampling

Purposive sampling was used to select six experts who met the inclusion criteria of nursing faculty with expertise in item construction and analysis as evidenced by publication related to item-writing guidelines. These experts were identified through review of the literature on item construction in nursing education. The experts were contacted by email, and five of the six experts responded agreeing to participate in the research study.

Instrumentation

The FIT (Appendix A) contains 41 item-writing guidelines identified through an extensive review of published higher education and nursing literature. The guidelines are categorized into five dimensions: bias in the stem (14 guidelines), bias in the options (12 guidelines), linguistic bias (four guidelines), structural bias (four guidelines), and cultural bias (eight guidelines).

A web-based survey (Appendix E) was designed using Research Electronic Database Capture (REDCap), a secure, web-based survey tool and database, available through the primary investigator's employer and supported by Center for Clinical and Translational Science and Training grant UL1-RR026314. The survey was constructed to collect data from the expert panel using a 4-point Likert scale to evaluate the relevance of each guideline along a continuum as follows: 1 = *not relevant*, 2 = *somewhat relevant*, 3 = *quite relevant*, 4 = *highly relevant*. The same scale was used to evaluate the tool's

organization, ease of use, and overall validity. The survey also contained write-in space for indicating additional items for inclusion in the tool and for general comments.

Data Collection

Each member of the expert panel received a personal code by email that enabled them to access the web-based study materials. Each expert evaluated the tool and individual guidelines within the tool. Expert feedback was incorporated into revision of the tool, and the members of the expert panel were then invited to evaluate the revised tool. Four of the five experts participated in the evaluation of the revised tool using the REDCap survey (Appendix F) and Likert scale to evaluate the relevance of each guideline, the tool's organization, ease of use, and overall validity. The survey again contained write-in space for indicating additional items for inclusion in the tool and for general comments.

Data Analysis

Data analysis for phase two was concerned with addressing the first research hypothesis. *Hypothesis 1: The Fairness of Items Tool (FIT) is a valid tool for identifying bias in multiple-choice examination items.*

A valid tool measures what it is supposed to measure – “the attributes of the construct under study” (DeVon et al., 2007, p. 155). Face validity concerns whether the tool “looks reasonable”, i.e. the items included in the tool are relevant (Bannigan & Watson, 2009, p. 3240). Content validity concerns whether the tool completely represents the attributes of the construct, including all relevant items and excluding irrelevant items (Bannigan & Watson, 2009). Construct validity is concerned with whether the tool provides a means of operationalizing abstract variables, i.e. what the tool is “*really*

measuring” (Polit & Beck, 2012, p. 339). Construct validity is a higher level of validity evidence, as it provides an objective assessment of a tool, whereas face and content validity are subjective judgments. The methods for establishing construct validity will be addressed in the discussion of data analysis for phase three.

Content validity was established through review of the FIT by the panel of experts. The responses to the Likert-scale items were reviewed in a table, and an item content validity index (I-CVI) was computed for each guideline by calculating the number of experts assigning a rating of 3 or 4 on the 4-point scale, divided by the total number of experts (Appendix G). I-CVIs should be .78 or higher to minimize the risk of chance agreement (Polit & Beck, 2012); therefore, any guideline with an I-CVI less than .78 was selected for validation through further literature review. The open responses were analyzed by sorting into themes and evaluating the frequency of similar responses (Appendix H). Themes noted by three or more experts were compared with the guidelines with I-CVIs less than .78 and were included in the validation through review of the literature.

All guidelines selected for further validation were recorded in a decision rubric (Appendix I). The rubric was designed to incorporate the frequency with which each guideline appears in the literature and its empirical support, noting the intent of the guideline and incorporating open responses from the expert panel. The FIT was then revised according to the decision rubric. The expert panel was invited to evaluate the revised tool (FITr), and four of the five experts participated in the second evaluation. I-CVIs were calculated for each guideline on the FITr. Face validity was established in a similar manner by analyzing the responses of the expert panel to the survey questions

about the appearance of the tool. The scale content validity index (S-CVI) was then computed by averaging the I-CVIs. An S-CVI of .90 or higher is desirable (Polit & Beck, 2012). When the expert panel contains less than six experts, additional measures of content validity are recommended (Lynn, 1986). The mean proportion of agreement, or average congruency percentage (ACP) was calculated by averaging the proportion of agreement for each expert. An ACP of .90 is considered acceptable (Waltz, Strickland, & Lenz, 2010). A stronger measure for a small expert panel is the universal calculation method for the S-CVI (S-CVI/UA) (Polit & Beck, 2012). The S-CVI/UA is a measure of the proportion of universal agreement by the experts, and a level of at least .90 is desired (Polit & Beck, 2012). Following analysis of the second review by the panel of experts, the research study proceeded to phase three in which the FITr was validated through use by nursing faculty. The details of the data analysis are presented in the next chapter.

Phase Three – Validating the FITr with Nursing Faculty

Phase three of this research study involved use of the FITr by nursing faculty to evaluate sample MC test items. This phase was concerned with establishing the reliability and construct validity of the FITr.

Population and Sampling

The FIT and FITr were designed for faculty use in identification and correction of MC items. The target population is nursing faculty members who use MC examinations for assessment of student learning. The accessible population is nursing faculty who are employed in American Academy of Colleges of Nursing (AACN) member schools. A list of participants was developed by accessing the websites of nursing programs from among the 760 AACN member schools. In order to ensure representation from all regions in the

United States, member programs were randomly ordered in each state, and faculty names and emails were obtained from the program websites. If a program's website did not contain faculty names and emails, the next nursing program on the list was used. A minimum of 100 names were collected using at least the first three randomly sorted nursing programs in each state. Only publicly available information was collected from program websites. The final list contained 5,786 potential participants from 195 different programs of nursing.

Inclusion criteria included active teaching in a nursing program and utilization of faculty-generated MC examinations for student assessment. Faculty-generated MC examinations include those that are developed by faculty through writing new test items, using test bank items, revising test items from any source, or any combination of these activities. Nursing faculty who were not actively teaching in nursing or who use only standardized MC examinations purchased through a testing service for student assessment were excluded from participation. Because item-writing guidelines are consistent across nursing programs that use MC examinations, it was not necessary to exclude participants based on the type of program.

The accessible population is relatively homogenous, and a small effect size was anticipated. In order to increase statistical power, a large sample size was needed. The larger the sample, the more representative of the population it is likely to be, and the smaller the sampling error (Froman, 2001; Polit & Beck, 2012). A common rule of thumb for scale development is to have 10 participants for every item contained in the scale ($N = 380$). Another rule of thumb is that a sample of at least 300 participants is usually

acceptable (Worthington & Whittaker, 2006). Through consultation with a statistician, it was determined that the target sample would be 60 participants per survey ($N = 300$).

Pilot Study

A pilot study with a convenience sample of the target population was conducted to test the survey and inform research procedures. A focus group comprised of five nursing faculty were asked to pre-test the survey and discuss their experiences. Three of the participants were also current doctoral students and two participants were doctoral-prepared. One doctoral-prepared participant was tenured, an experienced researcher, and editor of a national nursing journal. The other doctoral-prepared participant was experienced in the use of REDCap as a data collection tool. The doctoral student participants were also faculty in different undergraduate nursing programs – one in a small private health-system-based college in Ohio, one in a research-intensive state university in Ohio, and one in a regional state university in California. The pilot group completed a survey including demographic data and analysis of 20 MCQs using the FIT. The focus group was conducted online using AdobeConnect videoconferencing software, available through the primary investigator's employer, to allow all participants to virtually connect from remote locations using webcam and microphone.

Modifications in the survey and data collection plan were made following analysis of the results of the pilot study. Questions discussed with the pilot participants included: *What was it like to complete this survey? Were the directions clear? Did you understand the meaning of each guideline? How many MCQs can reasonably be evaluated within a 15-minute timeframe? What are the best/worst times of the year for you to receive an invitation to complete this survey?* Directed discussion with the focus

group assisted in determining how many MC test questions participants could reasonably be expected to analyze within a 20 to 30 minute time frame. The focus group made several suggestions about the directions for the survey and advised that both written and audio/video instructions be made available. Plans for data collection procedures were modified based on pilot group feedback to minimize respondent burden and improve participation rates. The pilot group also made suggestions related to distributing the survey at times most conducive to faculty availability and workloads.

Instrumentation

During the focus group discussions, the participants expressed concern about the length of time and participant fatigue, and they responded very favorably to the idea of dividing the FIT into dimensions with separate surveys addressing each dimension. Through the focus group discussions, it was decided that the study would be conducted with separate surveys addressing each dimension – stem, options, linguistic/structural, and cultural – with one survey designed with the comprehensive tool.

Five web-based surveys were designed using Research Electronic Database Capture (REDCap), a secure survey tool and database available through the primary investigator's employer and supported by Center for Clinical and Translational Science and Training grant UL1-RR026314. The surveys began with screening questions to ensure that participants met the inclusion criteria. Demographic data for each participant were then collected to assist in explaining results and testing assumptions. Examples of demographic data collected include age, gender, level of education, full-time equivalent years of academic teaching experience, clinical specialty, perceived level of teaching expertise, and perceived level of expertise in item writing (Appendix M).

Four surveys were designed to focus on each of the dimensions of bias within the FITr – stem, options, linguistic/structural, and cultural – with one survey designed for the comprehensive FITr. The REDCap surveys contained unmodified general knowledge MCQs selected from foundational nursing course textbook item banks and previously published research studies. The MCQs were purposefully selected to represent each dimension of the FITr and included both biased and unbiased test items. General knowledge MCQs were selected for which nurse educators can reasonably be expected to be knowledgeable, regardless of clinical specialty or teaching expertise. The comprehensive survey is included in Appendix N; the MCQs selected for each of the other surveys – stem, options, linguistic-structural, and cultural – are included in Appendix O.

Each survey contained the sample test items on a separate screen followed by the designated section of the FITr. In the case of the comprehensive survey, the sample test items were repeated for each section of the FITr, enabling each section to be evaluated within a single screen shot. The FITr guidelines were reworded into question format with a check box indicating yes or no responses to each question. In some cases, a yes response indicated that the item violated the guidelines (is biased), and in others, a no response indicated that the item violated the guidelines (is biased) (Appendix O). A response to each question was required before participants would be able to proceed to the next screen. These strategies were implemented as a result of the pilot study to increase clarity for the participants and improve the quality of the responses.

Data Collection

The second data collection step involved the use of the FITr by nursing faculty to evaluate MCQs. Participant recruitment was completed by the primary investigator using Mail Chimp, a password protected service available through the primary investigator's employer. Mail Chimp provides a platform for creating mass emails and tracking responses. The primary investigator obtained a free private subscription through the duration of participant recruitment and was the only person with account access.

Mail Chimp was used to design an email announcement that contained a brief introduction explaining the purpose of the study and the primary investigator's contact information (Appendix K). Interested participants were asked to submit a form containing contact information, verifying the inclusion criteria, and confirming their email address (Appendix L). To minimize sampling bias, eligible participant responses were reviewed each morning during the data collection period, randomly ordered, and systematically assigned to complete one of the five web-based surveys. The order of survey assignment was (1) comprehensive, (2) stem, (3) options, (4) linguistic-structural, (5) cultural. In order to ensure equal assignment to the surveys, order assignment was continuous from day-to-day. For example, if the last survey assigned on day x was options, the next day participants were randomly ordered, and the systematic assignment to survey began with linguistic-structural. Eligible participants then received an email invitation with a more detailed introduction, explanation of the criteria for participation, informed consent information, and a link to the web-based survey (Appendix P). Participants indicated their consent to participate by clicking the link to proceed with the survey.

Including a deadline date for completion of the survey and sending follow-up emails are strategies that have been recommended to increase response rates (Van Selm & Jankowski, 2006). Because previous research has demonstrated that the majority of responses to email invitations are received within four days from the time of mailing (Van Selm & Jankowski), follow-up emails were sent to non-responders one week and two weeks after the initial announcement. A deadline date of three to four weeks from the first email invitation was specified in the final contact. Follow-up emails were also sent at one-week intervals to participants who indicated an interest in participating and had not completed the survey or who had partially completed surveys. In order to provide an incentive to increase response rates, faculty participants were promised a copy of the FITr for their personal use following the completion of the research study.

Data Analysis

Data analysis in phase three will be addressed according to each research hypothesis. *Hypothesis 1: The Fairness of Items Tool (FIT) is a valid tool for identifying bias in multiple-choice examination items.*

A valid tool measures what it is supposed to measure – “the attributes of the construct under study” (DeVon et al., 2007, p. 155). Construct validity is concerned with whether the tool provides a means of operationalizing abstract variables, i.e. what the tool is “*really* measuring” (Polit & Beck, 2012, p. 339). Construct validity is a higher level of validity evidence, as it provides an objective assessment of a tool, whereas face and content validity are subjective judgments.

Construct validity was established by using the known groups comparison technique through the selection of the sample test items included in the REDCap survey.

A sample of MCQs that are known to be biased was purposively selected through review of previously published research studies. Similarly, a sample of MCQs that are known to be fair (unbiased) were also purposefully selected. Participants indicated the guidelines for which each test item was in violation, and the survey was designed to calculate descriptive statistics for each test item and dimension of item bias, as shown in Table 1.

Table 1

Descriptive Statistics for Data Analysis

Descriptive statistics for each test item

(T_P) = number of participants evaluating each test item

(T_S) = total score assigned to each test item by each participant

(T_G) = total number of times each guideline is selected as violated for each test item

(T_N) = total number of times each guideline is selected as not violated for each test item

(T_{KB}) = total score assigned to each test item known to be biased (unfair)

(T_{KF}) = total score assigned to each test item known to be fair (unbiased)

Descriptive statistics for each dimension of item bias

(B_{STEM}) = total number of guidelines selected for bias in the stem for each test item

$(B_{OPTIONS})$ = total number of guidelines selected for bias in the options for each test item

(B_{L-S}) = total number of guidelines selected for linguistic-structural bias for each test item

(B_C) = total number of guidelines selected for cultural bias for each test item

---

The scores for those questions that are known to be biased were contrasted with the scores for those questions that are known to be fair. Evident differences in these scores provided support for the construct validity of the FITr. A one-tailed Welch's *t*-test (independent samples assuming unequal variances) was calculated using the means for the pairs of scores (known and unknown) to determine the significance level of the differences ($p < .05$). A Welch's *t*-test is appropriate for testing the differences in means from unequal samples in which the variances cannot be assumed to be equal (Miles & Banyard, 2007).

Hypothesis 2: The Fairness of Items Tool (FIT) is a reliable tool for identifying bias in multiple-choice examination items.

A reliable tool consistently and dependably measures what it is supposed to be measuring (Polit & Beck, 2012). Reliability is assessed through multiple means to document the degree of stability, consistency, and equivalence of the tool. Stability concerns the extent to which the tool produces "similar results on separate occasions" (Polit & Beck, 2012, p. 331) and was assessed by calculating split-half reliability. Internal consistency is used to assess how well different items on the tool measure the same attributes of the construct (Bannigan & Watson, 2009) and was evaluated by calculating a Cronbach's alpha correlation coefficient for each MC test item contained on the comprehensive survey. Equivalence measures the degree to which different users of the tool obtain the same results (Polit & Beck, 2012) and was evaluated by testing the independence of scores and by calculating inter-rater agreement.

Reliability of the FITr was established through participant use of the tool to evaluate a sample of general knowledge nursing MC test items in the REDCap survey.

Participants indicated the guidelines for which each test item was in violation, and the survey was designed to calculate descriptive statistics for each test item and dimension of item bias, as shown in Table 1. The distribution of the guideline scores was not normal, necessitating the use of nonparametric tests (Mann Whitney U and Kruskal-Wallis tests) to determine if the distribution of the dimension (B_{STEM} , $B_{OPTIONS}$, B_{L-S} , and B_C) and total scores for each MC test item (T_S) in each survey was consistent among demographic variables. Nonparametric tests are appropriate for non-normal distributions (Polit & Beck, 2012) and are a more sensitive measure in these cases – when a nonparametric score is significant ($p < .05$), the equivalent parametric test will also be significant (J. Ying, personal communication, October 15, 2014). Using a parametric test for a non-normal distribution may lead to falsely elevated levels of significance, which increases the risk of a type II error where the null hypothesis is accepted when, in fact, differences actually exist (Polit & Beck, 2012; Qualls, Pallin, & Schuur, 2010). An equivalent tool measures the variable of interest consistently across demographic groups; therefore, no significant differences were predicted ($p < .05$).

Inter-rater reliability was evaluated by calculating the percent agreement between the participants for each guideline on each test item. Raw agreement indices are appropriate statistical tests for categorical dichotomous data and provide a means to summarize data in a meaningful and practical manner (Uerbersax, 2009).

Internal consistency reliability was established using the total scores for each dimension of bias (B_{STEM} , $B_{OPTIONS}$, B_{L-S} , and B_C) to calculate a Cronbach's alpha correlation coefficient for each sample test item in the comprehensive survey. Cronbach's alpha is an appropriate statistical test for expressing the degree and direction of

relationship between categorical variables (J. C. Schafer, personal communication, November 30, 2009).

Split-half reliability compares the means obtained from two different halves of the tool to estimate the stability of the overall tool and was established using the Kuder-Richardson (KR-20) reliability coefficient. The participants' yes/no responses for the comprehensive items (B-1, B-11, B-13, B-18, B-35, and F-10) were graded against the score that was pre-assigned for each guideline based on the identification of bias during survey development. The KR-20 yields the means of all possible split-halves within the tool. Split-half reliability is not affected by time-error variance and is therefore preferred over test-retest reliability (Fishman & Galguera, 2003). Split-half reliability also has the advantage of not requiring multiple administrations, which reduces the likelihood of participant attrition, length of time for data collection, and participant fatigue.

Protection of Human Subjects

A dissertation proposal hearing with the dissertation committee was held and approval of research methodology received in April 2012 prior to the commencement of any research activities. In addition, this dissertation study received approval as exempt through the Institutional Review Boards (IRB) at the University of Northern Colorado and the primary investigator's employing university following the pilot study. Initial approval was received in December 2013. An amendment was filed in May 2014 to report the modifications made to the surveys, recruitment process, and study design following Phase 2 (expert review). Data collection did not occur until after IRB approval of the amended proposal was granted. Copies of all IRB letters of approval are included in Appendix Q.

Risks to the participants were minimal and related to the time involved in completing the surveys and any difficulty they encountered with technical issues. Significant technical issues did occur on two occasions during the first week of data collection. On the first occasion, the REDCap site was offline for maintenance; however, there was no announcement of this down-time, and participants received an error message when attempting to access the survey during this time. An email communication was sent by the primary researcher to all registered participants to notify them of the downtime. A coding error during the maintenance window caused a software error that affected access to all surveys the next day. Participants could access page one of the survey, but the formatting of the surveys was not included, and participants were not able to respond to demographic questions. These issues were resolved within twelve hours. Two days after the first incidences, a firewall issue at the host site prevented access to the REDCap surveys for participants. Emails apologizing for and communicating about the issue were again sent to registered participants. The second issue was resolved within eight hours.

Every effort was employed in the survey design to minimize respondent burden. Participants were provided with a detailed introductory letter with explanation of the criteria for participation at the initial entry-point, followed by informed consent information. Participants indicated their consent to participate by proceeding with the survey after the introductory page. Participants were provided detailed instructions for completing the survey in both written and video format at the suggestion of the pilot focus group. YouTube was used to create a 15-minute web-based video with explanations of the FITr guidelines. A YouTube video was also created to demonstrate analysis of a

sample test item for each survey. These videos varied in length from 10 to 20 minutes. Participants were required to answer every question before proceeding to the next section; however, the survey was designed with the option of allowing participants to save their responses and return at a later time to complete the survey. This option was added immediately after the first survey downtime issue that occurred so that participant data would not be lost and to minimize respondent burden.

All expert and participant responses were coded and kept separate from identifying information. Research data were maintained in electronic format and password protected on the primary investigator's computer. Emails to multiple participants employed the batch mail merge option to protect the privacy of email addresses. All account data were erased at the end of participant recruitment. All email addresses were deleted from the email account at the end of participant recruitment. All data were reported in aggregate form, and no identifying information will be included in any written report of the research.

Summary

This chapter has explained the methods used for establishing the validity and reliability of the FIT. Development and validation of the FIT was a three-phase process. In the first phase, the tool was developed by the primary investigator through review of published higher education and nursing literature related to item-writing rules, examination bias, and cultural bias. This dissertation study comprised phases two and three, using systematic methods to establish the validity and reliability of the FIT. In phase two, content validity and face validity was established through review of the tool by a panel of item-writing experts. In phase three, reliability and construct validity was

established through testing of the tool by nursing faculty to evaluate sample MC test items. Every effort was made in the study design to minimize respondent burden and prevent bias. A discussion of the sample population demographics and findings from each of the statistical tests is presented in the next chapter.

CHAPTER IV

ANALYSIS OF RESULTS

The purpose of this dissertation study was to establish the validity and reliability of the Fairness of Items Tool (FIT) for its use by nursing faculty in the identification of bias in multiple-choice questions (MCQs). The FIT is intended to serve as a nursing discipline-specific taxonomy for use by educators in evaluating and revising multiple choice (MC) test items. This study examined the question: *Is the Fairness of Items Tool (FIT) a valid and reliable tool for identification of bias in multiple-choice examination items by nurse educators?*

Development and validation of the FIT was a three-phase process. In the first phase, the tool was developed by the primary investigator through review of published higher education and nursing literature related to item-writing rules, examination bias, and cultural bias. This dissertation study comprised phases two and three, using systematic methods to establish the validity and reliability of the FIT. In phase two, content validity and face validity was established through review of the tool by a panel of item-writing experts. In phase three, reliability and construct validity was established through testing of the tool by nursing faculty to evaluate sample MC test items. Statistical analyses for this study were completed using Microsoft Excel 2010 software and IBM SPSS Statistics software Version 21. This chapter presents a comprehensive summary of the data collected, analysis of demographic variables, and discussion of the statistical

evaluation that was performed. Following presentation of the results, the study question and hypotheses are evaluated relative to the statistical analysis.

Phase Two – Validating the FIT through Expert Review

Phase two of the development of the FIT comprised the first step in establishing validity and reliability. In this phase, the tool was evaluated by a panel of experts in item construction and analysis. Data analysis for phase two was concerned with addressing the first research hypothesis. *Hypothesis 1: The Fairness of Items Tool (FIT) is a valid tool for identifying bias in multiple-choice examination items.*

Results

Purposive sampling was used to select five experts who met the inclusion criteria of nursing faculty with expertise in item construction and analysis as evidenced by publication related to item-writing guidelines. Each expert evaluated the FIT and its guidelines by completing a web-based survey using a 4-point Likert scale to evaluate the relevance of each guideline and the tool's organization, ease of use, and completeness. The survey also contained write-in space for indicating additional items for inclusion in the tool and for general comments.

The responses to the Likert-scale items were reviewed in a table, and an item content validity index (I-CVI) was computed for each guideline (Appendix G). Twenty-eight guidelines had I-CVIs of 1.0, indicating perfect agreement by the experts that the guidelines are highly relevant. Eleven guidelines had an I-CVI of .8, which is an acceptable level, although perfect agreement is preferred with a sample of five experts (Polit & Beck, 2012). Four guidelines had an I-CVI less than .78 and were therefore selected for validation through further literature review. The scale item content validity

index (S-CVI) was also calculated by averaging the I-CVIs for each guideline in the FIT (see Table 2). The S-CVI is considered acceptable above .80 (Polit & Beck, 2012), and the results of this study meet that requirement; however, the number of I-CVIs below .78 in this review may not be reflected accurately in the S-CVI since the expert panel contained only five experts. The mean proportion of agreement, or average congruency percentage (ACP), and the universal calculation method for the S-CVI (S-CVI/UA) were also calculated to provide a better measure of expert agreement in a panel of less than six experts (Lynn, 1986). While the ACP and S-CVI were at or above acceptable levels for the first panel review, the S-CVI/UA does not meet acceptable criteria.

Table 2

Validity from Expert Panel: Items Rated 3 or 4 on a 4-point Relevance Scale

Validity Index	Review 1	Review 2
S-CVI	.90	.988
S-CVI/UA	.63	.97
Face Validity	1.0	.92
Proportion Relevant	Expert 1 = .93 Expert 2 = .98 Expert 3 = .73 Expert 4 = .90 Expert 5 = .98	Expert 1 = 1.0 Expert 2 = 1.0 Expert 3 = .97 Expert 4 = .97
ACP	.90	.99

Note. S-CVI = Scale item content validity index; S-CVI/UA = Universal calculation method for the scale item content validity index; ACP = Average congruency percentage.

Decision Rubric. The open responses were analyzed by sorting into themes and evaluating the frequency of similar responses (Appendix H). Themes noted by three or

more experts were compared with the guidelines with I-CVIs less than .78 and were included in the validation through review of the literature. All guidelines selected for further validation were recorded in a decision rubric (Appendix I). The rubric was designed to incorporate the frequency with which each guideline appears in the literature and its empirical support, noting the intent of the guideline and incorporating open responses from the expert panel. Each guideline and theme selected for validation was reviewed in light of the intent of the guideline, literature support for the guideline as originally incorporated in the FIT, and literature supporting the intent of the guideline. The tool was then revised according to the decision rubric. A version of the FIT with track changes that incorporates rationale for each revision is presented in Appendix R.

Revisions. Stem guideline 2: *Eliminate of the following* (I-CVI = .6) had strong support from the literature and empirical research to support the intent of eliminating extraneous words and unnecessary information. This guideline was reworded to reflect the intent with the phrase, *of the following*, included as an example of extraneous wording that should be removed: *Eliminate extraneous words (e.g., of the following)*. The result is a guideline that is more representative of the literature and more broadly applicable.

Stem guideline 6: *Best answer format – underline, capitalize, and bold key words* (**BEST, MOST**) (I-CVI = .4) had weak literature support from one main author with application specifically to non-native speakers of English – also known as English as additional language (EAL). Comments from the expert panel indicated that this practice was inconsistent with standardized examinations and national licensure examinations and

suggested using this strategy for negatively phrased terms only, which is already addressed in stem guideline 5. This guideline was therefore removed from the FIT.

Stem guideline 8: *Avoid conditional expressions (should/would) and passive voice* (I-CVI = .6) also had weak literature support from one author with application specifically to EAL students. The intent of this guideline is to address verb tense, which had some support, including a review of literature that did not specifically apply to EAL students. Comments from the expert panel indicated that should is a desirable term for nursing MC test items, although there was only reference to this position in a textbook authored by one of the expert panel participants. This guideline was revised to reflect the intent of addressing verb tense: *Use active verbs and present tense*.

Structural guideline 31: *Write items that can be read and comprehended easily on the first reading* (I-CVI = .6) similarly had weak literature support from one author and application specifically to EAL students. The intent of this guideline is that test items are understandable, comprehensible, and clear. While students still need to read test items carefully, they should not be worded in a way that requires multiple readings to be understandable, nor should they be trick items. There is adequate literature to support the intent of this guideline with one review of literature and pilot data. This item was revised and combined with stem guideline 6: *Avoid trick items* to better reflect the intent and literature support: *Write items that can be comprehended on the first reading. Avoid tricky or misleading items*.

Themes. Several themes were evident in the written feedback from the expert panel. Cultural guideline 40: *Use gender neutral language* was included in the validation because it received three similar comments during the expert review, despite having had

an acceptable I-CVI (.8). Comments from the experts addressed the fact that it is sometimes necessary to use gender-specific language, but references to gender should not be made when unrelated to the content. There was sufficient support from the literature to retain this guideline, so it was reworded to reflect the expert comments: *Use gender-specific language only when necessary to test nursing content.*

Another theme noted by three experts recommended that more examples be provided to clarify the guidelines. Examples were incorporated into stem guideline 10: *Write questions that require multi-logical thinking (require knowledge of more than one fact/concept)* and stem guideline 12: *Avoid testing student opinions (e.g., use nurse instead of you as the subject)*. The example listed in one guideline was changed based on expert feedback. In cultural guideline 37: *Use terminology from textbook, notes, and common words*, the example was revised to *home* vs. *abode* to present a clearer example of common versus uncommon terminology.

Three options. Two comments from the same expert suggested that the use of three options be included in the tool, and this suggestion has strong empirical support in the literature (Considine, Bottie, & Thomas, 2005; Delgado & Prieto, 1998; Haladyna, Downing, & Rodriguez, 2002; Moreno, Martínez, & Muñiz, 2004; Moreno, Martínez, & Muñiz, 2006; Rodriguez, 2005; Sidick, Barrett, & Doverspike, 1994; Tarrant & Ware, 2010; Weaver, 1982). In fact, the optimal number of options to use in MC test items has been addressed more in the literature than any other item-writing guideline (Haladyna et al., 2002). The use of three options was not incorporated into the original FIT, because four-option MC test items are used on the National Council Licensure Examination (NCLEX) and are therefore standard in nursing education (Oermann, Saewart, Charasika,

& Yarbrough, 2009; Tarrant & Ware). In light of this evidence, the literature was again reviewed to explore the use of three-option MCQs in nursing and higher education.

Haladyna and Downing's (1989a) taxonomy contained this advice: "Use as many functional distracters as possible" (p. 40-41), which was revised after review of literature to "Use as many functional distracters as feasible" (Haladyna & Downing, 1989b).

Rodriguez (2005) conducted a meta-analysis of 27 studies with 56 independent trials published between 1925 and 1999 and concluded that "three options are optimal for MC items in most settings" (p. 10). Three-option MCQs take less time to construct, are easier to write, reduce the probability of including weak distractors, and are as reliable as four-option MCQs (Considine et al., 2005; McDonald, 2014; Rodriguez, 2005; Rogausch, Hofer, & Krebs, 2010; Schneid, Armour, Park, Yudkowsky, & Bordage, 2014; Sidick et al., 1994; Tarrant & Ware, 2010). Research has demonstrated that removing distractors that perform poorly on item analysis improves the discrimination of the test item (Considine et al., 2005; Rodriguez, 2005; Weaver, 1982).

Several studies have demonstrated similar results in nursing education. Tarrant and Ware (2010) used an experimental design to test the use of three versus four options in nursing examinations. While the study has limited generalizability, the findings were consistent with those previously reported in the literature. Tarrant and Ware (2010) recommend the adoption of three-option items in nursing education for reasons of practicality – they are easier to write, take less time to develop and administer, and "perform equally as well" as four-option items (p. 542). Piasentin (2010) investigated the effect of reducing the number of options in a "high-stakes credentialing examination" by examining item analysis data post-administration and eliminating the weakest distractor

(p. 20). Statistical analysis demonstrated that there would be no significant impact on item difficulty, discrimination, or test reliability. Faculty with experience in item writing reported that developing the third distractor took the most time and was perceived as the most difficult part of the process of developing quality test items. As a result of these findings, Piasentin also advocated for three-option test items as being more efficient to develop while providing at least equal quality testing. Redmond, Hartigan-Rogers, and Cobbett (2012) administered examinations to two cohorts of nursing students in Nova Scotia to compare three- and four-option MCQs. Non-functioning distractors were removed from each MCQ by examining the results of the item-analyses from the previous three years. The results demonstrated no significant difference in item difficulty or discrimination between the groups, and mean examination averages also did not differ (Redmond et al., 2012). These researchers also strongly recommended the implementation of three-option MCQs by nurse educators and licensing bodies (Redmond et al., 2012).

It is acceptable to have test items with different numbers of options on the same test (Haladyna & Downing, 1985; King, 1978; McDonald, 2014), and it is better to use three plausible options, rather than write a fourth option for no other reason than to have uniform test items (McDonald, 2014; S. Morrison, personal communication, October 3, 2008). Using three options is an excellent alternative to all-of-the-above or none-of-the-above as the fourth option and supports the expert panel review. The three-option rule was thus incorporated into the FIT as an alternative in options guideline 16: *Avoid none-of-the-above and all-of-the-above. Use three options instead.*

Other revisions. During this review, it also became apparent that there was overlap between the guidelines within the dimensions of linguistic bias and structural bias, and the distinctions as to which dimension a guideline belonged were often unclear. A logical resolution was to combine these guidelines into a single dimension of linguistic/structural bias. The items were also necessarily reordered during this process.

Stem guideline 4: *Avoid absolute terms (always, never, all)* is not limited to the stem only. This rule should be applied to both the stem and options (Haladyna & Downing, 1985; Hansen & Dexter, 1997; McDonald, 2014). This error is one of irrelevant difficulty and also contributes to linguistic complexity, particularly for EAL students (Bosher, 2003). This guideline was therefore moved to the linguistic/structural bias dimension.

Options guideline 14: *Make sure options are similar in length and amount of detail* and guideline 15: *Make sure options are grammatically and visually similar* contained redundancies. These guidelines were therefore combined and reworded for clarity: *Make sure options are similar grammatically and in length and amount of detail*.

Revised Fairness of Items Tool. The Revised Fairness of Items Tool (FITr) (Appendix J) contains 38 item-writing guidelines categorized into four dimensions: bias in the stem (10 guidelines), bias in the options (11 guidelines), linguistic-structural bias (9 guidelines), and cultural bias (8 guidelines). The expert panel was invited to evaluate the FITr, and four of the five experts were able to participate in the second evaluation. The web-based survey used in the first expert panel review was modified to evaluate the relevance of each guideline, organization, ease of use, and completeness of the FITr.

Results from the second review. Item content validity indices (I-CVIs) were calculated for each guideline on the FITr (Appendix G). Results from the second review indicated improvement with the revisions. Cultural guideline 32: *Eliminate all names* was the only guideline with an I-CVI less than 1.0. The S-CVI and ACP were improved from the first review (.988 and .99 respectively); and the S-CVI/UA was markedly improved (.97), indicating almost perfect agreement by the expert panel (see Table 2). These results provide strong support for the content validity of the FITr.

Face validity was established in a similar manner by analyzing the responses of the expert panel to the three survey questions about the appearance of the tool – organization, ease of use, and completeness. Face validity was 1.0 and .92 for the FIT and FITr respectfully. One expert made comments about the usability of the tool that had not been addressed in the first review; therefore, the face validity of the FITr was lower than that of the FIT. There is no documented standard on which to base decisions about face validity, as it is considered subjective. The survey design for the expert review was intended to help quantify this subjective assessment. It is reasonable to follow the standard of .90 set for the CVI, and both determinations for face validity for this study are therefore acceptable.

Phase Three – Validating the FIT with Nursing Faculty

Following analysis of the second review by the panel of experts, the research study proceeded to phase three, in which reliability and construct validity of the FITr was established through use by nursing faculty to evaluate sample MC test items.

Participation Rates

The sample for this research study was drawn from the accessible population of nursing faculty employed in American Academy of Colleges of Nursing (AACN) member schools. A list of 5,786 names and email addresses were systematically sampled from AACN member school websites. Inclusion criteria included active teaching in a nursing program and utilization of faculty-generated MC examinations for student assessment. Faculty-generated MC examinations include those that are developed by faculty through writing new test items, using test bank items, revising test items from any source, or any combination of these activities. Nursing faculty who were not actively teaching in nursing or who use only standardized MC examinations purchased through a testing service for student assessment were excluded from participation.

During participant recruitment, 704 potential participant names were eliminated because they were duplicate entries, did not meet inclusion criteria, were unavailable during data collection, or for whom email addresses were undeliverable. Of the remaining sample of 5,082, the interest form was submitted by 695 eligible participants (14%), 489 of whom participated in the research study (10%), with 379 completing the survey entirely (7.5%). A participation rate of 10% is consistent with conservative estimates of response rates for email and web-based surveys, which range from 2% to 25% (J. C. Schafer, personal communication, November 30, 2009).

Of those eligible participants who submitted interest forms ($n = 695$), 70% participated in the study, and 55% completed the survey in its entirety. Seventy-eight percent of those who participated ($n = 489$) completed the entire survey ($n = 379$). Incomplete surveys were included in the data analysis for any test items for which all of

the guideline questions were answered, and partial test item responses were excluded. Complete and incomplete response totals were calculated for each survey and are presented in Table 3. The comprehensive survey was the lengthiest survey, with participants being required to analyze six MC test items according to all dimensions of the FITr. The comprehensive survey contained 38 guidelines and demographics, which meant that participants responded to 247 questions. Completion rates for the comprehensive survey were the lowest (63%). The cultural survey was the shortest survey, incorporating eight guidelines and 10 MC test items, and requiring participants to respond to 99 questions. Completion rates for the cultural survey were the highest (89%).

Table 3

Completion Data for FIT Surveys

Survey	Incomplete	Complete	Total	Completion Percentage
Comprehensive (COMP)	38	65	103	63.1%
Stem (S)	16	78	94	83%
Options (O)	27	78	105	74.3%
Linguistic-Structural (L-S)	18	72	90	80%
Cultural (C)	11	86	97	88.7%
Total	110	379	489	77.5%

Characteristics of the Sample

Raw demographic data were combined to enable description of the entire sample and by individual survey and is presented in Appendix S. Data were coded and categorized to facilitate comparison with the data that are available describing the demographics of the population of nursing faculty in the United States. Overall, the

demographic characteristics of the sample population were fairly representative of the general nursing faculty population, consisting primarily of educated white females over age 45. The sample population was more likely to have doctoral preparation, full-time and tenured or tenure track status, certification in academic nursing education, and hold higher academic rank than the general nursing faculty population. Males were slightly overrepresented in the sample, while African Americans were underrepresented. The sample represented all regions in the United States, over 162 programs of nursing, and diverse clinical specialties.

Gender and age. Male participants were slightly overrepresented in the sample population (7.4%, $n = 36$) when compared with the 2012 Annual Survey from the American Association of Colleges of Nursing (AACN) (2014a) in which males represented 5.4% of the general nursing faculty population. The participants reported ages ranging from 27 to 87 with a mean age of 53 years, which is consistent with the general nursing faculty population mean age of 56 years for master's and doctoral prepared faculty at the ranks of Assistant Professor, Associate Professor, and Professor in 2012-2013 (American Association of Colleges of Nursing, 2014b). The NLN 2009 Faculty Census revealed similar ages with 57% of part-time and "nearly" 76% of full-time faculty over the age of 45, with 16% of full-time educators over age 60 (National League for Nursing, 2010, p. 1).

Ethnicity and race. The minority representation of the sample population was similar to that of nursing faculty reported in the NLN 2009 Faculty Census with respect to Hispanic, American Indian/Alaska Native, and Asian minority groups (2%, 1.8%, and 0.4% respectively); however, the percentage of African-American participants (4%) was

half of that represented in the general nursing faculty population in 2009 (8%). Overall, racial-ethnic minorities accounted for 6.3% of the sample population, while 12.3% of the general nursing faculty belonged to a racial-ethnic minority in 2012 (American Association of Colleges of Nursing, 2014a).

Highest degree. The sample population represented a much larger proportion of doctoral preparation (64.5%, $n = 315$) than the general nursing faculty population. The NLN (2010) reported for full-time faculty in 2009 that 25% had a doctoral degree, compared with 67% master's-prepared. Part of this discrepancy may be related to the fact that the NLN reported only full-time faculty, while the sample population represents both full- and part-time faculty. The NLN did not specify whether their data represented only earned doctorates or also included in-progress degrees, which may also explain part of the discrepancy. This discrepancy may also be related to the fact that doctoral prepared faculty are more motivated to participate in nursing education research studies, a possibility that was supported by the email communications to the primary investigator offering encouragement and support for completion of the requirements for the doctorate degree.

Experience. The participants had extensive clinical nursing experience with 60% ($n = 293$) reporting 20 or more years and close to 90% ($n = 430$) reporting at least 10 years. The range of FTE years of clinical nursing experience was 0 to 50 years with a mean of 22.7 years (± 11.3). Participants had less academic nursing experience with slightly over half (50.6%, $n = 247$) reporting less than 10 years. The range of FTE years of academic nursing experience was 1 to 40 years with a mean of 11.9 years (± 8.9). A wide variety of clinical specialties were reported, and these were organized into seven

broad categories. The most common specialties reported were medical-surgical (including adult health and oncology), family health (including women's health, obstetrics, midwifery, maternal-child, and pediatrics), and critical care (including emergency, perioperative, and anesthesia).

Status. Almost all of the participants reported full-time employment (95.7%) and faculty status (98.4%) (including adjunct, visiting, tenure track, clinical track, tenured, and non-tenured), with the majority (59.5%) holding appointments outside of the tenure track (including adjunct, visiting, clinical track, and non-tenured). Almost half (47.5%) of the participants held the rank of Assistant Professor with another 30% holding higher ranks. These findings are consistent with the number of years of academic experience within the sample population, but there are some differences from the general nursing faculty population.

The NLN/Carnegie National Survey of Nurse Educators in 2006-2007 reported 90% of respondents holding full-time faculty positions (Kaufman, 2007). The NLN (2010) reported less than one third of nursing faculty holding tenure in 2009 with wide discrepancies between professors and associate professors for whom 75% and 65% are tenured respectively, compared with clinical faculty, 6% to 31% of whom are on the tenure track. The NLN (2010) reported faculty rank by race-ethnicity for 2009. Among the White Non-Hispanic full-time nursing faculty population, 35% were at the Instructor rank, 26% Assistant Professor, 15% Associate Professor, and 12% Professor (National League for Nursing, 2014b). The proportion of tenured and tenure track faculty and higher ranks in the sample population is consistent with the number of years of academic nursing experience and the percentage of faculty with doctoral preparation represented.

Expertise. Slightly less than 20% of the participants ($n = 97$) reported having earned the Certified Nurse Educator (CNE) credential, the majority of which were earned in 2010 or later (70.1%, $n = 68$). There are currently 4,220 CNEs (L. Simmons, personal communication, March 3, 2014), comprising 13.2% of the 32,000 nursing faculty reported in a recent national survey (McNeal, 2012). The percentage of CNEs in the sample population is higher than the general nursing faculty population, which is consistent with the higher levels of academic nursing experience, tenure status, and education also found in this population. Participants were also asked to rate their expertise using a Likert scale from novice to expert. Overall, the participants demonstrated more expertise in teaching than item writing, with 66% ($n = 323$) assigning ratings of at least proficient in teaching compared with 34% ($n = 168$) at this level in item writing. These findings are consistent with published reports that few nursing faculty members have formal preparation and expertise in assessment methods such as item construction (Tarrant, Knierem, Hayes, & Ware, 2006; Tarrant & Ware, 2008, Zungolo, 2008).

Demographics by survey. The sample populations within each survey category were similar across most of the demographic variables (Appendix S). A Z-test statistic was calculated to compare the proportions of demographic variables between survey groups ($p < .05$). The Z-test for proportions is appropriate to test whether large ($n > 30$) independent random samples differ on some categorical characteristic (Stangroom, 2014). Overall, the survey participants were very similar in terms of demographic characteristics. Differences between the samples were significant only with respect to specialty in the options survey, in which participants with community health specialties

were represented in significantly higher proportions (16.3%) than in the other surveys (4.4% to 7.2%). There were also fewer participants with critical care specialties represented in the options survey, although this difference was not statistically significant.

Adequacy of sample. The accessible population is relatively homogenous, and a small effect size was anticipated, so a large sample size was desired in order to increase statistical power. The larger the sample, the more representative of the population it is likely to be, and the smaller the sampling error (Froman, 2001; Polit & Beck, 2012). The common rule of thumb for scale development is to have 10 participants for every item contained in the scale ($N = 380$). Another rule of thumb is that a sample of at least 300 participants is usually acceptable (Worthington & Whittaker, 2006). For this research study, the sample contained 379 completed surveys and an additional 110 incomplete surveys with usable data, which meets the benchmark for a 10:1 ratio of participants per item contained in the scale. Separate surveys were administered for each dimension in the FITr in order to minimize respondent burden and improve completion rates, however, resulting in a sample size for the MC test items that were evaluated comprehensively (according to all dimensions of bias) that ranged from 64 to 163 participants when all data were combined.

Results

The results of the data analysis in phase three will be addressed according to each research hypothesis. *Hypothesis 1: The Fairness of Items Tool (FIT) is a valid tool for identifying bias in multiple-choice examination items.* Descriptive statistics were

calculated for each test item and dimension of item bias and were used for calculating the validity and reliability statistics and are presented in Appendix T.

Construct validity. The known groups comparison technique was used to establish construct validity of the FITr through the selection of the sample test items included in the REDCap survey. Samples of MCQs known to be biased and known to be fair were purposefully selected through review of previously published research studies. Participants indicated the guidelines for which each test item was in violation, and the scores for those questions known to be biased were contrasted with the scores for those questions known to be fair. Pairs of known biased (T_{GB}) and fair (T_{GF}) scores for each guideline were compared to test the hypothesis that the scores for the known fair items would be lower than the scores for the known biased items. Similar pairs were compared for the total scores in each dimension of bias (B_{DIM-B} , B_{DIM-F}) and for the total scores (T_{SB} , T_{SF}) for each test item. At the levels of the dimension and test items, the total scores were standardized by dividing by the number of guidelines to facilitate comparison across dimensions; all comparison scores ranged from 0 to 1. A one-tailed *t*-test for independent samples assuming unequal variances using the means for the pairs of scores was used for testing the following hypotheses ($p < .05$):

Guideline Level	$H_0: \mu T_{GF} = \mu T_{GB}$	$H_1: \mu T_{GF} < \mu T_{GB}$
Dimension Level	$H_0: \mu B_{DIM-F} = \mu B_{DIM-B}$	$H_1: \mu B_{DIM-F} < \mu B_{DIM-B}$
Test Item Level	$H_0: \mu T_{SF} = \mu T_{SB}$	$H_1: \mu T_{SF} < \mu T_{SB}$

It was expected that the scores for the known fair items would be closer to zero, and the scores for the known biased items would be closer to one. Overall, the items performed as expected with few exceptions. The guideline scores ranged from 0 to 25 for the known

biased items ($M = 2.7 \pm 2.4$) and 0 to 10 for the known fair items ($M = 0.92 \pm 1.4$). At the dimension level, the means of the standardized scores for the known biased items ranged from 0 to 5 ($M = .29 \pm .41$) and 0 to 0.73 ($M = .1 \pm .13$) for the known fair items. At the test item level, the mean total score for the known biased items was 7.84 $\pm$ 4.8 compared with 2.76 $\pm$ 2.84 for the known fair items. The comparisons of the means of the scores assigned to each guideline (T_G) total and standardized scores for each dimension of bias (B_{STEM} , $B_{OPTIONS}$, B_{L-S} , B_C) are presented in Appendix U.

Bias in the stem. Seven MC test items containing known bias and one item known to be fair in the stem were purposefully selected for this research study. Known bias was present in the selected MC test items for 7 of the 10 guidelines pertaining to the stem (ES). All of the scores demonstrated higher values for the known biased items than the known fair items. Guideline mean scores for the known biased items ranged from .15 to .95, while mean scores for the known fair items ranged from .007 to .14. For all of the guidelines pertaining to bias in the stem, mean scores for the known fair items were lower than scores for the known biased items ($p < .05$). The means of the standardized scores for known biased items ranged from .1 to .52, while the mean standardized score for the known fair item was .04. For the dimension of bias in the stem, all of the mean scores for known biased items were higher than those of known fair items ($p < .05$).

Bias in the options. Eleven MC test items containing known bias and two items known to be fair were purposefully selected for this research study. Known bias was present in the selected MC test items for 10 of the 11 guidelines pertaining to the options (EO). Scores for 17 of the 20 pairs demonstrated higher values for the known biased items than the known fair items. The range of guideline mean scores for the known

biased items was .125 to .96, while the range of the mean scores for the known fair items was .007 to .14. Guideline EO15: *Avoid repeating words in the stem and correct option* demonstrated mixed results in the known fair items, with one item (F-8, $M = .5$) scoring higher than the biased item and the other (F-10, $M = .34$) scoring lower than the biased item. Additionally, both known fair items scored higher than expected.

The known biased item for guideline EO21: *Write options that require a high level of discrimination to select the correct answer* (B-36) scored much lower than expected (.125) and lower than both of the known fair items. The range of means of the standardized scores for known biased items was .8 to .38, compared with mean scores of .08 and .27 for the known fair items. For the dimension of bias in the options, all of the mean scores for known biased items were higher than those of known fair items ($p < .05$).

Linguistic-structural bias. Seven MC test items containing known linguistic-structural bias and two items known to be fair were purposefully selected for this research study. Known bias was present in the selected MC test items for seven of the nine guidelines pertaining to linguistic-structural bias (LS). Scores for seven of the eight pairs demonstrated higher values for the known biased items than the known fair items; however, the difference in mean scores for the pair pertaining to guideline LS26: *Use straight-forward, uncomplicated language. Test nursing content, not vocabulary or reading* were not significant ($p = .16$).

The range of guideline mean scores for the known biased items was .9 to .75, while the range of the mean scores for the known fair items was 0 to .45. Guideline LS23: *Use correct grammar, punctuation, capitalization, and spelling* demonstrated mixed results in the known biased items, with one item (B-18, $M = .33$) scoring higher

than the fair item and the other biased item (B-10, $M = .14$) scoring lower than the fair item; item B-10 scored lower than expected for a known biased item. The range of means of the standardized scores for known biased items was .145 to .32 compared with a mean score of .144 for the known fair item. For the dimension of linguistic-structural bias, all of the mean scores for known biased items were higher than those of known fair items; however, the difference in means between one pairing (B-20 and F-10) was not significant ($p = .488$).

Cultural bias. Four MC test items containing known cultural bias and two items known to be fair were purposefully selected for this research study. Known bias was present in the selected MC test items for five of the eight guidelines pertaining to cultural bias (C). All of the mean scores for the guidelines pertaining to cultural bias demonstrated higher values for the known biased items than the known fair items ($p < .05$). The range of guideline mean scores for the known biased items was .27 to .97, while the range of the mean scores for the known fair items was 0 to .06. For all of the guidelines pertaining to cultural bias, mean scores for the known fair items were lower than scores for the known biased items ($p < .05$). The range of means of the standardized scores for known biased items was .083 to .52, compared with mean scores of .007 and .019 for the known fair items. For the dimension of cultural bias, all of the mean scores for known biased items were higher than those of known fair items ($p < .05$).

Test item level. Three MC test items containing known bias and two items known to be fair were purposefully selected for this research study and evaluated comprehensively according to all four dimensions of bias. The comparison of the means of the standardized item total scores (T_s) is presented in Table 4. All of the total item

scores for the known fair items were lower than scores for the known biased items ($p < .05$). Overall, the items for the known groups comparison performed as expected with few exceptions. Scores for the known fair items were lower than scores for the known biased items, and mean scores for the fair items were close to zero at all levels of the analysis. Mean scores for biased items at the guideline level were closer to one than those at the dimension and item levels. The results of this analysis support the construct validity of the FITr. Discussion of the conclusions and explanations for the exceptions will be explored in the next chapter.

Table 4

Known Groups Comparison: Difference of Means of Test Item Scores

Biased Item	ST $T_{SB}(\mu)$	Fair Item	ST $T_{SF}(\mu)$	p
B-11	20.3 (.265)	F-10	4.8 (.073)	+
B-13	15.1 (.224)	F-10	4.8 (.073)	+
B-18	16.5 (.246)	F-10	4.8 (.073)	+

+ $p < .05$

Hypothesis 2: The Fairness of Items Tool (FIT) is a reliable tool for identifying bias in multiple-choice examination items. Reliability was assessed through multiple means to document the degree of stability, consistency, and equivalence of the FIT. Stability was assessed by calculating split-half reliability for the comprehensive items (those that were evaluated according to all four dimensions of bias). Internal consistency was evaluated by calculating a Cronbach's alpha correlation coefficient for each MC test item contained on the comprehensive survey. Equivalence was evaluated with

nonparametric measures of the independence of scores and by calculating inter-rater agreement for each guideline on each test item.

Equivalence. Independence of scores and inter-rater reliability were used to test the hypothesis that the FITr produces similar results for different users. Nonparametric tests (Mann Whitney *U* and Kruskal-Wallis) were used to explore the distribution of yes/no scores across demographic variables, because the distribution of the yes/no responses identifying violations of item-writing guidelines were highly skewed, depending on whether the test item was biased or fair for that guideline. If the yes/no responses and demographic variables demonstrate independence, there is strong support for the equivalence reliability of the FITr.

Analysis of independence. Twenty-eight MC test items were evaluated with the FITr in this research study, each selected to represent specific guidelines and dimensions of bias. This study used the scores obtained from five different surveys (comprehensive, stem, options, linguistic-structural, and cultural). The comprehensive survey contained six MC items to evaluate all 38 guidelines, and each of the other surveys contained 10 MC items to evaluate the guidelines within the selected dimension of bias (the number of guidelines in each survey were 10, 11, 9, and 8 respectfully). Seventeen demographic variables were included in this analysis. A total of 1,190 values for the independence of scores were obtained with this analysis: 170 per survey for the dimensions and 510 for the comprehensive survey; 70 per demographic variable; 212 per dimension of bias; and 102 for the T_S from the items in the comprehensive survey. These results are presented in Appendix V. Overall, independence of scores was demonstrated in over 95% of the correlations for this research study ($N = 1,136$) ($p < .05$). These results demonstrate

strong support for the hypothesis that the FITr produces consistent results when used by nursing faculty, regardless of user demographics such as gender, ethnicity, level of education, experience in academic nursing education, and level of expertise.

Analysis of agreement. Inter-rater reliability was evaluated by calculating the agreement between the participants ($N = 513$ items). Raw agreement indices were evaluated according to the following scale, with good to perfect agreements providing support for the tool’s equivalence (see Table 5).

Table 5

Interpretation of Raw Agreement Indices

Perfect agreement if the agreement coefficient is .9 to 1.0.
Excellent agreement if the agreement coefficient is .8 to .89.
Very good agreement if the agreement coefficient is .7 to .79.
Good agreement if the agreement coefficient is .6 to .69.
Fair agreement if the agreement coefficient is .5 to .59.
Poor agreement if the agreement coefficient is below .5.

Note. J. Ying, personal communication, October 7, 2014.

Agreement indices for the items are presented in Appendix W and organized according to the dimension of bias. Overall, raw agreement indices of at least .6 were demonstrated in 90% of the items ($n = 463$) with perfect agreement for almost half (47%). Within each dimension of bias, good to perfect agreements were demonstrated in 88% to 94% of the items. Items within the dimension of cultural bias demonstrated the highest agreement (94%), followed by linguistic-structural bias (92%) and bias in the stem and options (88% in each). Perfect agreements were demonstrated in 32% to 74% of

items, with the highest number of perfect agreements demonstrated in the dimension of cultural bias (83 of 112 items) and the lowest number of perfect agreements in linguistic-structural bias (40 of 126 items).

Only two items demonstrated poor agreement, and both scores were .49, which is just shy of fair agreement. Both of these items represented the dimension of bias in the options. Poor agreement was demonstrated for a known fair item pertaining to guideline EO15: *Avoid repeating words in the stem and correct option* (F-8) and a biased item pertaining to guideline EO21: *Write options that require a high level of discrimination to select the correct answer* (B-1). These two items also demonstrated guideline scores (T_G) that failed to meet the expectations of scores closer to zero for fair items and closer to one for biased items.

Agreements indices were also sorted by guideline to explore the number of items pertaining to each guideline at each level of the scale. Overall, the guidelines contained in the FITr demonstrated strong agreements. Over one third of the guidelines ($n = 13$) demonstrated agreements of at least .6 for 100% of the relevant items. Four guidelines demonstrated agreements of at least .8 for all of the relevant items:

ES1: *Use a question format.*

ES4: *Avoid negatively phrased questions, double negatives, and the use of except.*

EO12: *Avoid none-of-the-above and all-of-the-above. Use three options instead.*

C32: *Eliminate all names.*

Nine additional guidelines demonstrated agreements of at least .6 for all of the relevant items:

ES9: *Avoid testing student opinions (e.g., use nurse instead of you as the subject).*

ES10: *Test important content and avoid trivia.*

EO17: *Eliminate multiple-multiples.*

LS23: *Use correct grammar, punctuation, capitalization, and spelling.*

LS24: *Use precise terms (avoid frequently, appropriate).*

LS25: *Avoid absolute terms (always, never, all).*

LS30: *Use consistent spacing, question numbering/lettering, page numbering.*

Make sure options appear on the same page as the question.

C33: *Eliminate all slang.*

C38: *Present the person first, not the diagnosis.*

Only one guideline failed to demonstrate agreement of good and above for at least 60% of the relevant items. ES3: *Present a single, clearly defined question with the problem in the stem* demonstrated fair agreement with 55% of relevant items scoring .5 to .59. Overall, the results of the raw agreements demonstrate strong inter-rater reliability and provide support for the equivalence of the FITr.

Consistency. Internal consistency reliability was established using the total scores for each dimension of bias (B_{STEM} , B_{OPTIONS} , $B_{\text{L-S}}$, and B_{C}) to calculate a Cronbach's alpha correlation coefficient (α) for each sample test item for which responses for the comprehensive tool were available (see Table 6). For this study, α demonstrated acceptable internal consistency ($\alpha > .60$) for five of the six test items evaluated, and three of the test items had α coefficients greater than .70 ($p < .05$). The known fair test item (F-10) had the lowest correlation coefficient ($\alpha = .598$, $p < .05$). This test item also had high agreement indices and mean guideline scores (T_{G}) near zero.

Table 6

Comparison of Cronbach's Alpha Coefficients

Test Item	α	n	Dimension	α if Item Deleted	$\mu \pm$ SD
B-18	.737	67	B18-S	.601	2.55, 1.743
			B18-O	.585	3.48, 2.149
			B18-LS	.571	2.99, 1.942
			B18-C	.82	0.33, 0.613
B-13	.73	70	B13-S	.697	2.71, 1.589
			B13-O	.567	3.09, 1.924
			B13-LS	.617	1.83, 1.769
			B13-C	.741	0.91, 1.004
B-1	.706	66	B1-S	.648	3.2, 1.511
			B1-O	.566	1.18, 1.201
			B1-LS	.611	1.02, 1.583
			B1-C	.712	0.27, 0.621
B-11	.694	80	B11-S	.519	4.14, 1.682
			B11-O	.603	3.03, 1.713
			B11-LS	.569	2.6, 1.946
			B11-C	.741	0.19, 0.576
B-35	.651	65	B35-S	.595	5.14, 2.106
			B35-O	.504	3.46, 1.985
			B35-LS	.578	1.03, 1.541
			B35-C	.638	2.11, 1.522
F-10	.598	67	F10-S	.513	0.39, 0.797
			F10-O	.416	0.97, 1.18
			F10-LS	.404	1.3, 1.557
			F10-C	.644	0.06, 0.239

Note. B = Biased item; F = Fair item; S = Bias in the Stem; O = Bias in the Options; LS = Linguistic-Structural Bias; C = Cultural Bias.

The dimension of cultural bias demonstrated the weakest correlation for all of the items evaluated in items for which no cultural bias was present ($B_C = 0$). Cultural bias had the strongest relationship with the other dimensions in item B-35, an item that did

contain cultural bias and for which agreement indices were lower. Overall, correlations were higher when bias was present than when bias was not present in a test item. These results suggest that the dimensions represent similar constructs of item bias and provide support for the internal consistency reliability of the FITr.

Stability. Split-half reliability was measured using the Kuder-Richardson (KR-20) reliability coefficient to examine the hypothesis that the FITr will produce similar results on different occasions. The participants' yes/no responses for the comprehensive items (B-1, B-11, B-13, B-18, B-35, and F-10) were graded against the score that was pre-assigned for each guideline based on the identification of bias during survey development. A KR-20 of .799 ($\alpha = .05$) was calculated using the graded responses, which is above the benchmark of .70 for a reliable test. The results support the hypothesis that the FITr will produce similar results on different occasions.

Additional Findings

Significant technical issues occurred on two occasions during the first week of data collection; therefore, the period for data collection was extended to six weeks in order to meet the target sample of 60 participants for each survey ($N = 300$). Patterns of participant enrollment and survey completion were reviewed in order to explore whether the technical issues had any impact. Almost half of the interest forms (46.8%) were submitted on or before the date of the first instance of technical issues; however the majority of participation (77.8%) took place after the technical issues were resolved. Close to half of the incomplete surveys were started before the technical issues (41.2%), and only one participant with an incomplete survey before the technical issues finished

the survey at a later date, compared with 27 who returned to complete surveys that were started after the technical issues.

Complicating this analysis is the fact that participant recruitment took place during the summer, a time when many faculty are off contract and therefore check email infrequently or are inaccessible. The time period for recruitment took place toward the end of the summer when some faculty were busy preparing for fall semester and had less time to participate. Feedback from the pilot study participants indicated that there is no ideal time for faculty; however, a few weeks after the beginning of a term through the middle of the term was identified as a time in which faculty may be more available. It is suspected that both technical issues and the timing of the participant recruitment had a negative impact on participation rates for this study.

Summary of the Findings

This dissertation study used systematic methods to establish the validity and reliability of the Fairness of Items Tool as part of a multi-phase process of tool development. Two hypotheses were proposed, and the study was designed to use multiple measures to address each hypothesis.

Hypothesis 1: The Fairness of Items Tool (FIT) is a valid tool for identification of bias in multiple-choice examinations by nurse educators. Content validity and face validity were established through review of the tool by a panel of item-writing experts. The FIT was revised using systematic methods based on the analysis of the data from the expert panel. The second review demonstrated strong support for content and face validity. Construct validity was established through testing of the FITr (Appendix F) by nursing faculty to evaluate sample MC test items. The known groups comparison

technique was used to compare responses to known biased and known fair items and provided support for the hypothesis that the FITr is a measure of item bias. Analysis of the data provided strong support for the tool's construct validity.

Hypothesis 2: The Fairness of Items Tool (FIT) is a reliable tool for identifying bias in multiple-choice examination items. Reliability was established through testing of the tool by nursing faculty to evaluate sample MC test items. Tests for independence of scores supported the hypothesis that scores obtained using the FITr do not vary according to demographic variables. Analysis of agreement indices supported the hypothesis that different users of the FITr would obtain the same results. These measures supported the tool's equivalence reliability. The KR-20 as a measure of stability supported the hypothesis that repeat use of the FITr to evaluate the same items would have similar results. Cronbach's alpha was calculated to establish the tool's internal consistency reliability. Correlation coefficients demonstrated adequate reliability for a newly developed tool. The results of the analysis supported the hypothesis that the FITr reflects the constructs and dimensions of bias in MC test items. Further development of the FITr will improve its ability to measure the construct of interest.

Overall, the results of this research study support the hypothesis that the FITr is a valid and reliable tool for identifying bias in MC examination items. A more detailed discussion of the findings and implications for nursing education and research is presented in the next chapter.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

The purpose of this dissertation study was to test an intervention to improve the quality of nursing examinations – specifically, to evaluate the Fairness of Items Tool (FIT) and, subsequently, the Revised Fairness of Items Tool (FITr) for its use in the identification of bias in multiple-choice questions (MCQs). This study examined the question: *Is the Fairness of Items Tool (FIT) a valid and reliable tool for identification of bias in multiple-choice examination items by nurse educators?* This chapter presents a discussion and analysis of the findings and limitations of this research study in light of the current literature and theoretical frameworks. The implications for nursing education are discussed, followed by recommendations for future research and conclusions.

Discussion of the Findings

The Fairness of Items Tool (FIT and FITr) was developed to address the need within nursing education for a discipline-specific tool to assist faculty in improving the quality of MC test items. For the FITr to meet this need, it must meet several characteristics:

- Valid – Does the FITr measure what it is supposed to measure – bias in MC test items?

- Reliable – Does the FITr measure bias in MC test items consistently and dependably?
- Practical – Does the FITr provide a clear and concise description of the most relevant item-writing guidelines in an easy-to-use format?

This dissertation research study was designed to address the first two characteristics, establishing the validity and reliability of the FIT through expert review and comparison of faculty scores on MC test items. Overall, the results of this research study support the hypothesis that the FITr is a valid and reliable tool for identifying bias in MC examination items.

This research study also demonstrated that participants made similar decisions when using the FITr to evaluate MC test items. These findings are consistent with previous research reports that provide evidence that the use of clearly written guidelines facilitates faculty agreement on the quality of test items. Previous research evaluating inter-rater reliability reported high agreements among faculty using similar guidelines to evaluate test items. Ellsworth, Dunnell, and Duell (1990) and Hansen and Dexter (1997) reported agreement with two reviewers on 96% to 97% of items respectively, and Downing (2005) reported that three judges independently classified test items with “few disagreements” (p. 135). Six nursing faculty reviewers in Masters et al. (2001) also reported 97% agreement on a sample of test items. This study demonstrated similar high agreements among much larger numbers of reviewers. Good agreements were demonstrated on 90% of the items with 66 to 87 reviewers. These findings suggest that faculty using the FITr to evaluate MC test items will be able to reach similar conclusions about the presence of bias in those test items.

The findings for internal consistency reliability were not as conclusive as the other reliability and validity measures. Internal consistency indicates that the tool represents similar constructs of item bias and was evaluated for this study at the item level to yield information about the correlation of the dimensions to each other and to the total tool for each of the test items evaluated comprehensively. The Cronbach's α was at or above the benchmark of .60 for five of the six items (α for the sixth item = .0598). Further analysis of the results showed that there may be a relationship between the presence of bias within a dimension and the α coefficient. Comparing the α coefficient with the means shows that the closer the mean is to zero, the less likely the dimension or item is to demonstrate a strong α coefficient. Items with means closer to zero indicate bias is less likely to be present, and the scores for these items contain many values of zero that may be affecting statistical analysis. Further research is needed to explore this phenomenon.

Practicality

This research study was not specifically designed to examine the practicality of the FITr; however, there are some inferences that can be made based on the results. For a tool to be effective, it must be used. For it to be used, it must provide a clear and concise description of the most relevant guidelines in an easy-to-use format that facilitates writing and revising fair, valid, and reliable MC test items within a nurse educator's full workload. The results of the evaluation of validity and reliability demonstrate that the FITr provides a clear and concise description of the most relevant guidelines. It is not so clear whether faculty will make time for the FITr in their full workloads. During the review by the expert panel, concerns were expressed about the length of the FITr, and

one expert commented that it was not reasonable to expect faculty to use it to evaluate every test item.

Analysis of the survey completion rates may support this concern. The comprehensive survey was the lengthiest survey and the only one to use the complete FITr to evaluate MCQs; this survey had the lowest completion rate at 64%, compared with 75% to 89% for the other surveys. Comments from the pilot group indicated that the FIT was difficult to use at first but became much easier as they progressed through the questions. Participant instructions for the surveys included the pilot group's advice to "stick with it" through the first few questions. The surveys did not contain space for participant comments, but a few participants emailed the primary investigator stating that the survey was taking "too long" or that they did not have time to complete it. Conversely, the primary researcher also received feedback from participants commenting on how much fun the survey was, how much they learned from the process, and how excited they were to use the FITr to evaluate their own test items. Survey completion rates likely related more to completing a survey during a very busy time of the year than the utility of using the FITr to evaluate MCQs; however, usability of the tool cannot be ruled out as a factor. Future research studies should include space for participant comments to gain insight into the thought processes of faculty as they evaluate MCQs and the time investment involved in the process.

Previous researchers have discussed the time requirement for writing quality test items (Clifton & Schriener, 2010; Morrison & Free, 2001). Authors have also discussed the relative lack of time designated for the item-writing process in faculty workloads; however, only one published research study was found that examined item writing from

the perspective of faculty time commitments. Piasentin (2010) surveyed 75 faculty members who participated in MCQ writing workshops to develop MC test items for a national licensure examination. Participants reported spending, on average, 52 minutes to write one test item with supported rationale. This research was conducted with experienced item writers, so it is likely that item writing is a longer process for the typical faculty member. Further research needs to be designed to evaluate the time that faculty spend on the process of test development, as well as determining the impact that implementation of the FITr has on the time required for test item development.

In summary, the findings of this research study provide evidence of the validity and reliability of the FITr for identifying bias in MC examination items by nursing faculty. Further research is needed to explore the relationship between the presence or absence of bias in a test item and the internal consistency reliability of the FITr. Future research studies also should incorporate space for participant comments to gain insight into the thought processes of faculty as they use the FITr to evaluate MCQs. Finally, research studies need to be designed to investigate the time investment faculty make in the item-writing process before and after implementing the FITr. Additional recommendations for future research are discussed later in this report.

Other Findings

The study findings were reviewed to identify patterns among poorly performing test items, dimensions, and guidelines. Two guidelines were selected for further analysis: EO15: *Avoid repeating words in the stem and correct option*; and EO21: *Write options that require a high level of discrimination to select the correct answer*. Both of these guidelines demonstrated unanticipated results for both inter-rater reliability and the

known groups comparison. Further examination of the test items offers some explanation for these findings.

Guideline EO15 demonstrated mixed results in the known fair items, with item F-8 scoring higher than the biased item, item F-10 scoring lower than the biased item, and both items scoring higher than expected. Item F-8 also demonstrated poor agreement for guideline EO15. The test item F-8 was identified as a fair item by Morrison, Nibert, & Flick (2006):

The nurse notes that a client does not exhibit the defining characteristics of the priority problem identified in the plan of care. What action does the nurse implement?

- A. Document that the client's defining characteristics are inconsistent with the priority problem.
- B. Change the plan of care to include the problem that is consistent with the client's defining characteristics.*
- C. Revise the plan of care so that the identified problem is a high-risk problem rather than a priority problem (p. 36).

The correct response for this item is B. The intent of this guideline is to avoid providing clues to the correct answer that enable testwise students to select the correct response without having the required ability (McDonald, 2014). For this MCQ, every option repeats words from the stem; therefore, the repeated words do not provide a clue to lead students to the correct answer, and guideline EO15 is not violated. There is no way to know what participants were thinking when they responded to this item; however, a significant number either failed to recognize the word repeats in all of the options or did not understand the intent of this guideline.

For guideline EO21, poor agreement (.49) was demonstrated for item B-1. Item B-36 scored much lower than expected (.125) for a known biased item and lower than both of the known fair items for this guideline; however, participant agreement was

excellent for this test item (.81) related to this guideline. An explanation for these discrepancies may found be in the definition of the guideline itself. Discrimination in the options relates to the effectiveness of the distracters and is established through analysis of the frequency of distracter selection and overall item response statistics following test administration. To be discriminating, the distracters must be plausible so that all options are equally appealing to test takers who lack knowledge of the constructs being tested (McDonald, 2014). In this case, both of the test items contained bias for guideline EO21; however, it may have been difficult for participants to evaluate option discrimination for these test items without the post-examination response statistics. Bias in guideline EO21 was present and correctly identified by participants in eight other test items. These results are important to consider when selecting test items for future research studies for the FITr. Consideration may be given to evaluating the FITr as one component of the item development process that also includes post-administration item analysis. This process will be discussed later in this report.

Finally, there were two guidelines from the dimension of linguistic-structural bias that were not evaluated with this research study:

LS29: *Be specific and clear with directions.*

LS30: *Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.*

For this research study, participants were presented with individual test items. These linguistic-structural guidelines relate more to the overall structure of the examination and are difficult to capture in a single MCQ. Future research can address this gap by having participants use the FITr to evaluate a sample test containing a limited number of MCQs.

Theoretical Implications

The findings of this research study are consistent with its foundational theoretical frameworks. According to the Framework for Quality Assessment adapted for this study from Quinn's (2000) Cardinal Criteria of Assessment and Scheuneman's (1984) conceptualization of bias, every effective assessment must be valid, reliable, discriminating, practical, and unbiased. Well-written MCQs are designed to fulfill these criteria as one component of a comprehensive testing program. The process of item writing and evaluation used in this research study is based on The Conceptual Model for Test Development (see Figure 1), which identifies a clear process for constructing high quality test items within the domain of nursing. The identification and revision of biased items is an integral component of the item-writing and evaluation phases of the model.

The results of this research study demonstrate that the FITr can be instrumental in the identification of biased items as part of the process of item development, but this study did not evaluate whether the FITr will assist faculty in writing and revising test items as well. An additional essential component of the process identified by Khan, Danish, Awan, and Anwar (2013) is "repetition and practice" (p. 718). It is speculated that use of the FITr by nursing faculty to identify bias in MC test items will also facilitate improvement of those items through revision, which will then lead to writing new items using the same process.

An important component of the Conceptual Model for Test Development is faculty expertise. Faculty must be clinical experts and be proficient in item-writing practices, and both are pre-requisites for developing reliable, valid, discriminating, and unbiased assessments of student learning. Faculty development in item writing is

therefore a critical component of the test development process. This component is often overlooked or left up to the faculty member (Tarrant & Ware, 2012). Previous research in multiple disciplines suggests that faculty development in item-writing principles, combined with the use of pre-established guidelines, results in significant improvement in the quality of MC test items (Caldwell & Pate, 2012; Jozefowicz et al., 2002; Naeem, van der Vleuten, & Alfaris, 2011; Reese, 1988; Van Ort & Hazzard, 1985; Wallach, Crespo, Hotzman, Galbraith, & Swanson, 2006). Implementing the FITr as a component of faculty development and a clear process for constructing high quality test items, along with repetition and practice, will logically lead to improvement in MC test items. Further research needs to be designed to evaluate this relationship, however, by testing the use of the FITr in writing and revising test items as a component of this model.

Limitations

This research study had several limitations. First, there were potential sampling biases. Eighty-six percent of the invited eligible participants did not respond to the survey invitation. It is not clear whether the characteristics of the non-respondents were different than those of the participants. It is also not clear whether the sample population from which participants were invited contained only nursing faculty names and email addresses. Several responses from invited participants declared their ineligibility due to non-faculty status, non-nursing status, retirement, and administrative/staff roles. It is highly likely that there were others on the list who were not nursing faculty and who did not declare their ineligibility. Another contributor to non-response is the fact that participant recruitment took place during the summer, a time when many faculty are off contract and therefore check email infrequently or are inaccessible. Systematic sampling

bias was also a factor during participant recruitment. The primary researcher was notified that participants from two large programs of nursing were ineligible without approval from the Institutional Review Boards (IRB) at both institutions. These participants were therefore systematically prevented from participating in the research study.

It is likely that the significant technical issues that occurred during the first week of data collection also contributed to non-response and participant attrition. Patterns of participant enrollment and survey completion suggest that the technical issues may have negatively impacted both participant enrollment and survey completion. Other unanticipated technical issues that were identified during data collection were with individual participants' technology – several reports of screen freezing and links that directed participants to random webpages were investigated by the primary investigator. It is highly likely that additional events occurred that were not reported. A final technical issue was the fact that the REDCap survey was not available on mobile devices, an issue that may have prevented access for some participants. While some technical issues are expected whenever technology is used for a project of this magnitude, it is suspected that the negative impact on participation rates for this study was more significant. For future research, alternate survey and database management software should be investigated and close attention should be paid to planning participant recruitment around scheduled outages.

Finally, following the recommendation of the pilot focus group, participants were provided detailed instructions for completing the survey in both written and video format. These instructions were optional, and there was no way to track who viewed them. As previously discussed, research has demonstrated that the quality of MC items can be

improved through faculty training in principles of item writing (Jozefowicz et al., 2002; Khan et al., 2013; Naeem et al., 2012). It is possible that participants who viewed these video instructions performed differently on the survey than they would have without the instructions or from those who did not view the instructions. Further research studies should be designed to test the impact of both the video instructions and more extensive education for faculty to improve the quality of MC test items.

Generalizability

Much of the discussion in the nursing literature related to improving test items is focused on preparation for the National Council Licensure Examination (NCLEX); however, these item-writing guidelines are consistent across nursing programs that use MC examinations. Therefore, the results of this study are generalizable within nursing education for writing and revising MC test items. The findings of this research study are not generalizable beyond nursing faculty. This study was designed specifically to evaluate a discipline-specific tool for nursing faculty use in identifying bias in MC test items. Previous published research reports have identified similar needs in other practice disciplines, such as medicine and pharmacy (Al-Faris, Alorainy, Abdel-Hameed, & Al-Rukban, 2010; Breitbach, 2010; Caldwell & Pate, 2012; Downing, 2002; Downing, 2005; Jozefowicz et al., 2002). There is also evidence that item-writing guidelines may be applied across practice disciplines (Naeem et al., 2012). Future research should examine whether the FITr can be applied to other practice disciplines.

Importance for Nursing Education

The FITr was developed to meet the need for discipline-specific guidelines to assist nursing faculty in improving the quality of MC test items. As previously discussed,

the FITr is one component of a comprehensive testing program, the implementation of which has the potential to transform assessment practices in schools of nursing. This statement may appear to be overly enthusiastic; however, the potential impact on stakeholders is far reaching. It is outside the boundaries of this research study to generalize beyond the implementation of the FITr in assessment practices, however, so the discussion of the importance for nursing education will be limited in scope.

All of the guidelines with the FITr are consistent with those used for MCQs on standardized and licensure examinations with one exception. As previously discussed, the use of three options was incorporated into the FITr as an alternative to using none-of-the-above or all-of-the-above. The efficacy of three-option test items is strongly supported by empirical data in educational and nursing literature (Considine, Bottie, & Thomas, 2005; McDonald, 2014; Piasentin, 2010; Redmond, Hartigan-Rogers, & Cobbett, 2012; Rodriguez, 2005; Sidick et al., 1994; Tarrant & Ware, 2010; Weaver, 1982) and needs to be implemented in teaching practice, which is unlikely until it is adopted by licensing bodies, and specifically, NCLEX. As previously discussed, research demonstrates that three-option test items are psychometrically comparable to four-option items but have the advantage of saving significant faculty time in item writing. Using three-option test items means that more items can be included on a test, which will more comprehensively measure the constructs being tested and provide better assessment of student learning. The use of three-option items must be implemented as a standard alternative in nursing education in order to ensure that assessment practices continue to be based on the best available evidence.

Previous research suggests there is a relationship between the use of clearly written item-writing guidelines, faculty development in item writing, and improved quality of MC test items (Caldwell & Pate, 2013; Hansen & Dexter, 1997; Morrison & Free, 2001). The most obvious impact of the implementation of the FITr as a component of assessment practices is on students and faculty. Previous studies have demonstrated that the presence of flawed test items negatively impacts student success and may particularly impact high achieving students and those for whom English is an additional language (EAL) (Bosher & Bowles, 2008; Downing, 2005; Tarrant & Ware, 2008). Improving the quality of MC test items used in nursing examinations has the potential to improve student success and better prepare all nursing students for licensure and certification examinations (Clifton & Schriener, 2010; McDonald, 2014). Indirectly, the FITr has the potential to increase the quality, quantity, and diversity of nurses joining the workforce. These improvements in student success also have a positive impact on nursing program accreditation rates and ability to recruit high quality students.

For faculty, increased student success equates to improved evaluations of faculty teaching effectiveness and less time devoted to remediating students who are performing poorly on examinations containing biased test items. Previous research has established that, although faculty frequently use textbook test bank items in assessments, these item banks are not secure and contain flawed test items; therefore, they should not be used for examination purposes without revision (Clifton & Schriener, 2010; Cross, 2000; Masters et al, 2001; Tarrant, Knierim, Hayes, & Ware, 2006). This research study has demonstrated that the FITr is useful for identifying bias in MC test items, and this usefulness has the potential to assist faculty in revising biased items obtained from

textbook test banks, saving faculty time in test item development and enhancing test security through modifying test items that are readily available to students. The FITr can also be instrumental for faculty in developing item banks of quality MC test items both through revising test items and writing new items. Having a readily available test bank of high quality MC test items can save faculty time in test development and provides a means to incorporate pilot questions in examinations to continually improve and add to the item bank.

Finally, a discussion of the impact on nursing faculty must include the benefits of implementing the FITr as one component of a systematic test development process (based on The Conceptual Model for Test Development), the most obvious of which is assistance in developing high quality MC test items. Improving the quality of test items is only relevant if those test items accurately reflect the curriculum and learning outcomes through deliberate planning during the test development process (Tarrant & Ware, 2012; Ware & Vik, 2009). However, there are broader implications for faculty as well – implementing a college-wide systematic test development process provides a means for recognizing the value of high quality assessments and the time commitment from faculty to develop these assessments. Such a process provides a means for documenting the workload impacts of item writing, sharing the responsibility for development and peer review among faculty, and accessing much-needed resources (such as item-banking and analysis software) and faculty development funds. Continued research needs to be conducted, however, to evaluate the impact of improving MC test items on licensure exam pass rates, progression, and faculty workload to provide support for the implementation of these changes.

Recommendations for Further Research

The next step in validating the FITr is to use confirmatory factor analysis (CFA) methods to strengthen the inferences about internal consistency. This research study demonstrated internal consistency reliability that was adequate for a newly developed tool ($\alpha > .60$). A much larger sample size will be needed for a research study using CFA. The common rule of thumb for scale development is to have 10 participants for every item contained in the scale ($n = 380$); however, some authors recommend 20 participants per item for factor analysis ($n = 760$) (Bannigan & Watson, 2009). In this research study, data were collected from over 380 participants; however, these were divided among the five surveys, so the maximum number of participants evaluating a test item with the complete FITr was 80. Subsequent research will need to draw from a much larger pool of potential participants ($N > 10,000$) in order to achieve a sufficient sample size for CFA. A research design in which all of the participants evaluate the same sample test is recommended to meet the requirements for CFA and to address the previous recommendations for including all guidelines in the analysis.

The FITr was designed for evaluating the quality of MCQs. During this research study, feedback from one of the expert reviewers suggested expanding the FITr to incorporate guidelines for the development of alternate item formats as well. An alternate item format, also known as innovative item type, is a test item that uses technology to deliver items in “a format other than the standard, four-option, multiple-choice items” (National Council of State Boards of Nursing, 2014a, ¶ 4). Examples of alternate item formats currently in use include multiple response items that require examinees to select multiple correct responses, calculation questions using fill-in-the-blank, “hot spot” items

in which examinees identify areas on a picture or graph, and ranking items (National Council of State Boards of Nursing, 2014a). Alternate item formats have been used on the national licensure examinations for nursing (NCLEX) since 2003 (National Council of State Boards of Nursing, 2014a). A recent survey by the National Board of Certification and Recertification for Nurse Anesthetists (NBCRNA) (2013) found that these items are also being used on specialty certification exams in nursing. Research studies evaluating alternate item formats have demonstrated that these items are psychometrically comparable to MC test items and provide a means to test higher level cognitive processes and constructs that are not possible with MC test items (McDonald, 2014; Wendt, 2008; Wendt & Harnes, 2009a; Wendt & Harnes, 2009b; Wendt & Kenny, 2009). McDonald (2014) recommends that students have practice with alternate item formats prior to the licensure exam, and nurse educators commonly believe that MC tests prepare students for the licensure and certification exams (Walloch, 2006). It is logical to conclude, then, that there is a need for guidelines related to the development of these items as well. A review of literature related to alternate item formats yielded only empirical studies related to items used in NCLEX. Subsequent research needs to be designed to investigate these item types and identify valid and reliable guidelines for their development.

This research study confirms the validity and reliability of the FITr for identifying bias in MC test items; however, this is only the first step toward improving the quality of test items. Further research needs to evaluate the use of the FITr for faculty use in writing and revising quality test items. Research studies should also be designed using the FITr as a framework to validate each of the item-writing guidelines empirically. Further

research is also needed to evaluate how a systematic process for test development and evaluation that incorporates the FITr impacts student progression and success on licensure examinations.

Conclusion

There is a need for development of a valid and reliable tool for use by nurse educators in evaluating and revising MC test items. This dissertation study contributed to the body of knowledge by establishing the validity and reliability the Fairness of Items Tool (FITr) for nursing faculty use in identifying bias in MC test questions. The FITr provides a means to facilitate systematic research to validate item-writing guidelines, testing procedures, and the actual quality of test items. Use of the FITr in nursing education has the potential to improve MC assessments, better prepare students for success on the licensure examination, and enhance the quantity and diversity of the nursing workforce.

REFERENCES

- Abedi, J. (2006). Language issues in item development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.
- Aiken, L. R. (1987). Testing with multiple-choice items. *Journal of Research and Development in Education*, 20(4), 44-58.
- Al-Faris, E. A., Alorainy, I. A., Abdel-Hameed, A. A., & Al-Rukban, M. O. (2010). A practical discussion to avoid common pitfalls when constructing multiple choice questions items. *Journal of Family and Community Medicine*, 17(2), 91-97.
doi:10.4103/1319-1683.71992
- American Association of Colleges of Nursing. (1997). *AACN position statement: Diversity and equality of opportunity*. Retrieved from <http://www.aacn.nche.edu/Publications/positions /diverse.htm>
- American Association of Colleges of Nursing. (2014a). *Enhancing diversity in the nursing workforce*. Retrieved from <http://www.aacn.nche.edu/media-relations/diversityFS.pdf>
- American Association of Colleges of Nursing. (2014b). *The nursing faculty shortage*. Retrieved from <http://www.aacn.nche.edu/media-relations/fact-sheets/nursing-faculty-shortage>
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

- Andrews, D. R. (2003). Lessons from the past: Confronting past discriminatory practices to alleviate the nursing shortage through increased professional diversity. *Journal of Professional Nursing*, 19(5), 289-294. doi:10.1016/S8755-7223(03)00099-1
- Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, 20(2), 153-170.
- Ayoola, A. (2013). *Why diversity in the nursing workforce matters*. Robert Wood Johnson Foundation. Retrieved from http://www.rwjf.org/en/blogs/human-capital-blog/2013/05/why_diversity_inthe.html
- Bailey, P. H., Mossey, S., Moroso, S., Cloutier, J D., & 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
- Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, 18(23), 3237-3243. doi:10.1111/j.1365-2702.2009.02939.x
- Begum, T. (2012). A guideline on developing effective multiple choice questions and construction of single best answer format. *Journal of Bangladesh College of Physicians and Surgeons*, 30, 159-166.
- Billings, D. M., & Halstead, J. A. (2009). *Teaching in nursing: A guide for faculty* (3rd ed.). St. Louis, MO: Saunders.
- Boland, R. J., Lester, N. A., & Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry*, 34(4), 310-316.

- Bosher, S. (2003). Barriers to creating a more culturally diverse nursing profession: Linguistic bias in multiple-choice nursing exams. *Nursing Education Perspectives*, 24(1), 25-34.
- Bosher, S. D. (2009). Removing language as a barrier to success on multiple-choice nursing exams. In S. D. Bosher & M. D. Pharris (Eds.), *Transforming nursing education: The culturally inclusive environment* (pp. 259-284). New York, NY: Springer.
- Bosher, S., & Bowles, M. (2008). The effects of linguistic modification on ESL students' comprehension of nursing course test items. *Nursing Education Perspectives*, 29(3), 165-172.
- Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, 5, 238-242.
- Breitbach, A. P. (2010). Creating effective multiple choice items. *Athletic Therapy Today*, 15(3), 18-22.
- Buerhaus, P., Staiger, D., & Auerbach, D. (2008). *The future of the nursing workforce: Data, trends, and implications*. Boston, MA: Jones & Bartlett.
- Bureau of Labor Statistics, U.S. Department of Labor. (2014). Registered nurses. *Occupational Outlook Handbook, 2014-15 Edition*. Retrieved from <http://www.bls.gov/ooh/healthcare/registered-nurses.htm>
- Burns, C. M. (2009). Sold! Web-based auction sites have just compromised your test bank. *Nurse Educator*, 34(3), 95-96.

- Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education*, 30(1), 65-72.
doi:10.1080/0260293042000243904
- Caldwell D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, 77(4), 1-5.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.
- Campbell, D. E. (2011). How to write good multiple-choice questions. *Journal of Paediatrics and Child Health*, 47(6), 322-325. doi:10.1111/j.1440-1754.2011.02115.x
- Case, S. M., & Donahue, B. E. (2008). Developing high-quality multiple-choice questions for assessment in legal education. *Journal of Legal Education*, 58(3), 372-387.
- Case, S. M., & Swanson, D. B. (2002). *Constructing writing test questions for the basic and clinical sciences* (3rd ed. revised). Philadelphia, PA: National Board of Medical Examiners.
- Chenevey, B. (1988). Constructing multiple-choice examinations: Item writing. *The Journal of Continuing Education in Nursing*, 19(5), 201-204.
- Clifton, S. L., & Schriener, C. L. (2010). Assessing the quality of multiple-choice test items. *Nurse Educator*, 35(1), 12-16.
- Considine, J., Bottie, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, 12(1), 19-24.

- Cross, K. J. W. (2000). *Cognitive levels of multiple-choice items on teacher-made tests in nursing education* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9992186)
- Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, 14(3), 197-201.
- Demetrulias, D. A. M., & McCubbin, L. E. (1982). Constructing test questions for higher level thinking. *Nurse Educator*, 7(5), 13-17.
- DePew, D. D. (2001). *Validity and reliability in nursing multiple-choice tests and the relationship to NCLEX-RN success: An Internet survey* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3040905)
- DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., . . . Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, 39(2), 155-164.
- Downing, S. M. (2002a). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77(10), S103-S104.
- Downing, S. M. (2002b). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, 7(3), 235-241.
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830-837. doi:10.1046/j.1365-2923.2003.01594.x

- Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, 10(2), 133-143. doi:10.1007/s10459-004-4019-5
- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.
- Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38(3), 327-333. doi:10.1046/j.1365-2923.2004.01777.x
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.
- Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, 83(5), 289-293.
- Evans, B. C., & Greenberg, E. (2006). Atmosphere, tolerance, and cultural competence in a baccalaureate nursing program: Outcomes of a nursing workforce diversity grant. *Journal of Transcultural Nursing*, 17(3), 298-305. doi:10.1177/1043659606288374
- Farley, J. K. (1989). The multiple-choice test: Writing the questions. *Nurse Educator*, 14(6), 10-12, 39.
- Fishman, J. A., & Galguera, T. (2003). *Introduction to test construction in the social and behavioral sciences: A practical guide*. Lanham, MD: Rowman & Littlefield.

- Flynn, M. K., & Reese, J. L. (1988). Development and evaluation of classroom tests: A practical application. *Journal of Nursing Education*, 27(2), 61-65.
- Frey, B. B., Petersen, S., Edwards, L. M., Pedrotti, J. T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, 21(4), 357-364. doi:10.1016/j.tate.2005.01.0084
- Froman, R. D. (2001). Elements to consider in planning the use of factor analysis. *Southern Online Journal of Nursing Research*, 5(2), 20 pages. Retrieved from <http://www.snrs.org>
- Gaberson, K. B. (1996). Test design: Putting all the pieces together. *Nurse Educator*, 21(4), 28-33.
- Giddens, J. F. (2009). Changing paradigms and challenging assumptions: Redefining quality and NCLEX-RN pass rates. *Journal of Nursing Education*, 48(3), 123-124.
- Gronlund, N. E. (1998). *Assessment of student achievement* (6th ed.). Needham Heights, MA: Allyn & Bacon.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Haladyna, T. M., & Downing, S. M. (1985, April). *A quantitative review of research on multiple-choice item writing*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 37-50.

- Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51-78.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334.
- Hambleton, R. K., & Rodgers, H. J. (2005). *Developing an item bias review form*. Retrieved from Clearinghouse on Assessment and Evaluation:
<http://ericae.net/ft/tamu/biaspub2.htm>
- Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business*, 73(2), 94-97.
- Hays, R. B., Coventry, P., Wilcock, D., & Harley, K. (2009). Short and long multiple-choice question stems in a primary care oriented undergraduate medical curriculum. *Education for Primary Care*, 20, 173-177.
- Hicks, N. A. (2011). Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. *Nurse Educator*, 36(6), 266-270.
doi:10.1097/NNE.0b013e3182333fd2
- Institute of Medicine. (2011). *The future of nursing: Leading change, advancing health*. Washington, DC: National Academies Press.
- Joynt, J., & Kimball, B. (2008). *Blowing open the bottleneck: Designing new approaches to increase nurse education capacity* (Report). Washington, DC: Center to Champion Nursing in America.

- Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, 77(2), 156-161.
- Kaufman, K. A. (2010). Findings from the 2009 faculty census: Study confirms reported demographic trends and inequities in faculty salaries. *Nursing Education Perspectives*, 11(6), 404-405.
- Kaufman, K. (2007). Introducing the NLN/Carnegie national survey of nurse educators: Compensation, workload, and teaching practice. *Nursing Education Perspectives*, 28(3), 164-169.
- Kelly, A. L. (1998). *The use of constructed response testing in nursing education* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9834118)
- Khan, H. F., Danish, K. F., Awan, A. S., & Anwar, M. (2013). Identification of technical item flaws leads to improvement of the quality of single best multiple choice questions. *Pakistan Journal of Medical Science*, 29(3), 715-718.
- King, E. C. (1978). Constructing classroom achievement tests. *Nurse Educator*, 3(5), 30-36. doi:10.1097/00006223-197809000-00020
- Klisch, M. L. (1994). Guidelines for reducing bias in nursing examinations. *Nurse Educator*, 19(2), 35-39.
- Lampe, S., & Tsaouse, B. (2010). Linguistic bias in multiple-choice test questions. *Creative Nursing*, 16(2), 63-67.
- Layton, J. M. (1986). Validity and reliability of teacher-made tests. *Journal of Nursing Staff Development*, 2(3), 105-109.

- Lynn, M.R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–385.
- Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review process improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), 919-931.
- Martínez, R. J., Moreno, R., Martín, I., & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema*, 21(2), 326-330.
- Masters, J. C., Hulsmeier, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education*, 40(1), 25-32.
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, 26(8), 709-712. doi:10.1080/01421590400013495
- McDonald, M. E. (2014). *The nurse educator's guide to assessing learning outcomes* (3rd ed.). Sudbury, MA: Jones and Bartlett.
- McNeal, G. (2012). The nurse faculty shortage. *ABNF Journal*, 23(2), 23.
- Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology: A practical introduction*. Los Angeles, CA: Sage.
- Moreno, R., Martínez, R. J., & Muñiz, J. (2004). *Brief summary of the 12 guidelines for the construction of multiple choice test items*. Retrieved from <http://www.personal.us.es/rmoreno/resdiri.htm>

- Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, 2(2), 65-72. doi:10.1027/1614-1881.2.2.65
- Morrison, S. (2005). Chapter 5: Improving NCLEX-RN pass rates through internal and external curriculum evaluation. *Annual Review of Nursing Education*, 3, 77-94.
- Morrison, S., & Free, K. W. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education*, 40(1), 17-24.
- Morrison, S., Nibert, A., & Flick, J. (2006). *Critical thinking and test item writing* (2nd ed.). Houston, TX: Health Education Systems.
- Naeem, N., van der Vleuten, C., & Alfaris, E. A. (2011). Faculty development on item writing substantially improves item quality. *Advances in Health Sciences Education*, 17, 369-376.
- National Board of Certification and Recertification for Nurse Anesthetists. (2013). *A study of alternate item formats in accredited certification programs*. Retrieved from <http://www.nursingcertification.org/pdf/Alternate%20Item%20Formats%20Study%20-%20Combined%20Summary%202014%2001%2030.pdf>
- National Council of State Boards of Nursing. (2014a). *Alternate item format FAQs*. Retrieved from <https://www.ncsbn.org/2334.htm>
- National Council of State Boards of Nursing. (2014b). *Quarterly examination statistics*. Retrieved from <http://www.ncsbn.org>
- National League for Nursing. (2009). *Executive summary: Findings from the 2009 faculty census*. Retrieved from http://www.nln.org/researchgrants/slides/fc_exec_summary_0809.pdf

- National League for Nursing. (2013). *Findings from the annual survey of schools of nursing academic year 2011-2012*. Retrieved from <http://www.nln.org/researchgrants/slides/execsummary2012.pdf>
- National League for Nursing. (2014a). Percentage of programs that are highly selective by program type, 2012. *NLN DataView^{TM}*. Retrieved from http://www.nln.org/researchgrants/slides/pdf/AS1112_F20.pdf
- National League for Nursing. (2014b). Rank of full-time nurse educators by race-ethnicity, 2009. *NLN DataView^{TM}*. Retrieved from http://www.nln.org/researchgrants/slides/pdf/FC0809_F11.pdf
- Nibert, A. T. (2003). *Predicting NCLEX success with the HESI exit exam: Results from four years of study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3084182)
- Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, 15(8), 873-885. doi:10.1080/09658210701746621
- Oermann, M. H., Saewert, S. J., Charasika, M., & Yarbrough, S. S. (2009). Assessment and grading practices in schools of nursing: National survey findings part I. *Nursing Education Perspectives*, 30(5), 274-278.
- O'Neill, T. R., Marks, C., & Liu, W. (2006, Winter). Assessing the impact of English as a second language status on licensure examinations. *CLEAR Exam Review*, 17(1), 14-18.
- Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Norwell, MA: Kluwer Academic.

- Paxton, M. (2000). A linguistic perspective on multiple choice questioning. *Assessment & Evaluation in Higher Education*, 25(2), 109-119.
- Petress, K. (2007). How to make college tests more relevant, valid, and useful for instructors and students. *College Student Journal*, 41(4), 1098-1101.
- Polit, D. F., & Beck, C. T. (2012). *Nursing research: Generating and assessing evidence for nursing practice* (9th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
- Qualls, M., Pallin, D. J., & Schuur, J. (2010). Parametric versus nonparametric statistical tests: The length of stay example. *Academic Emergency Medicine*, 17(10), 1113-1121. doi:10.1111/j.1553-2712.2010.00874.x
- Quinn, F. M. (2000). *The principles and practice of nurse education* (4th ed.). Cheltenham, United Kingdom: Stanley Thornes.
- Quinn, F. M., & Hughes, S. J. (2007). *Quinn's principles and practice of nurse education* (5th ed.). Cheltenham, United Kingdom: Nelson Thornes.
- Redmond, S. P., Hartigan-Rogers, J. A., & Cobbett, S. (2012). High time for a change: Psychometric analysis of multiple-choice questions in nursing. *International Journal of Nursing Education Scholarship*, 9(1), 1-16.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.
- Rogausch, A., Hofer, R., & Krebs, R. (2010). Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: A simulation and survey. *BMC Medical Education*, 10(1), 85-93. doi:10.1186/1472-6920-10-85

- Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychologist*, 19(4), 219-225.
- Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics and standard setting. *Medical Education*, 48(10), 1020-1027. doi:10.1111/medu.12525
- Schroeder, J. M. (2007). *A study of improving critical thinking skills with multiple choice tests and first semester associate degree nursing students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3278154)
- Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology*, 47(4), 829-835.
- Sitzman, K. L. (2007). Diversity and the NCLEX-RN: A double-loop approach. *Journal of Transcultural Nursing*, 18(3), 271-276. doi:10.1177/1043659607301302
- Stangroom, J. (2014). Z-test for 2 population proportions. *Social Science Statistics*. Retrieved from <http://www.socscistatistics.com/>
- Stanton, J. P. H. (1983). Objective test construction – A must for nurse educators. *Journal of Nursing Education*, 22(8), 338-339.
- Stuart, C. C. (2013). *Mentoring, learning, and assessment in clinical practice* (3rd ed.). Oxford, United Kingdom: Churchill Livingstone.
- Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice*, 6(6), 354-363.

- Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42, 198-206. doi:10.1111/j.1365-2923.2007.02957.x
- Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today*, 30(6), 539-543. doi:10.1016/j.nedt.2009.11.002
- Tarrant, M., & Ware, J. (2012). A framework for improving the quality of multiple-choice assessments. *Nurse Educator*, 37(3), 98-104.
- Taylor, A. K. (2005). Violating conventional wisdom in multiple choice test construction. *College Student Journal*, 39(1), 141-148.
- Taxis, J. (2002). The underrepresentation of Hispanics/Latinos in nursing education: A deafening silence. *Research and Theory for Nursing Practice*, 16(4), 249-262.
- Uebersax, J. (2009). *Statistical methods for rater and diagnostic agreement*. Retrieved from <http://www.john-uebersax.com/stat/agree.htm>
- Vacc, N. A., Loesch, L. C., & Lubik, R. E. (2001). *Writing multiple-choice test items*. Retrieved from ERIC database. (ED457440)
- Van Ort, S., & Hazzard, M. E. (1985). A guide for evaluation of test items. *Nurse Educator*, 10(5), 13-15.
- Van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. *Quality and Quantity*, 40(3), 435-456. doi:10.1007/s11135-005-8081-8

- Wallach, P. M., Crespo, L. M., Hotzman, K. Z., Galbraith, R. M., & Swanson, D. B. (2006). Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education, 11*(1), 61-68.
doi:10.1007/s10459-004-7515-8
- Walloch, J. A. (2006). *Assessment practices in the nursing classroom: An exploration of educators' assessment of students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3233950)
- Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). New York, NY: Springer.
- Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher, 31*(3), 238-243. doi:10.1080/01421590802155597
- Weaver, P. (1982). Tutorial on multiple choice items. *Journal of Marketing Education, 4*(1), 62-67. doi:10.1177/027347538200400111
- Wendt, A. (2008). Investigation of the item characteristics of innovative item formats. *CLEAR Exam Review, 19*(1), 22-28.
- Wendt, A., & Harnes, J. C. (2009a). Developing and evaluating innovative items for the NCLEX: Item characteristics and cognitive processing. *Nurse Educator, 34*(3), 109-113.
- Wendt, A., & Harnes, J. C. (2009b). Evaluating innovative items for the NCLEX: Usability and pilot testing. *Nurse Educator, 34*(2), 56-59.
- Wendt, A., & Kenny, L. E. (2009). Alternate item types: Continuing the quest for authentic testing. *Journal of Nursing Education, 48*(3), 150-156.

- Wendt, A., & Worcester, P. (2000). The National Council Licensure Examinations differential item functioning process. *Journal of Nursing Education*, 39(4), 185-197.
- Woo, A., & Dragan, M. (2012). Ensuring validity of NCLEX with differential item functioning analysis. *Journal of Nursing Regulation*, 2(4), 29-31.
- Worthington, R., & Whittaker, T. (2006). Scale development research: A content analysis and recommendations for best practices. *Counseling Psychologist*, 34(6), 806-838. doi:10.1177/0011000006288127
- Zieky, M. (2006). Fairness review in assessment. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.
- Zungolo, E. (2008, November). *National League of Nursing: Issues related to faculty shortage*. Presented at the 119th meeting of the National Advisory Council on Nurse Education and Practice, Bethesda, MD.

APPENDIX A
FAIRNESS OF ITEMS TOOL

Fairness of Items Tool (FIT)

Copyright 2012 by Nikole A. Hicks

Evaluate the Stem

1. Use a question format.
2. Eliminate *of the following*.
3. Present a single, clearly defined question with the problem in the stem.
4. Avoid absolute terms (always, never, all).
5. Avoid negatively phrased questions, double negatives, and the use of *except*.
6. Best answer format – underline, capitalize, and bold key words (**BEST**, **MOST**).
7. Avoid trick questions.
8. Avoid conditional expressions (should, would) and passive voice.
9. Write questions at the application or above cognitive level.
10. Write questions that require multi-logical thinking.
11. Make sure content is current.
12. Avoid testing student opinions.
13. Test important content and avoid trivia.

Evaluate the Options

14. Make sure options are similar in length and amount of detail.
15. Make sure options are grammatically and visually similar.
16. Avoid none-of-the-above and all-of-the-above.
17. Avoid negatively phrased options.
18. Avoid repeating material in the options – move repetitive words to the stem.
19. Avoid repeating words in the stem and correct option.
20. Make sure there is one, and only one, correct answer.
21. Eliminate multiple-multiples.
22. Make sure all distracters are plausible.
23. If the stem asks what should be done first or which action is best, all options must be correct with only one option being the first or best.
24. Avoid overlapping options.
25. Write options that require a high level of discrimination to select the correct answer.

Linguistic Bias

26. Use a parsimonious style and short simple sentences.
27. Use precise terms (avoid frequently, appropriate).
28. Use straight-forward, uncomplicated language. Test nursing content, not vocabulary or reading.
29. Ensure that items are independent of each other.

Structural Bias

30. Use correct grammar, punctuation, capitalization, and spelling.
31. Write items that can be read and comprehended easily on the first reading.
32. Be specific and clear with directions.
33. Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.

Cultural Bias

34. Avoid dominant culture (literature, music, movies, sports, foods) unless essential to safe, effective nursing practice.
35. Eliminate all names.
36. Eliminate all slang.
37. Use terminology from textbook, notes, and common words (toilet vs. commode).
38. Eliminate humor.
39. Avoid stereotyping and over-representation of cultural groups.
40. Use gender-neutral language.
41. Present the person first, not the diagnosis.

APPENDIX B

PERMISSION TO USE CONTENT FROM
EDUCATIONAL PSYCHOLOGIST



[Welcome](#) |
 [Log in](#) |
 [Cart \(0\)](#) |
 [Manage Account](#) |
 [Feedback](#) |
 [Help](#) |
 [Live Help](#)

[GET PERMISSION](#) |
 [LICENSE YOUR CONTENT](#) |
 [PRODUCTS AND SOLUTIONS](#) |
 [PARTNERS](#) |
 [EDUCATION](#) |
 [ABOUT US](#)

Get Permission / Find Title

[Advanced Search Options](#)

Educational psychologist

ISSN:	0046-1520	Language:	English
Publication year(s):	1963 - present	Country of publication:	United States of America
Author/Editor:	AMERICAN PSYCHOLOGICAL ASSOCIATION		
Publication type:	Journal		
Publisher:	LAWRENCE/ERLBAUM ASSOCIATES, INC.		
Rights holder:	TAYLOR & FRANCIS INFORMA UK LTD - JOURNALS		

Permission type selected: Republish or display content

Type of use selected: reuse in a thesis/dissertation

[Select different permission](#)

Article title: A theoretical framework for the exploration of causes and effects of bias in testing

Author(s): Scheuneman, Janice Dowd

DOI: 10.1080/00461528409529298

Date: Sep 1, 1984

Volume: 19

Issue: 4

[Select different article](#)

Taylor & Francis is pleased to offer reuses of its content for a thesis or dissertation free of charge contingent on resubmission of permission request if work is published.

APPENDIX C

PERMISSION TO USE CONTENT FROM HEALTH
EDUCATION SYSTEMS, INC. / ELSEVIER

RE: Permission to use Conceptual Model

Nibert, Ainslie (ELS-HOU) [a.nibert@elsevier.com]

Sent: Friday, August 19, 2011 8:47 AM

To: HESI-info (ELS) [HESI-info@elsevier.com]; Nikole Hicks

Cc: Willson, Pamela (ELS-HOU) [Pa.Willson@elsevier.com]

Thanks for your inquiry, Nikole, and I would be happy for you to use the model in your dissertation. Please provide the appropriate citation for the source --- that is really all that is

necessary, since the book was published by HESI, not by Elsevier. (For example, the APA-formatted citation is: Morrison, S., Nibert, A. & Flick, J. (2006). *Critical Thinking and Test Item*

Writing, 2nd ed.. Houston, TX: Health Education Systems, Inc.

I am also copying Dr. Pam Willson, who is our Director of Research, on this message, since I know that she will be very interested in hearing more about your research study. We are

always interested in hearing about new nursing education science, and Pam has an extensive network of other nursing education faculty members that she can introduce you to who are

also leading advances in nursing education research, particularly in the areas of test and measurement.

Good luck with your dissertation study, and we will be interested in hearing periodically about your progress with it! --- Ainslie

ELSEVIER

Ainslie Nibert, PhD, RN • Vice President • Review and Testing/HESI • Health Sciences

Office: 800.950.2728, ext. 6910 • Mobile: 832.675.1561 • a.nibert@elsevier.com

APPENDIX D

PERMISSION TO USE QUINN'S (2000) *CARDINAL
CRITERIA FOR ASSESSMENT*

From: Woods, Amy [mailto:Amy.Woods@cengage.com]

Sent: Tuesday, February 07, 2012 5:52 AM

To: Hicks, Nikole (hicksnk)

Subject: RE: Permission Requested

Dear Nikole Hicks

The Principles and Practice of Nurse Education 4th edition by Quinn - 0748738959

1. The material to be quoted/produced was published without credit to another source. If another source is acknowledged, please apply directly to that source for permission clearance.
2. Permission is for non-exclusive, English language rights, and covers use in your dissertation only (print and/or electronic format of dissertation) only. Any further use shall be the subject of a separate application for permission.
3. Full acknowledgement must be given to the original source, with full details of figure/page numbers, title, author(s), publisher and year of publication.

Yours sincerely

Amy Woods

Permissions & Subsidiary Rights Administrator

Cengage Learning (EMEA) Ltd

Cheriton House, North Way, Andover, Hampshire, SP10 5BE, United Kingdom

(t) +44 (0) 1264 342873 | (f) +44 (0) 1264 342792 | (e) amy.woods@cengage.com

APPENDIX E
EXPERT PANEL SURVEY

Confidential

Page 1 of 8

Validation of the Fairness of Items Tool (FIT) - Expert Panel

Thank you for agreeing to participate in this dissertation study to validate the Fairness of Items Tool (FIT). You have been selected to participate on an expert panel for this dissertation study because of your expertise in test item construction and analysis in nursing. Your feedback is appreciated and will contribute to the advancement of nursing science in test item construction.

There is a need for development of a valid and reliable tool for use by nursing faculty in evaluating bias in multiple-choice test items. Developing a clear and concise guideline for nursing faculty will improve the quality of nursing assessments, thereby improving the quality of the decisions made based on these assessments. This dissertation study will contribute to the body of knowledge by establishing the validity and reliability of such a tool. This tool can then be used for further research to validate item-writing guidelines, evaluate the impact of item bias on student success, and better prepare nurse educators to design valid and reliable assessments of student learning.

The Fairness of Items Tool (FIT) was developed by the primary researcher through an exhaustive search of the literature from 1978-2011. The FIT is intended to serve as a discipline-specific taxonomy for use by nurse educators in evaluating and revising multiple-choice test items. The FIT contains 41 item-writing guidelines according to five dimensions of bias: the stem, the options, linguistic, structural, and cultural.

In this survey, you will be asked to review the FIT and evaluate the relevance of each guideline, the tool's organization, ease of use, and completeness. The survey contains write-in space for listing additional items for inclusion in the tool and for general comments/feedback. If needed, you will have the option to save and return later to complete the survey.

Please review the Fairness of Items Tool (FIT) (attachment below) and evaluate the relevancy of each guideline, the tool's organization, ease of use, and completeness. If you wish to comment on any of your replies, there is space below each for you to do so.

[Attachment: "Fairness of Items Tool.pdf"]

Please evaluate the relevance of each guideline under EVALUATE THE STEM.

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use a question format.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Eliminate of the following.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Present a single, clearly defined question with the problem in the stem.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Avoid absolute terms (always, never, all).	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Avoid negatively phrased questions, double negatives, and the use of except.	☐	☐	☐	☐

Comments

Best answer format - underline, capitalize, and bold key words (BEST, MOST).	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Avoid trick questions.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Avoid conditional expressions (should, would) and passive voice.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Write questions at the application or above cognitive level.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Write questions that require multi-logical thinking.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Make sure content is current.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Avoid testing student opinions.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Test important content and avoid trivia.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				

Please evaluate the relevance of each guideline under EVALUATE THE OPTIONS.

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Make sure options are similar in length and amount of detail.	☐	☐	☐	☐

Comments

Make sure options are grammatically and visually similar.	☐	☐	☐	☐

Comments

Avoid none-of-the-above and all-of-the-above.	☐	☐	☐	☐

Comments

Avoid negatively phrased options.	☐	☐	☐	☐

Comments

Avoid repeating material in the options - move repetitive words to the stem.	☐	☐	☐	☐

Comments

Avoid repeating words in the stem and correct option.	☐	☐	☐	☐

Comments

Make sure there is one, and only one, correct answer.	☐	☐	☐	☐

Comments

Eliminate multiple-multiples.	☐	☐	☐	☐

Comments

Make sure all distracters are plausible.	☐	☐	☐	☐

Comments

If the stem asks what should be done first or which action is best, all options must be correct with only one option being the first or best.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Avoid overlapping options.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Write options that require a high level of discrimination to select the correct answer.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Please evaluate the relevance of each guideline under LINGUISTIC BIAS.

Use a parsimonious style and short simple sentences.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Use precise terms (avoid frequently, appropriate).

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Use straight-forward, uncomplicated language. Test nursing content, not vocabulary or reading.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Ensure that items are independent of each other.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Please evaluate the relevance of each guideline under STRUCTURAL BIAS.

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use correct grammar, punctuation, capitalization, and spelling.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Write items that can be read and comprehended easily on the first reading.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Be specific and clear with directions.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.	☐	☐	☐	☐

Comments

Please evaluate the relevance of each guideline under CULTURAL BIAS.

Avoid dominant culture (literature, music, movies, sports, foods) unless essential to safe, effective nursing practice.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Eliminate all names.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Eliminate all slang.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Use terminology from textbook, notes, and common words (toilet vs. commode).	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Eliminate humor.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Avoid stereotyping and over-representation of cultural groups.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Use gender-neutral language.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Present the person first, not the diagnosis.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐

Comments

Please evaluate the following attributes considering the overall FIT as a tool for nursing faculty to evaluate and revise multiple-choice test items.

The FIT's overall organization.	Not organized ☐	Somewhat organized ☐	Quite organized ☐	Highly organized ☐

Comments

The FIT's ease of use.	Not easy to use ☐	Somewhat easy to use ☐	Quite easy to use ☐	Highly easy to use ☐

Comments

The FIT's completeness.	Not complete ☐	Somewhat complete ☐	Quite complete ☐	Highly complete ☐

Comments

Please comment on any additional guidelines that should be included in the FIT.

Please use this space for additional feedback or general comments not included above.

APPENDIX F

EXPERT PANEL SURVEY- REVISED FIT

Confidential

Page 1 of 8

Revised Fairness of Items Tool (FIT) - Expert Panel

Thank you for agreeing to participate in this dissertation study to review the Revised Fairness of Items Tool (FIT). Your feedback is appreciated and will contribute to the advancement of nursing science in test item construction.

The Fairness of Items Tool (FIT) was revised by the primary researcher based on data analysis from the expert panel. Several changes were made to the FIT based on expert feedback and analysis of literature. The revised FIT contains 38 item-writing guidelines according to four dimensions of bias: the stem, the options, linguistic/structural, and cultural.

In this survey, you will be asked to review the revised FIT and evaluate the relevance of each guideline, the tool's organization, ease of use, and completeness. The survey contains write-in space for listing additional items for inclusion in the tool and for general comments/feedback. If needed, you will have the option to save and return later to complete the survey.

Please review the revised Fairness of Items Tool (FIT) (attachment below) and evaluate the relevancy of each guideline, the tool's organization, ease of use, and completeness. If you wish to comment on any of your replies, there is space below each for you to do so.

[Attachment: "Fairness of Items Tool - Revised.pdf"]

Please evaluate the relevance of each guideline under EVALUATE THE STEM.

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use a question format.	☐	☐	☐	☐
Comments				
Eliminate extraneous words (e.g., of the following).	☐	☐	☐	☐
Comments				
Present a single, clearly defined question with the problem in the stem.	☐	☐	☐	☐
Comments				
Avoid negatively phrased questions, double negatives, and the use of except.	☐	☐	☐	☐
Comments				
Use active verbs and present tense.	☐	☐	☐	☐
Comments				
Write questions at the application or above cognitive level.	☐	☐	☐	☐
Comments				

Write questions that require multilogical thinking (require knowledge of more than one fact/concept).	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Make sure content is current.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Avoid testing student opinions (e.g., use nurse instead of you as the subject).	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Test important content and avoid trivia.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Avoid overlapping options.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Eliminate multiple-multiples.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				
Make sure all distracters are plausible.	Not relevant ☐	Somewhat relevant ☐	Quite relevant ☐	Highly relevant ☐
Comments				

Please evaluate the relevance of each guideline under EVALUATE THE OPTIONS.

Make sure options are similar grammatically and in length and amount of detail.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Avoid none-of-the-above and all-of-the-above. Use three options instead.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Avoid negatively phrased options.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Avoid repeating material in the options - move repetitive words to the stem.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Avoid repeating words in the stem and correct option.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Avoid overlapping options.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Eliminate multiple-multiples.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Make sure all distracters are plausible.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

If the stem asks what should be done first or which action is best, all options must be correct with only one option being the first or best.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Make sure there is only one correct answer.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Write options that require a high level of discrimination to select the correct answer.

Not relevant ☐   Somewhat relevant ☐   Quite relevant ☐   Highly relevant ☐

Comments

Please evaluate the relevance of each guideline under LINGUISTIC/STRUCTURAL BIAS.

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use a parsimonious style and short simple sentences.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use correct grammar, punctuation, capitalization, and spelling.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use precise terms (avoid frequently, appropriate).	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Avoid absolute terms (always, never, all).	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use straight-forward, uncomplicated language. Test nursing content, not vocabulary or reading.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Write items that can be comprehended on the first reading. Avoid tricky or misleading items.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Ensure that items are independent of each other.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Be specific and clear with directions.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.	☐	☐	☐	☐

Comments

Please evaluate the relevance of each guideline under CULTURAL BIAS.

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Avoid dominant culture (literature, music, movies, sports, foods) unless essential to safe, effective nursing practice.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Eliminate all names.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Eliminate all slang.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use terminology from textbook, notes, and common words (home vs. abode).	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Eliminate humor.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Avoid stereotyping and over-representation of cultural groups.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Use gender-specific language only when necessary to test nursing content.	☐	☐	☐	☐

Comments

	Not relevant	Somewhat relevant	Quite relevant	Highly relevant
Present the person first, not the diagnosis.	☐	☐	☐	☐

Comments

Please evaluate the following attributes considering the revised FIT as a tool for nursing faculty to evaluate and revise multiple-choice test items.

	Not organized	Somewhat organized	Quite organized	Highly organized
The revised FIT's overall organization.	☐	☐	☐	☐

Comments

	Not easy to use	Somewhat easy to use	Quite easy to use	Highly easy to use
The revised FIT's ease of use.	☐	☒	☐	☐

Comments

	Not complete	Somewhat complete	Quite complete	Highly complete
The revised FIT's completeness.	☐	☐	☐	☐

Comments

Please comment on any additional guidelines that should be included in the FIT.

Please use this space for additional feedback or general comments not included above.

APPENDIX G
RESULTS OF EXPERT PANEL SURVEY 1 & 2

Table G.1

Ratings from Expert Panel Review 1: Items Rated 3 or 4 on a 4-point Relevance Scale

Item	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Number in Agreement	Item CVI
1	X	X	--	X	X	4	.8
2	X	X	--	--	X	3	.6
3	X	X	X	X	X	5	1.0
4	X	X	X	X	X	5	1.0
5	X	X	X	X	X	5	1.0
6	--	--	X	X	--	2	.4
7	X	X	X	X	X	5	1.0
8	--	X	X	--	X	3	.6
9	X	X	--	X	X	4	.8
10	X	X	X	X	X	5	1.0
11	X	X	--	X	X	4	.8
12	X	X	--	X	X	4	.8
13	X	X	X	X	X	5	1.0
14	X	X	X	X	X	5	1.0
15	X	X	X	X	X	5	1.0
16	X	X	X	X	X	5	1.0
17	X	X	X	X	X	5	1.0
18	X	X	X	X	X	5	1.0
19	X	X	X	X	X	5	1.0
20	X	X	X	X	X	5	1.0
21	X	X	--	X	X	4	.8
22	X	X	X	X	X	5	1.0
23	X	X	X	X	X	5	1.0
24	X	X	X	X	X	5	1.0
25	X	X	--	X	X	4	.8
26	X	X	X	X	X	5	1.0
27	X	X	X	X	X	5	1.0
28	X	X	X	X	X	5	1.0
29	X	X	X	X	X	5	1.0
30	X	X	X	X	X	5	1.0
31	--	X	--	X	X	3	.6
32	X	X	X	X	X	5	1.0
33	X	X	X	X	X	5	1.0
34	X	X	--	X	X	4	.8
35	X	X	X	--	X	4	.8
36	X	X	X	X	X	5	1.0
37	X	X	--	X	X	4	.8
38	X	X	X	--	X	4	.8
39	X	X	X	X	X	5	1.0
40	X	X	X	X	X	5	1.0
41	X	X	--	X	X	4	.8
Organization	X	X	X	X	X	5	1.0
Ease of Use	X	X	X	X	X	5	1.0
Completeness	X	X	X	X	X	5	1.0

Table G.1

Ratings from Expert Panel Review 2: Items Rated 3 or 4 on a 4-point Relevance Scale

Item	Expert 1	Expert 2	Expert 3	Expert 4	Number in Agreement	Item CVI
1	X	X	X	X	4	1.0
2	X	X	X	X	4	1.0
3	X	X	X	X	4	1.0
4	X	X	X	X	4	1.0
5	X	X	X	X	4	1.0
6	X	X	X	X	4	1.0
7	X	X	X	X	4	1.0
8	X	X	X	X	4	1.0
9	X	X	X	X	4	1.0
10	X	X	X	X	4	1.0
11	X	X	X	X	4	1.0
12	X	X	X	X	4	1.0
13	X	X	X	X	4	1.0
14	X	X	X	X	4	1.0
15	X	X	X	X	4	1.0
16	X	X	X	X	4	1.0
17	X	X	X	X	4	1.0
18	X	X	X	X	4	1.0
19	X	X	X	X	4	1.0
20	X	X	X	X	4	1.0
21	X	X	X	X	4	1.0
22	X	X	X	X	4	1.0
23	X	X	X	X	4	1.0
24	X	X	X	X	4	1.0
25	X	X	X	X	4	1.0
26	X	X	X	X	4	1.0
27	X	X	X	X	4	1.0
28	X	X	X	X	4	1.0
29	X	X	X	X	4	1.0
30	X	X	X	X	4	1.0
31	X	X	X	X	4	1.0
32	X	X	--	X	4	.75
33	X	X	X	X	4	1.0
34	X	X	X	X	4	1.0
35	X	X	X	X	4	1.0
36	X	X	X	X	4	1.0
37	X	X	X	X	4	1.0
38	X	X	X	X	4	1.0
Organization	X	X	X	X	4	1.0
Ease of Use	X	X	X	X	4	1.0
Completeness	X	X	X	X	4	1.0

APPENDIX H
THEMES FROM EXPERT PANEL REVIEW 1

Notes from Expert Panel 1 – Open Comments

Theme	Comments
Provide Examples	4 – Without some examples, I’m not clear on what this guideline is referring to (#2) 3 – Not certain of the meaning here (#8) 2 – Not sure what you mean by “high level of discrimination” 2 – Give some item examples (i.e. multilogical thinking)
Options Similar	2 – Similar does not mean exact 2 – Similar does not mean exact 5 – Avoid use of the exact same words in every option
Three Options	2 – Use three options instead (#15) 2 – Additional guidelines – three options are acceptable.
Gender Neutral Language	1 – Avoid he/she – if a pronoun makes the reading easier, then state the client’s sex, for example: “a male client” or “an adult female” 2 – Whenever you can – if it makes the question clearer you can use wife, husband, son – just vary it 5 – If inclusion is needed to test content presented within the question, then I have no objection to identification of gender and use of he/she. If not necessary for selection of the correct answer is then extraneous.
Organization	1 – Too long.
Ease of Use	1 – Doubt faculty will use this tool for every question.
Additional Guidelines	2 – Alternative formats 2 – Clinical relevance – why is this important in practice? 2 – Use quotes extensively to make it realistic and have students analyze.
Feedback	4 – Expand to workbook/package for new/experienced faculty members develop high quality test items. 1 – Good job of operationalizing Morrison’s book.

APPENDIX I
EXPERT PANEL DECISION RUBRIC

Decision Rubric

Guideline	Literature (Frequency)	Empirical Support
STEM-2: Eliminate of the following. (0.6)	Morrison et al. (2006) specifies of the following.	Guideline
<i>Intent is extraneous words, unnecessary information</i>	Bosher (2003) Case & Donahue (2008) Haladyna & Downing (1985) Haladyna & Downing (1989a, 1989b); Haladyna et al. (2002) Klisch (1994) Masters et al. (2001) Tarrant et al. (2006) Tarrant & Ware (2008) Vacc et al. (2001) Van Ort & Hazzard (1985)	Review of literature/data - EAL Not research Validated through review of empirical research Comparative review Not research - EAL Review of literature Review of literature Developed from previous study Not research Review of literature – piloted with graduate students
STEM-6: Best answer format – underline, capitalize, and bold key words (BEST , MOST). (0.4)	Bosher (2003) Bosher & Bowles (2008)	Review of literature/data - EAL Research – linguistic modification - EAL
<i>Comments indicated this is inconsistent with standardized exams and NCLEX – suggest using this strategy for negatively phrased terms only (addressed in preceding guideline).</i>		
STEM-8: Avoid conditional expressions (should/would) and passive voice. (0.6)	Bosher (2003) – conditional	Review of literature/data - EAL
<i>Intent is verb tense.</i>	Bosher (2003) McDonald (2007) Morrison et al. (2006)	Review of literature/data – EAL Textbook Review of literature
<i>Comments indicate <u>should</u> is desirable.</i>	McDonald (2007)	Textbook
STRUCTURAL-31: Write items that can be read and comprehended easily on the first reading. (0.6)	Klisch (1994)	Not research - EAL
<i>Intent is understandable, comprehensible, clear – still needs to be read carefully.</i>	Bosher (2003) Brady (2005) Klisch (1994) McCoubrie (2004) Van Ort & Hazzard (1985)	Review of literature/data - EAL Based on Quinn framework Not research – EAL Not research Review of literature – piloted with graduate students
Use gender-neutral language. (3 similar comments)	Anthony (2004) Boland et al. (2010) McDonald (2007) Morrison et al. (2006) Vacc et al. (2001)	Literature review Textbook Review of literature Not research
<i>Unless necessary to test nursing content.</i>		

APPENDIX J

REVISED FAIRNESS OF ITEMS TOOL (FITr)

Fairness of Items Tool (FITr)

Copyright 2014 by Nikole A. Hicks

Evaluate the Stem

1. Use a question format.
2. Eliminate extraneous words (e.g., *of the following*).
3. Present a single, clearly defined question with the problem in the stem.
4. Avoid negatively phrased questions, double negatives, and the use of *except*.
5. Use active verbs and present tense.
6. Write questions at the application or above cognitive level.
7. Write questions that require multilogical thinking (require knowledge of more than one fact/concept).
8. Make sure content is current.
9. Avoid testing student opinions (e.g., use *nurse* instead of *you* as the subject).
10. Test important content and avoid trivia.

Evaluate the Options

11. Make sure options are similar grammatically and in length and amount of detail.
12. Avoid none-of-the-above and all-of-the-above. Use three options instead.
13. Avoid negatively phrased options.
14. Avoid repeating material in the options – move repetitive words to the stem.
15. Avoid repeating words in the stem and correct option.
16. Avoid overlapping options.
17. Eliminate multiple-multiples.
18. Make sure all distracters are plausible.
19. If the stem asks what should be done first or which action is best, all options must be correct with only one option being the first or best.
20. Make sure there is only one correct answer.
21. Write options that require a high level of discrimination to select the correct answer.

Linguistic/Structural Bias

22. Use a parsimonious style and short simple sentences.
23. Use correct grammar, punctuation, capitalization, and spelling.
24. Use precise terms (avoid frequently, appropriate).
25. Avoid absolute terms (always, never, all).
26. Use straight-forward, uncomplicated language. Test nursing content, not vocabulary or reading.
27. Write items that can be comprehended on the first reading. Avoid tricky or misleading items.
28. Ensure that items are independent of each other.
29. Be specific and clear with directions.
30. Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.

Cultural Bias

31. Avoid dominant culture (literature, music, movies, sports, foods) unless essential to safe, effective nursing practice.
32. Eliminate all names.
33. Eliminate all slang.
34. Use terminology from textbook, notes, and common words (home vs. abode).
35. Eliminate humor.
36. Avoid stereotyping and over-representation of cultural groups.
37. Use gender-specific language only when necessary to test nursing content.
38. Present the person first, not the diagnosis.

APPENDIX K
PARTICIPANT ANNOUNCEMENT EMAIL

«Dear First Name, Last Name»

You are invited to participate in a research study to evaluate a tool designed to assist nursing faculty in improving multiple-choice test items.

I am a candidate for the PhD in Nursing Education at the University of Northern Colorado and an Assistant Professor of Clinical Nursing at the University of Cincinnati. I developed the Fairness of Items Tool (FIT) and am conducting a research study to determine if the FIT is a valid tool for nursing faculty to use in identifying bias in multiple-choice questions.

Nurse educators who are currently teaching in a program of nursing and use faculty-generated multiple-choice examinations are eligible to participate. Faculty-generated MC examinations include those that are developed by faculty through writing new test items, using test bank items, revising test items from any source, or any combination of these activities. Participants will complete an online questionnaire that will take approximately 30 minutes. All participants will receive a final copy of the FIT for their personal use following the completion of the research study.

To participate in the study, please click below to complete and submit the interest form. You may also copy and paste this link into your browser: <http://eepurl.com/WWMcj>

I'm interested.

Please note that after you complete the interest form, you will be required to verify your email address before I will receive your information. Please watch for a confirmation email and respond to verify your email address. You will then receive the link to the survey within the next day.

Thank you in advance for your time and dedication to advancing the science of academic nursing education.

Nikole Hicks, PhD(c), RNC, CNE
Email: Nikole.Hicks@bears.unco.edu

APPENDIX L
PARTICIPANT INTEREST FORM

Study Participation Interest - FIT

I am interested in participating in this research study. Please send me a link to the survey.

Email Address

First Name

Last Name

Are you currently actively teaching in nursing?

☐

Yes

☐

No

Have you used faculty-generated MC exams?

☐

Yes

☐

No

Subscribe to list

MailChimp

APPENDIX M
FIT SURVEY – DEMOGRAPHICS

Validating the Fairness of Items Tool (FIT)

Thank you for participating in this dissertation research.

The Fairness of Items Tool (FIT) was developed through an extensive review of literature. The FIT was designed to assist nursing faculty to identify bias in multiple-choice (MC) test items and revise them to improve faculty-developed tests administered in nursing programs. The FIT contains 38 item-writing guidelines organized into four dimensions: bias in the stem, bias in the options, linguistic/structural bias, and cultural bias.

If you would like to view a 15-minute video that provides an overview of the FIT and each of the guidelines, Click [Here](http://www.youtube.com/watch?v=4mP_CYgVOH0) or copy the URL into your web browser: http://www.youtube.com/watch?v=4mP_CYgVOH0

You will be participating in a survey to evaluate whether the FIT is a valid and reliable tool for nursing faculty use. The first section of the survey contains demographic questions to assist the researcher in comparing the study population with the target population of academic nurse educators.

You will be asked to review six (6) MC test questions and evaluate each question by using the FIT. An attachment with a copy of the FIT is included in the survey for you to download and use, since the survey tool did not allow me to use formatting features like bold, italics, and underlining.

Each MC question will be followed by a section of the FIT guidelines: Evaluate the Stem, Evaluate the Options, Linguistic/Structural Bias, and Cultural Bias. In order to make it easier for you to evaluate each section, you will see each sample test item four times – once for each section of the FIT.

To assist you in using the FIT, each of the guidelines have been written in question format. Please evaluate the MC test item and answer Yes/No to each guideline question. You will need to answer all questions on a page to proceed to the next page.

Each sample test item and corresponding sections of the FIT will appear on a separate page, and you will have the ability to go back and view previous sections until you submit the last page, and then you will not be able to back-track.

If you would like to view a 14-minute video demonstrating how to evaluate the questions using the FIT, please Click [Here](https://www.youtube.com/watch?v=JAKmTITMfs4) or copy this URL into your web browser: <https://www.youtube.com/watch?v=JAKmTITMfs4>

When this survey was piloted, the participants wanted me to be sure to tell you to stick with it – the task seems difficult at the beginning, but after the first couple of questions, the pilot group said they enjoyed the activity and learned from it.

Thank you again for agreeing to participate in this dissertation study.

Please respond to the following demographic questions.

Are you currently actively teaching in a program of nursing?

- ☐ Yes
☐ No

Have you ever used faculty-generated multiple-choice examinations for student assessment? (Faculty-generated examinations are developed by faculty through writing new test items, using test bank items, revising test items from any source, or any combination of these activities.)

- ☐ Yes
☐ No

What is your gender?

- ☐ Female
☐ Male

What is your age?

Please indicate your ethnicity:

- ☐ Hispanic or Latino
☐ Not Hispanic or Latino

Please indicate your race (select all that apply):

- ☐ American Indian or Alaska Native
☐ Asian
☐ Black or African American
☐ Native Hawaiian or Other Pacific Islander
☐ White

Please select the state in which you are employed as an academic nurse educator. If more than one state, please select the primary state.

- ☐ Alabama ☐ Alaska ☐ Arizona ☐ Arkansas ☐ California ☐ Colorado ☐ Connecticut
☐ Delaware ☐ Florida ☐ Georgia ☐ Hawaii ☐ Idaho ☐ Illinois ☐ Indiana ☐ Iowa
☐ Kansas ☐ Kentucky ☐ Louisiana ☐ Maine ☐ Maryland ☐ Massachusetts ☐ Michigan
☐ Minnesota ☐ Mississippi ☐ Missouri ☐ Montana ☐ Nebraska ☐ Nevada ☐ New Hampshire
☐ New Jersey ☐ New Mexico ☐ New York ☐ North Carolina ☐ North Dakota ☐ Ohio
☐ Oklahoma ☐ Oregon ☐ Pennsylvania ☐ Rhode Island ☐ South Carolina ☐ South Dakota
☐ Tennessee ☐ Texas ☐ Utah ☐ Vermont ☐ Virginia ☐ Washington ☐ West Virginia
☐ Wisconsin ☐ Wyoming

Please indicate your educational background - select the highest degree completed or in progress:

- ☐ Diploma in nursing
☐ Associate degree in nursing
☐ BSN
☐ Non-nursing bachelor's degree
☐ MSN
☐ Non-nursing master's degree
☐ PhD in nursing
☐ DNP
☐ Other nursing doctoral degree
☐ Non-nursing doctoral degree
☐ Other (please list below)

Please list other degree

In what year did you/will you receive your highest degree?

Please estimate your full-time equivalent years of clinical nursing experience.

What is your clinical specialty?

Please estimate your full-time equivalent years of experience in academic nursing education.

Please indicate your faculty status (select all that apply):

- ☐ Adjunct
- ☐ Visiting
- ☐ Part-time
- ☐ Full-time
- ☐ Administration
- ☐ Staff
- ☐ Tenure track faculty
- ☐ Clinical track faculty
- ☐ Tenured faculty
- ☐ Faculty, non-tenure-granting institution
- ☐ Other (please list below)

Please list other faculty status:

Please indicate your rank:

- ☐ Lecturer
- ☐ Instructor
- ☐ Assistant Professor
- ☐ Associate Professor
- ☐ Professor
- ☐ Other (please list below)

Please list other rank:

Are you a certified academic nurse educator (CNE)?

- ☐ Yes
- ☐ No

In what year were you certified in academic nursing education?

Please reflect upon your own expertise in teaching and test item writing. Indicate how you rate yourself in each category below.

	Novice	Advanced Beginner	Competent	Proficient	Expert
Teaching Expertise	☐	☐	☐	☐	☐
Test Item Writing Expertise	☐	☐	☐	☐	☐

APPENDIX N

SAMPLE FIT SURVEY – COMPREHENSIVE

You will now be reviewing six (6) MC test items and evaluating each using the FIT.

A copy of the FIT is attached here for you to download and use, as the survey tool does not allow formatting such as bold, underlined, and italicized fonts.

Please review each sample test item. The correct answer option is indicated with the * symbol.

Each test item will be followed by a section of the FIT guidelines: Evaluate the Stem, Evaluate the Options, Linguistic/Structural Bias, and Cultural Bias. In order to make it easier for you to evaluate each section, you will see each sample test item four times - once for each section of the FIT.

To assist you in using the FIT, each of the guidelines have been written in question format on the survey. Please evaluate each MC test item and answer Yes/No to each guideline question. You will need to answer all questions on a page to proceed to the next page.

Each sample test item and corresponding sections of the FIT will appear on a separate page, and you will have the ability to go back and view previous sections until you submit the last page, and then you will not be able to back-track.

Please stick with it - the task seems difficult at the beginning, but after the first couple of questions, the pilot group said they enjoyed the activity and learned from it.

[Attachment: "Fairness of Items Tool.pdf"]

Multiple Choice Question 1

Evaluate the Stem

Which of the following would not be a characteristic of an adult who may have potential for abusing children?

- A. The person is challenged by chronic stress.
- B. The person is socially isolated.
- C. The person is in a stable environment with good support.*
- D. The person was treated abusively as a child.

	Yes - it meets the guideline.	No - it violates the guideline.
Is the stem written in a question format?	☐	☐
Does the item present a single, clearly defined question with the problem in the stem?	☐	☐
Is the stem written in the present tense with active forms of the verbs?	☐	☐
Is the stem written at the application or above cognitive level?	☐	☐
Is multi-logical thinking required to answer the test item correctly?	☐	☐
Does the test item contain current content?	☐	☐
Does the item test important content rather than trivial information?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.

Does the stem contain extraneous words (for example, "of the following")?	☐	☐
Does the stem contain a negatively phrased question, a double negative, or use "except"?	☐	☐
Does the item test student opinions?	☐	☐

Evaluate the Options

Which of the following would not be a characteristic of an adult who may have potential for abusing children?

- A. The person is challenged by chronic stress.
- B. The person is socially isolated.
- C. The person is in a stable environment with good support.*
- D. The person was treated abusively as a child.

	Yes - it meets the guideline.	No - it violates the guideline.
Are the options similar grammatically and in length and amount of detail?	☐	☐
Are all distracters plausible?	☐	☐
If the stem asks what should be done first or what action is best, are all options correct with only one option being the first or best?	☐	☐
Is a high level of discrimination required to select the correct answer from the options?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Do the options include none-of-the-above or all-of-the-above?	☐	☐
Are any of the options phrased negatively?	☐	☐
Do the options contain repeating material or repetitive words that should be moved to the stem?	☐	☐
Does the correct option contain a word(s) that is repeated in the stem?	☐	☐
Are any of the options overlapping?	☐	☐
Does the item contain complex multiple-multiples combinations in the options?	☐	☐
Is there more than one correct answer?	☐	☐

Evaluate Linguistic/Structural Bias

Which of the following would not be a characteristic of an adult who may have potential for abusing children?

- A. The person is challenged by chronic stress.
- B. The person is socially isolated.
- C. The person is in a stable environment with good support.*
- D. The person was treated abusively as a child.

	Yes - it meets the guideline.	No - it violates the guideline.
Is the test item written in a parsimonious style with short simple sentences?	☐	☐
Does the test item contain straight-forward, uncomplicated language that tests nursing content, rather than vocabulary or reading?	☐	☐
Is the test item comprehended easily on the first reading without being tricky or misleading?	☐	☐
Is the test item independent of other items?	☐	☐
Are the directions specific and clear?	☐	☐
Is there consistent spacing, question numbering/lettering, and page numbering with all options appearing on the same page as the question?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain any errors in grammar, punctuation, capitalization, and spelling?	☐	☐
Does the test item contain imprecise terms (such as frequently and appropriate)?	☐	☐
Does the test item contain absolute terms (such as always, never, or all)?	☐	☐

Evaluate Cultural Bias

Which of the following would not be a characteristic of an adult who may have potential for abusing children?

- A. The person is challenged by chronic stress.
- B. The person is socially isolated.
- C. The person is in a stable environment with good support.*
- D. The person was treated abusively as a child.

	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain a reference(s) to dominant culture (literature, music, movies, sports, foods) that is NOT essential to safe, effective nursing practice?	☐	☐
Does the test item contain individuals' names?	☐	☐
Does the test item include slang expressions?	☐	☐
Does the test item contain terminology that is NOT commonly used or found in the textbook or notes (for example home vs. abode)?	☐	☐
Does the test item contain humor?	☐	☐
Does the test item stereotype or over-represent cultural groups?	☐	☐
Does the test item contain gender-specific language that is NOT needed to test nursing content?	☐	☐
Does the test item present the diagnosis before the person?	☐	☐

Multiple Choice Question 2

Evaluate the Stem

Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

	Yes - it meets the guideline.	No - it violates the guideline.
Is the stem written in a question format?	☐	☐
Does the item present a single, clearly defined question with the problem in the stem?	☐	☐
Is the stem written in the present tense with active forms of the verbs?	☐	☐
Is the stem written at the application or above cognitive level?	☐	☐
Is multi-logical thinking required to answer the test item correctly?	☐	☐
Does the test item contain current content?	☐	☐
Does the item test important content rather than trivial information?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the stem contain extraneous words (for example, "of the following")?	☐	☐
Does the stem contain a negatively phrased question, a double negative, or use "except"?	☐	☐
Does the item test student opinions?	☐	☐

Evaluate the Options
Evaluate the Options

Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

	Yes - it meets the guideline.	No - it violates the guideline.
Are the options similar grammatically and in length and amount of detail?	☐	☐
Are all distracters plausible?	☐	☐
If the stem asks what should be done first or what action is best, are all options correct with only one option being the first or best?	☐	☐
Is a high level of discrimination required to select the correct answer from the options?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Do the options include none-of-the-above or all-of-the-above?	☐	☐
Are any of the options phrased negatively?	☐	☐
Do the options contain repeating material or repetitive words that should be moved to the stem?	☐	☐
Does the correct option contain a word(s) that is repeated in the stem?	☐	☐
Are any of the options overlapping?	☐	☐
Does the item contain complex multiple-multiple combinations in the options?	☐	☐
Is there more than one correct answer?	☐	☐

Evaluate Linguistic/Structural Bias

Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

	Yes - it meets the guideline.	No - it violates the guideline.
Is the test item written in a parsimonious style with short simple sentences?	☐	☐
Does the test item contain straight-forward, uncomplicated language that tests nursing content, rather than vocabulary or reading?	☐	☐
Is the test item comprehended easily on the first reading without being tricky or misleading?	☐	☐
Is the test item independent of other items?	☐	☐
Are the directions specific and clear?	☐	☐
Is there consistent spacing, question numbering/lettering, and page numbering with all options appearing on the same page as the question?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain any errors in grammar, punctuation, capitalization, and spelling?	☐	☐
Does the test item contain imprecise terms (such as frequently and appropriate)?	☐	☐
Does the test item contain absolute terms (such as always, never, or all)?	☐	☐

Evaluate Cultural Bias

Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain a reference(s) to dominant culture (literature, music, movies, sports, foods) that is NOT essential to safe, effective nursing practice?	☐	☐
Does the test item contain individuals' names?	☐	☐
Does the test item include slang expressions?	☐	☐
Does the test item contain terminology that is NOT commonly used or found in the textbook or notes (for example home vs. abode)?	☐	☐
Does the test item contain humor?	☐	☐
Does the test item stereotype or over-represent cultural groups?	☐	☐
Does the test item contain gender-specific language that is NOT needed to test nursing content?	☐	☐
Does the test item present the diagnosis before the person?	☐	☐

Multiple Choice Question 3

Evaluate the Stem

Which of the following would be the best intervention(s) for persons who may not have oral fluids and are experiencing thirst as a result of intracellular volume depletion?

- A. providing unlimited ice chips
- B. providing ice water mouth rinses*
- C. providing lemon wedges to suck on
- D. all of the above

	Yes - it meets the guideline.	No - it violates the guideline.
Is the stem written in a question format?	☐	☐
Does the item present a single, clearly defined question with the problem in the stem?	☐	☐
Is the stem written in the present tense with active forms of the verbs?	☐	☐
Is the stem written at the application or above cognitive level?	☐	☐
Is multi-logical thinking required to answer the test item correctly?	☐	☐
Does the test item contain current content?	☐	☐
Does the item test important content rather than trivial information?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the stem contain extraneous words (for example, "of the following")?	☐	☐
Does the stem contain a negatively phrased question, a double negative, or use "except"?	☐	☐
Does the item test student opinions?	☐	☐

Evaluate the Options

Which of the following would be the best intervention(s) for persons who may not have oral fluids and are experiencing thirst as a result of intracellular volume depletion?

- A. providing unlimited ice chips
- B. providing ice water mouth rinses*
- C. providing lemon wedges to suck on
- D. all of the above

	Yes - it meets the guideline.	No - it violates the guideline.
Are the options similar grammatically and in length and amount of detail?	☐	☐
Are all distracters plausible?	☐	☐
If the stem asks what should be done first or what action is best, are all options correct with only one option being the first or best?	☐	☐
Is a high level of discrimination required to select the correct answer from the options?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Do the options include none-of-the-above or all-of-the-above?	☐	☐
Are any of the options phrased negatively?	☐	☐
Do the options contain repeating material or repetitive words that should be moved to the stem?	☐	☐
Does the correct option contain a word(s) that is repeated in the stem?	☐	☐
Are any of the options overlapping?	☐	☐
Does the item contain complex multiple-multiples combinations in the options?	☐	☐
Is there more than one correct answer?	☐	☐

Evaluate Linguistic/Structural Bias

Which of the following would be the best intervention(s) for persons who may not have oral fluids and are experiencing thirst as a result of intracellular volume depletion?

- A. providing unlimited ice chips
- B. providing ice water mouth rinses*
- C. providing lemon wedges to suck on
- D. all of the above

	Yes - it meets the guideline.	No - it violates the guideline.
Is the test item written in a parsimonious style with short simple sentences?	☐	☐
Does the test item contain straight-forward, uncomplicated language that tests nursing content, rather than vocabulary or reading?	☐	☐
Is the test item comprehended easily on the first reading without being tricky or misleading?	☐	☐
Is the test item independent of other items?	☐	☐
Are the directions specific and clear?	☐	☐
Is there consistent spacing, question numbering/lettering, and page numbering with all options appearing on the same page as the question?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain any errors in grammar, punctuation, capitalization, and spelling?	☐	☐
Does the test item contain imprecise terms (such as frequently and appropriate)?	☐	☐
Does the test item contain absolute terms (such as always, never, or all)?	☐	☐

Evaluate Cultural Bias

Which of the following would be the best intervention(s) for persons who may not have oral fluids and are experiencing thirst as a result of intracellular volume depletion?

- A. providing unlimited ice chips
- B. providing ice water mouth rinses*
- C. providing lemon wedges to suck on
- D. all of the above

	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain a reference(s) to dominant culture (literature, music, movies, sports, foods) that is NOT essential to safe, effective nursing practice?	☐	☐
Does the test item contain individuals' names?	☐	☐
Does the test item include slang expressions?	☐	☐
Does the test item contain terminology that is NOT commonly used or found in the textbook or notes (for example home vs. abode)?	☐	☐
Does the test item contain humor?	☐	☐
Does the test item stereotype or over-represent cultural groups?	☐	☐
Does the test item contain gender-specific language that is NOT needed to test nursing content?	☐	☐
Does the test item present the diagnosis before the person?	☐	☐

Multiple Choice Question 4

Evaluate the Stem

The nurse administers acetaminophen (Tylenol) 650 mg orally to a client with type 2 diabetes and urosepsis whose temperature is 104 degrees F. One hour later, the client is diaphoretic. Based on these findings, which client assessment is it MOST important for the nurse to obtain?

- A. Temperature.*
- B. Serum glucose.
- C. Pain level.
- D. Blood pressure.

	Yes - it meets the guideline.	No - it violates the guideline.
Is the stem written in a question format?	☐	☐
Does the item present a single, clearly defined question with the problem in the stem?	☐	☐
Is the stem written in the present tense with active forms of the verbs?	☐	☐
Is the stem written at the application or above cognitive level?	☐	☐
Is multi-logical thinking required to answer the test item correctly?	☐	☐
Does the test item contain current content?	☐	☐
Does the item test important content rather than trivial information?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the stem contain a negatively phrased question, a double negative, or use "except"?	☐	☐
Does the item test student opinions?	☐	☐

Evaluate the Options

The nurse administers acetaminophen (Tylenol) 650 mg orally to a client with type 2 diabetes and urosepsis whose temperature is 104 degrees F. One hour later, the client is diaphoretic. Based on these findings, which client assessment is it MOST important for the nurse to obtain?

- A. Temperature.*
- B. Serum glucose.
- C. Pain level.
- D. Blood pressure.

	Yes - it meets the guideline.	No - it violates the guideline.
Are the options similar grammatically and in length and amount of detail?	☐	☐
Are all distracters plausible?	☐	☐
If the stem asks what should be done first or what action is best, are all options correct with only one option being the first or best?	☐	☐
Is a high level of discrimination required to select the correct answer from the options?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Do the options include none-of-the-above or all-of-the-above?	☐	☐
Are any of the options phrased negatively?	☐	☐
Do the options contain repeating material or repetitive words that should be moved to the stem?	☐	☐
Does the correct option contain a word(s) that is repeated in the stem?	☐	☐
Are any of the options overlapping?	☐	☐
Does the item contain complex multiple-multiples combinations in the options?	☐	☐
Is there more than one correct answer?	☐	☐

Evaluate Linguistic/Structural Bias

The nurse administers acetaminophen (Tylenol) 650 mg orally to a client with type 2 diabetes and urosepsis whose temperature is 104 degrees F. One hour later, the client is diaphoretic. Based on these findings, which client assessment is it MOST important for the nurse to obtain?

- A. Temperature.*
- B. Serum glucose.
- C. Pain level.
- D. Blood pressure.

	Yes - it meets the guideline.	No - it violates the guideline.
Is the test item written in a parsimonious style with short simple sentences?	☐	☐
Does the test item contain straight-forward, uncomplicated language that tests nursing content, rather than vocabulary or reading?	☐	☐
Is the test item comprehended easily on the first reading without being tricky or misleading?	☐	☐
Is the test item independent of other items?	☐	☐
Are the directions specific and clear?	☐	☐
Is there consistent spacing, question numbering/lettering, and page numbering with all options appearing on the same page as the question?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain any errors in grammar, punctuation, capitalization, and spelling?	☐	☐
Does the test item contain imprecise terms (such as frequently and appropriate)?	☐	☐
Does the test item contain absolute terms (such as always, never, or all)?	☐	☐

Evaluate Cultural Bias

The nurse administers acetaminophen (Tylenol) 650 mg orally to a client with type 2 diabetes and urosepsis whose temperature is 104 degrees F. One hour later, the client is diaphoretic. Based on these findings, which client assessment is it MOST important for the nurse to obtain?

- A. Temperature.*
- B. Serum glucose.
- C. Pain level.
- D. Blood pressure.

	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain a reference(s) to dominant culture (literature, music, movies, sports, foods) that is NOT essential to safe, effective nursing practice?	☐	☐
Does the test item contain individuals' names?	☐	☐
Does the test item include slang expressions?	☐	☐
Does the test item contain terminology that is NOT commonly used or found in the textbook or notes (for example home vs. abode)?	☐	☐
Does the test item contain humor?	☐	☐
Does the test item stereotype or over-represent cultural groups?	☐	☐
Does the test item contain gender-specific language that is NOT needed to test nursing content?	☐	☐
Does the test item present the diagnosis before the person?	☐	☐

Multiple Choice Question 5

Evaluate the Stem

Which one of the following is the main, overarching goal for Healthy People 2010?

- A. Reduction of health care costs
- B. Elimination of health disparities*
- C. Investigation of substance abuse
- D. Determination of acceptable morbidity rates

	Yes - it meets the guideline.	No - it violates the guideline.
Is the stem written in a question format?	☐	☐
Does the item present a single, clearly defined question with the problem in the stem?	☐	☐
Is the stem written in the present tense with active forms of the verbs?	☐	☐
Is the stem written at the application or above cognitive level?	☐	☐
Is multi-logical thinking required to answer the test item correctly?	☐	☐
Does the test item contain current content?	☐	☐
Does the item test important content rather than trivial information?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the stem contain a negatively phrased question, a double negative, or use "except"?	☐	☐
Does the item test student opinions?	☐	☐

Evaluate the Options

Which one of the following is the main, overarching goal for Healthy People 2010?

- A. Reduction of health care costs
- B. Elimination of health disparities*
- C. Investigation of substance abuse
- D. Determination of acceptable morbidity rates

	Yes - it meets the guideline.	No - it violates the guideline.
Are the options similar grammatically and in length and amount of detail?	☐	☐
Are all distracters plausible?	☐	☐
If the stem asks what should be done first or what action is best, are all options correct with only one option being the first or best?	☐	☐
Is a high level of discrimination required to select the correct answer from the options?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Do the options include none-of-the-above or all-of-the-above?	☐	☐
Are any of the options phrased negatively?	☐	☐
Do the options contain repeating material or repetitive words that should be moved to the stem?	☐	☐
Does the correct option contain a word(s) that is repeated in the stem?	☐	☐
Are any of the options overlapping?	☐	☐
Does the item contain complex multiple-multiples combinations in the options?	☐	☐
Is there more than one correct answer?	☐	☐

Evaluate Linguistic/Structural Bias

Which one of the following is the main, overarching goal for Healthy People 2010?

- A. Reduction of health care costs
- B. Elimination of health disparities*
- C. Investigation of substance abuse
- D. Determination of acceptable morbidity rates

	Yes - it meets the guideline.	No - it violates the guideline.
Is the test item written in a parsimonious style with short simple sentences?	☐	☐
Does the test item contain straight-forward, uncomplicated language that tests nursing content, rather than vocabulary or reading?	☐	☐
Is the test item comprehended easily on the first reading without being tricky or misleading?	☐	☐
Is the test item independent of other items?	☐	☐
Are the directions specific and clear?	☐	☐
Is there consistent spacing, question numbering/lettering, and page numbering with all options appearing on the same page as the question?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain any errors in grammar, punctuation, capitalization, and spelling?	☐	☐
Does the test item contain imprecise terms (such as frequently and appropriate)?	☐	☐
Does the test item contain absolute terms (such as always, never, or all)?	☐	☐

Evaluate Cultural Bias

Which one of the following is the main, overarching goal for Healthy People 2010?

- A. Reduction of health care costs
- B. Elimination of health disparities*
- C. Investigation of substance abuse
- D. Determination of acceptable morbidity rates

	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain a reference(s) to dominant culture (literature, music, movies, sports, foods) that is NOT essential to safe, effective nursing practice?	☐	☐
Does the test item contain individuals' names?	☐	☐
Does the test item include slang expressions?	☐	☐
Does the test item contain terminology that is NOT commonly used or found in the textbook or notes (for example home vs. abode)?	☐	☐
Does the test item contain humor?	☐	☐
Does the test item stereotype or over-represent cultural groups?	☐	☐
Does the test item contain gender-specific language that is NOT needed to test nursing content?	☐	☐
Does the test item present the diagnosis before the person?	☐	☐

Multiple Choice Question 6

Evaluate the Stem

Men should be encouraged to enter nursing primarily because:

- A. They work with physicians better
- B. They have physical strength
- C. They can't get pregnant
- D. They will change the perception of nursing*

	Yes - it meets the guideline.	No - it violates the guideline.
Is the stem written in a question format?	☐	☐
Does the item present a single, clearly defined question with the problem in the stem?	☐	☐
Is the stem written in the present tense with active forms of the verbs?	☐	☐
Is the stem written at the application or above cognitive level?	☐	☐
Is multi-logical thinking required to answer the test item correctly?	☐	☐
Does the test item contain current content?	☐	☐
Does the item test important content rather than trivial information?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the stem contain a negatively phrased question, a double negative, or use "except"?	☐	☐
Does the item test student opinions?	☐	☐

Evaluate the Options

Men should be encouraged to enter nursing primarily because:

- A. They work with physicians better
- B. They have physical strength
- C. They can't get pregnant
- D. They will change the perception of nursing*

	Yes - it meets the guideline.	No - it violates the guideline.
Are the options similar grammatically and in length and amount of detail?	☐	☐
Are all distracters plausible?	☐	☐
If the stem asks what should be done first or what action is best, are all options correct with only one option being the first or best?	☐	☐
Is a high level of discrimination required to select the correct answer from the options?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Do the options include none-of-the-above or all-of-the-above?	☐	☐
Are any of the options phrased negatively?	☐	☐
Do the options contain repeating material or repetitive words that should be moved to the stem?	☐	☐
Does the correct option contain a word(s) that is repeated in the stem?	☐	☐
Are any of the options overlapping?	☐	☐
Does the item contain complex multiple-multiples combinations in the options?	☐	☐
Is there more than one correct answer?	☐	☐

Evaluate Linguistic/Structural Bias

Men should be encouraged to enter nursing primarily because:

- A. They work with physicians better
- B. They have physical strength
- C. They can't get pregnant
- D. They will change the perception of nursing*

	Yes - it meets the guideline.	No - it violates the guideline.
Is the test item written in a parsimonious style with short simple sentences?	☐	☐
Does the test item contain straight-forward, uncomplicated language that tests nursing content, rather than vocabulary or reading?	☐	☐
Is the test item comprehended easily on the first reading without being tricky or misleading?	☐	☐
Is the test item independent of other items?	☐	☐
Are the directions specific and clear?	☐	☐
Is there consistent spacing, question numbering/lettering, and page numbering with all options appearing on the same page as the question?	☐	☐
	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain any errors in grammar, punctuation, capitalization, and spelling?	☐	☐
Does the test item contain imprecise terms (such as frequently and appropriate)?	☐	☐
Does the test item contain absolute terms (such as always, never, or all)?	☐	☐

Evaluate Cultural Bias

Men should be encouraged to enter nursing primarily because:

- A. They work with physicians better
- B. They have physical strength
- C. They can't get pregnant
- D. They will change the perception of nursing*

	Yes - it violates the guideline.	No - it meets the guideline.
Does the test item contain a reference(s) to dominant culture (literature, music, movies, sports, foods) that is NOT essential to safe, effective nursing practice?	☐	☐
Does the test item contain individuals' names?	☐	☐
Does the test item include slang expressions?	☐	☐
Does the test item contain terminology that is NOT commonly used or found in the textbook or notes (for example home vs. abode)?	☐	☐
Does the test item contain humor?	☐	☐
Does the test item stereotype or over-represent cultural groups?	☐	☐
Does the test item contain gender-specific language that is NOT needed to test nursing content?	☐	☐
Does the test item present the diagnosis before the person?	☐	☐

APPENDIX O

MC TEST ITEM SELECTION FOR FIT SURVEYS – STEM, OPTIONS, LINGUISTIC-STRUCTURAL, CULTURAL

Stem Survey

B-1. Which one of the following is the main, overarching goal for Healthy People 2010?

- A. Reduction of health care costs
- B. Elimination of health disparities*
- C. Investigation of substance abuse
- D. Determination of acceptable morbidity rates

B-5. All of the following are correct statements about the American Nurses Association **EXCEPT:**

- A. Is a professional organization whose membership consists of physicians, nurses, and citizens interested in improving health care
- B. Works to improve the quality of nursing practice
- C. Identifies the appropriate academic credentials for entry into nursing practice*
- D. Fosters the development of nursing theory by promoting nursing research

B-12. Mr. Stone is scheduled for lithotripsy. The nurse develops a teaching plan in which the procedure is described as the:

- A. Surgical removal of stones
- B. Capture of stones via scope
- C. Fragmentation of stones by electrical charge*
- D. Dissolution of stones with medication

B-13. Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

B-18. Which of the following would be the best intervention(s) for persons who may not have oral fluids and are experiencing thirst as a result of intracellular volume depletion?

- A. providing unlimited ice chips
- B. providing ice water mouth rinses*
- C. providing lemon wedges to suck on
- D. all of the above

F-10. The nurse administers acetaminophen (Tylenol) 650 mg orally to a client with type 2 diabetes and urosepsis whose temperature is 104 degrees F. One hour later, the client is diaphoretic. Based on these findings, which client assessment is it MOST important for the nurse to obtain?

- A. Temperature.*
- B. Serum glucose.
- C. Pain level.
- D. Blood pressure.

B-27. The nurse is assessing clients in a mental health clinic. Major depression is the greatest risk for a

- A. man who was widowed in the last year.
- B. person who recently moved to this country.
- C. man who retired from the military one month ago.
- D. woman who is unemployed because of poor health.*

B-30. Prior to assisting an elderly client to take a tub bath, the nurse should complete all of the following interventions EXCEPT

- A. Check the bath water temperature.
- B. Close the bathroom door.*
- C. Remind the client to void.
- D. Provide extra towels.

B-35. Men should be encouraged to enter nursing primarily because:

- A. They work with physicians better
- B. They have physical strength
- C. They can't get pregnant
- D. They will change the perception of nursing*

B-36. You are teaching a client who has an obstructed bile duct secondary to cholelithiasis. What changes in bowel movements will you tell the client to expect?

- A. Clay-colored stool with fatty streaks.*
- B. Hard, liquid brown stool with bloody streaks.
- C. Liquid, yellow stool.
- D. Black, tarry stool.

Options Survey

B-2. The care manager role is demonstrated when the nurse:

- A. Helps a diabetic client learn to give her own injection
- B. Meets with the client's family prior to discharge
- C. Organizes and manages a client's plan of care*
- D. Changes a client's wound dressing

B-3. Which description of the breast examination is true:

- A. Postmenopausal women do not need to do a breast exam.
- B. Palpate the breast tissue systematically in a clockwise motion.*
- C. Male breasts are not examined because males can't develop breast cancer.
- D. Percussion is used to further assess any palpable breast masses.

B-10. In which of the following situations should the nurse have a high index of suspicion for water intoxication?

- A. Persons experiencing SIADH
- B. Persons who have experienced head trauma
- C. Persons with a diagnosis of lung cancer
- D. All of the above*

B-11. Which of the following would not be a characteristic of an adult who may have potential for abusing children?

- A. The person is challenged by chronic stress.
- B. The person is socially isolated.
- C. The person is in a stable environment with good support.*
- D. The person was treated abusively as a child.

B-14. A medication order reads: "Digoxin, 0.125 mg PO qod." The nurse correctly gives this drug:

- A. Daily before bedtime
- B. By mouth every other day*
- C. Twice a day by way of the oral route
- D. Once a week after recording an apical rate

B-29. A client diagnosed with congestive heart failure complains of thirst. Which intervention is **MOST** important for the nurse to implement?

- A. Provide small sips of water as needed for thirst.
- B. Remind the client that fluid is being restricted.*
- C. Document the client's hourly intake and output.
- D. Slowly increase the peripheral IV infusion rate.

F-8. The nurse notes that a client does not exhibit the defining characteristics of the priority problem identified in the plan of care. What action does the nurse implement?

- A. Document that the client's defining characteristics are inconsistent with the priority problem.
- B. Change the plan of care to include the problem that is consistent with the client's defining characteristics.*
- C. Revise the plan of care so that the identified problem is a high-risk problem rather than a priority problem.

B-31. May B. High is 36 weeks pregnant. The nurse should conduct further assessment for pregnancy induced hypertension (PIH) based on which finding?

- A. A blood pressure reading of 160/90 with the client in a supine position.*
- B. A client complaint of swelling in the lower extremities.
- C. A systolic blood pressure 30 points higher than the previous reading 4 weeks ago.*
- D. A white blood cell count (WBC) of 15,000 mm³.

B-34. What steps will you implement before starting a blood transfusion on your client?

- 1. Discontinue saline solution and hang dextrose.
  - 2. Identify proper type blood and correct client with another RN.
  - 3. Use a central IV line since a peripheral line cannot be used.
  - 4. Assess vital signs and skin integrity of face, chest, and back.
  - 5. If refrigerated, allow blood to warm for several hours before starting infusion.
- A. 1, 2, 4   B. 2, 3, 4   C. 2, 4, 5   D. 2, 4*

B-36. You are teaching a client who has an obstructed bile duct secondary to cholelithiasis. What changes in bowel movements will you tell the client to expect?

- A. Clay-colored stool with fatty streaks.*
- B. Hard, liquid brown stool with bloody streaks.
- C. Liquid, yellow stool.
- D. Black, tarry stool.

Linguistic/Structural Survey

B-10. In which of the following situations should the nurse have a high index of suspicion for water intoxication?

- A. Persons experiencing SIADH
- B. Persons who have experienced head trauma
- C. Persons with a diagnosis of lung cancer
- D. All of the above*

B-13. Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

B-18. Which of the following would be the best intervention(s) for persons who may not have oral fluids and are experiencing thirst as a result of intracellular volume depletion?

- A. providing unlimited ice chips
- B. providing ice water mouth rinses*
- C. providing lemon wedges to suck on
- D. all of the above

B-20. When an individual experiences metabolic acidosis, which of the following potassium fluctuations would the nurse expect to see initially?

- A. Increased serum potassium levels*
- B. Decreased serum potassium levels
- C. Acidosis has no influence on potassium

B-21. To avoid infection after receiving a puncture wound to the hand, a nurse should:

- A. Always go to the immunization center to receive a tetanus shot.
- B. Be treated with an antibiotic only if the wound is painful.
- C. Ensure that no foreign object has been left in the wound.*
- D. Never wipe the wound with alcohol unless it is still bleeding.

B-22. Severe obesity in early adolescence

- A. usually responds dramatically to dietary regimens
- B. often is related to endocrine disorders
- C. has a 75% chance of clearing spontaneously
- D. shows a poor prognosis*
- E. usually responds to pharmacotherapy and intensive psychotherapy

B-23. Following a second episode of infection, what is the likelihood that a woman is infertile?

- A. Less than 20%
- B. 20 to 30%
- C. Greater than 50%
- D. 90%
- E. 75%

B-25. When instilling a client's eye drops, which technique is used by the nurse?

- 1. Cleanse the eyelid by wiping from inner to outer canthus.
  - 2. Gently compress the outer canthus after the instillation.
  - 3. Hold the medication dropper six inches above the eye.
  - 4. Keep the opposite eye open while instilling the drops.
  - 5. Ask the client to look up while instilling the eye drops.
  - 6. Carefully drop the medication on the client's cornea.
- A. 1, 2, 3, and 6.
  - B. 2, 3, and 5.
  - C. 3 and 6 only.
  - D. 1 and 5 only.*
  - E. All of the above.

B-32. A nurse should recognize that a client who has elevated intracranial pressure will most likely receive which of these medications?

- A. Mannitol (Osmitrol).*
- B. Digoxin (Lanoxin).
- C. Indomethacin (Indocin).
- D. Nadolol (Corgard).

B-33. The nurse should plan to monitor the client for side effects of the medication in the previous question, which include

- A. hyponatremia.*
- B. bradycardia.
- C. hematuria.
- D. agranulocytosis.

Cultural Survey

B-2. The care manager role is demonstrated when the nurse:

- A. Helps a diabetic client learn to give her own injection
- B. Meets with the client's family prior to discharge
- C. Organizes and manages a client's plan of care*
- D. Changes a client's wound dressing

B-10. In which of the following situations should the nurse have a high index of suspicion for water intoxication?

- A. Persons experiencing SIADH
- B. Persons who have experienced head trauma
- C. Persons with a diagnosis of lung cancer
- D. All of the above*

B-12. Mr. Stone is scheduled for lithotripsy. The nurse develops a teaching plan in which the procedure is described as the:

- A. Surgical removal of stones
- B. Capture of stones via scope
- C. Fragmentation of stones by electrical charge*
- D. Dissolution of stones with medication

B-13. Before her patient goes to surgery, the nurse obtains and records the patient's vital signs. This is important because it provides:

- A. Routine information needed from all hospitalized patients
- B. Information the doctor will use when deciding where to place the patient after completion of the surgical procedure
- C. A time for the nurse to get acquainted with the patient before he/she goes to surgery
- D. Baseline data for comparison during and after surgery*

B-15. A six-year-old is scheduled for surgery to repair a ventricular septal defect. The child is placed on a low sodium diet. The nurse teaches the mother that the menu containing the lowest sodium content is:

- A. Hot dog and baked beans
- B. Beef patty and baked potato*
- C. Tomato soup and tossed salad
- D. Bologna sandwich and French fries

F-8. The nurse notes that a client does not exhibit the defining characteristics of the priority problem identified in the plan of care. What action does the nurse implement?

- A. Document that the client's defining characteristics are inconsistent with the priority problem.
- B. Change the plan of care to include the problem that is consistent with the client's defining characteristics.*
- C. Revise the plan of care so that the identified problem is a high-risk problem rather than a priority problem.

B-16. When Sotheby's auctioned off items from the Jackie Kennedy Onassis estate, those who paid "top dollar" for items were most likely using the behavioral mechanism of:

- A. Projection
- B. Identification *
- C. Rationalization
- D. Reaction formation

B-28. The nurse administers acetaminophen (Tylenol) 650 mg orally to a diabetic client with urosepsis whose temperature is 104°F. One hour later, the client is diaphoretic. Based on these findings, which intervention should the nurse implement?

- A. Assess the client's temperature.*
- B. Assess the client's serum glucose.
- C. Assess the client's pain level.
- D. Assess the client's blood pressure.

B-31. May B. High is 36 weeks pregnant. The nurse should conduct further assessment for pregnancy induced hypertension (PIH) based on which finding?

- A. A blood pressure reading of 160/90 with the client in a supine position.*
- B. A client complaint of swelling in the lower extremities.
- C. A systolic blood pressure 30 points higher than the previous reading 4 weeks ago.*
- D. A white blood cell count (WBC) of 15,000 mm³.

B-35. Men should be encouraged to enter nursing primarily because:

- A. They work with physicians better
- B. They have physical strength
- C. They can't get pregnant
- D. They will change the perception of nursing*

APPENDIX P

PARTICIPANT INVITATION EMAIL

Dear Colleague,

Thank you for your interest in participating in my dissertation research study to evaluate the Fairness of Items Tool (FIT) for its use in identifying bias in multiple-choice (MC) questions.

Nurse educators who are currently teaching in a program of nursing and who use faculty-generated MC examinations are eligible to participate in this research study. Faculty-generated MC examinations include those that are developed by faculty through writing new test items, using test bank items, revising test items from any source, or any combination of these activities. If you agree to participate in this research study, you will be asked to complete a web-based survey. I anticipate that completion of the survey will take approximately 30 minutes.

The first part of the survey contains questions about your nursing and teaching experience and expertise in developing test items. In the second part of the survey, you will be presented with sample MC questions followed by guidelines from the FIT. You will be asked to review the test item and indicate if the item violates one or more of the guidelines. Your obligation is then concluded.

I foresee no risks to participants beyond those normally encountered in web-based surveys, including the risk of technical difficulties and the time you will need to spend in completing survey. As a token of my appreciation for your time and effort in participating in this research study, I would like to send you a copy of the FIT following the completion of this research study.

Names and identifying information will not appear in any professional report of this research. Participation is voluntary. You may decide not to participate in this study, and if you begin participation, you may decide to stop and withdraw at any time. Your decision will be respected and will not result in loss of benefits to which you are otherwise entitled. Having read the above and having had the opportunity to ask any questions, please click the link below if you would like to participate in this research, and you will be directed to the web-based survey. By completing the web-based survey, you will grant permission for your participation. You may keep this form for future reference.

Please feel free to phone or email me if you have any questions or concerns about this research. If you have any concerns about your selection or treatment as a research participant, please contact the Sponsored Programs and Academic Research Center, Kepner Hall, University of Northern Colorado, Greeley, CO 80639; 970-351-1907.

I appreciate your participation in this research study and your support of the nursing profession.

Sincerely,
Nikole Hicks, PhD(c), RNC, CNE

Project Title: Establishing Validity and Reliability of the Fairness of Items Tool (FIT)
Researcher: Nikole Hicks, PhD Candidate, School of Nursing
Research Advisor: Janice Hayes, PhD, School of Nursing
Phone: (606) 759-6677
Email: Nikole.Hicks@uc.edu

You may open the survey in your web browser by clicking the link below:
[Validating the Fairness of Items Tool \(FIT\) b](#)

If the link above does not work, try copying the link below into your web browser:
<https://redcap.research.cchmc.org/surveys/?s=XX00xxXXXx>

This link is unique to you and should not be forwarded to others.

APPENDIX Q
INSTITUTIONAL REVIEW BOARD
LETTERS OF APPROVAL



Institutional Review Board

DATE: December 12, 2013

TO: Nikole Hicks

FROM: University of Northern Colorado (UNCO) IRB

PROJECT TITLE: [532951-1] Establishing the Validity and Reliability of the Fairness of Items Tool (FIT)

SUBMISSION TYPE: New Project

ACTION: APPROVAL/VERIFICATION OF EXEMPT STATUS

DECISION DATE: December 10, 2013

Thank you for your submission of New Project materials for this project. The University of Northern Colorado (UNCO) IRB approves this project and verifies its status as EXEMPT according to federal IRB regulations.

Nikole -

Hello and thank you for a very clear and thorough IRB application.

Best wishes with your research and don't hesitate to contact me with any IRB-related questions or concerns.

Sincerely,

Dr. Megan Stellino, UNC IRB Co-Chair

We will retain a copy of this correspondence within our records for a duration of 4 years.

If you have any questions, please contact Sherry May at 970-351-1910 or Sherry.May@unco.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Northern Colorado (UNCO) IRB's records.

Institutional Review Board - Federalwide Assurance #00003152**University of Cincinnati**

Date: 3/5/2014
 From: UC IRB
 To: Principal Investigator: Nikole Hicks
 CON AD for Academic Nursing
 Re: Study ID: 2013-8080
 Study Title: Establishing the Validity and Reliability of the Fairness of Items Tool (FIT)

The Institutional Review Board (IRB) received the above referenced proposal. After review of the materials submitted, the UC IRB has determined that this research is **EXEMPT** from IRB review in accordance with **45 CFR 46.101 (b)** (see below) on 3/3/2014. Ongoing IRB oversight is not required. This study has also been review through the University of Northern Colorado IRB. Due to the Exempt status determined at UNCO and that IRB oversight is not required UNCO declined to sign the Authorization Agreement form.

Please note the following requirements:

AMENDMENTS: The principal investigator is responsible for notifying the IRB of any changes in the protocol, participating investigators, procedures, recruitment, consent forms, FDA status, or conflicts of interest. Approval is based on the information as submitted. New procedures cannot be initiated until IRB approval has been given. If you wish to change any aspect of this study, please submit an Amendment via ePAS to the IRB, providing a justification for each requested change.

UNANTICIPATED PROBLEMS: The investigator is responsible for reporting **unanticipated problems** promptly to the IRB via ePAS according to current reporting policies.

Please note: This approval is through the IRB only. You may be responsible for reporting to other regulatory officials (e.g. VA Research and Development Office, UC Health – University Hospital). Please check with your institution and department to ensure you have met all reporting requirements.

Statement regarding International conference on Harmonization and Good clinical Practices.

The Institutional Review Board is duly constituted (fulfilling FDA requirements for diversity), has written procedures for initial and continuing review of clinical trials: prepares written minutes of convened meetings and retains records pertaining to the review and approval process; all in compliance with requirements defined in 21 CFR Parts 50, 56 and 312 Code of Federal Regulations. This institution is in compliance with the ICH GCP as adopted by FDA/DHHS.

Thank you for your cooperation during the review process.

---

§46.101 (b) (2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless (i)



Institutional Review Board

DATE: June 5, 2014

TO: Nikole Hicks

FROM: University of Northern Colorado (UNCO) IRB

PROJECT TITLE: [532951-3] Establishing the Validity and Reliability of the Fairness of Items Tool (FIT)

SUBMISSION TYPE: Amendment/Modification

ACTION: APPROVAL/VERIFICATION OF EXEMPT STATUS

DECISION DATE: June 4, 2014

Thank you for your submission of Amendment/Modification materials for this project. The University of Northern Colorado (UNCO) IRB approves this project and verifies its status as EXEMPT according to federal IRB regulations.

Nikole -

Thank you for a clear explanation of the modifications. These changes are approved exempt. Please be sure to utilize the described revisions to your participant recruitment and newly developed forms as you go forward with this data collection.

Don't hesitate to contact me with any IRB-related questions or concerns.

Sincerely,

Dr. Megan Stellino, UNC IRB Co-Chair

We will retain a copy of this correspondence within our records for a duration of 4 years.

If you have any questions, please contact Sherry May at 970-351-1910 or Sherry.May@unco.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Northern Colorado (UNCO) IRB's records.

**Institutional Review Board - Federalwide Assurance #00003152
University of Cincinnati**

Date: 7/15/2014
 From: UC IRB
 To: Principal Investigator: Nikole Hicks
 CON AD for Academic Nursing
 Re: Study ID: 2013-8080
 Study Title: Establishing the Validity and Reliability of the Fairness of Items Tool (FIT)

The Institutional Review Board (IRB) received an amendment to the above referenced proposal.

Revised Study Documents

NE Surveys.pdf
 Original CNE Surveys
 Original Invitation Announcement
 Revised Cover Letter and Informed Consent.docx
 Revised Invitation Announcement - Track Changes.docx
 Revised Invitation Announcement.docx
 Revised IRB Application Narrative.docx
 Revised NE Surveys - Track Changes.pdf

Amendment Documents:

Modification Participant Recruitment - UC.docx
 IRB Amendment Approval UNCO.pdf

IRB Review Date: 7/15/2014

It was determined that this research remains **EXEMPT** from IRB Review in accordance with **45 CFR 46.101 (b)**.

Please note the following: Ongoing IRB oversight is not required.

AMENDMENTS: The principal investigator is responsible for notifying the IRB of any changes in the proposal. Approval is based on the information as submitted. New procedures cannot be initiated until IRB approval has been given. If you wish to change any aspect of this study, please submit an Amendment via ePAS to the IRB, providing a justification for each requested change.

Statement regarding International conference on Harmonization and Good Clinical Practices:
 The Institutional Review Board is duly constituted (fulfilling FDA requirements for diversity), has written procedures for initial and continuing review of clinical trials; prepares written minutes of convened meetings, and retains records pertaining to the review and approval process; all in compliance with requirements defined in 21 CFR Parts 50, 56 and 312 Code of Federal Regulations. This institution is in compliance with the ICH GCP as adopted by FDA/DHHS.

Thank you for your cooperation during the review process.

APPENDIX R
RATIONALE FOR REVISIONS TO
FAIRNESS OF ITEMS TOOL

Fairness of Items Tool (FIT)

Evaluate the Stem

1. Use a question format.
2. ~~Eliminate extraneous words (e.g., of the following).~~
3. Present a single, clearly defined question with the problem in the stem.
4. ~~Avoid absolute terms (always, never, all).~~
5. Avoid negatively phrased questions, double negatives, and the use of *except*.
6. ~~Best answer format – underline, capitalize, and bold key words (BEST, MOSTY).~~
7. ~~Avoid trick questions.~~
8. ~~Avoid conditional expressions (should, would) and passive voice.~~ Use active verbs and present tense.
9. Write questions at the application or above cognitive level.
10. Write questions that require multi-logical thinking (require knowledge of more than one fact/concept).
11. Make sure content is current.
12. Avoid testing student opinions (e.g., use nurse instead of you as the subject).
13. Test important content and avoid trivia.

Evaluate the Options

14. Make sure options are similar grammatically and in length and amount of detail.
15. ~~Make sure options are grammatically and visually similar.~~
16. Avoid none-of-the-above and all-of-the-above. Use three options instead.
17. Avoid negatively phrased options.
18. Avoid repeating material in the options – move repetitive words to the stem.
19. Avoid repeating words in the stem and correct option.
20. ~~Make sure there is one, and only one, correct answer.~~ Avoid overlapping options.
21. Eliminate multiple-multiples.
22. Make sure all distracters are plausible.
23. If the stem asks what should be done first or which action is best, all options must be correct with only one option being the first or best.
24. ~~Make sure there is only one, and only one, correct answer.~~ Avoid overlapping options.
25. Write options that require a high level of discrimination to select the correct answer.

Linguistic/Structural Bias

26. Use a parsimonious style and short simple sentences.
27. ~~Use correct grammar, punctuation, capitalization, and spelling.~~
28. Use precise terms (avoid frequently, appropriate).
29. ~~Avoid absolute terms (always, never, all).~~
30. Use straight-forward, uncomplicated language. Test nursing content, not vocabulary or reading.
31. ~~Write items that can be comprehended on the first reading. Avoid tricky or misleading items.~~
32. Ensure that items are independent of each other.
33. ~~Be specific and clear with directions.~~
34. ~~Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.~~

Structural Bias

35. ~~Use correct grammar, punctuation, capitalization, and spelling.~~
36. ~~Write items that can be read and comprehended easily on the first reading.~~
37. ~~Be specific and clear with directions.~~
38. ~~Use consistent spacing, question numbering/lettering, page numbering. Make sure options appear on the same page as the question.~~

Cultural Bias

39. Avoid dominant culture (literature, music, movies, sports, foods) unless essential to safe, effective nursing practice.
40. Eliminate all names.
41. Eliminate all slang.
42. Use terminology from textbook, notes, and common words (toilet vs. commode; home vs. abode).
43. Eliminate humor.
44. Avoid stereotyping and over-representation of cultural groups.
45. Use gender-neutral specific language only when necessary to test nursing content.
46. Present the person first, not the diagnosis.

Comment [NH1]: I-CVI 0.6

Feedback indicated this guideline was unclear. Extraneous words was added to clarify (as suggested) with the following as an example.

Comment [NH2]: I-CVI 0.4

Feedback indicated that this guideline was inconsistent with standardized tests and NCLEX and may create confusion and inconsistency when used in other ways. Literature analysis revealed that this guideline appeared in literature by Boshier and Boshier & Bowles related to ELL population. Other experts recommend bolding and capitalizing only if negative expressions are used, which is addressed in the previous guideline. This guideline was eliminated for lack of support.

Comment [NH3]: This guideline was moved to structural bias for clarity – see comment.

Comment [NH4]: I-CVI 0.6

Feedback supported using active verbs and present tense – comments related to use of should vs. would. Literature analysis revealed that use of should as a conditional expression was mentioned by Boshier and Boshier & Bowles. Other nursing literature supported using should and not would. Guideline revised to address verb tense.

Comment [NH5]: Expert comment recommended clarifying this term.

Comment [NH6]: This phrase was added based on feedback to provide an example of a way to avoid testing student opinion.

Comment [NH7]: These two guidelines overlapped somewhat so were combined for simplicity.

Comment [NH8]: One expert recommended adding a guideline about number of options. Empirical literature suggests that 3-4 options is ideal. Three options if preferable to creating a fourth option that is not good quality. However, standardized tests and NCLEX still use four options. Rather than creating a new guideline, it was decided to add the suggestion of three options in this guideline. Number of options will be revisited if changes are made in practice.

Comment [NH9]: This guideline was moved for clarity – it logically follows the previous guideline.

Comment [NH10]: Categories combined because there are overlapping guidelines (feedback from pilot). Guidelines reordered to flow logically.

Comment [nah11]: Rationale: Absolute terms need to be avoided in both the stem and options. While it is a testwise cue, use of absolute terms also contributes to linguistic complexity.

Comment [NH12]: I-CVI 0.6 Reworded for clarity. Avoid trick questions combined with this guideline for clarity and to convey the intent.

Comment [NH13]: I-CVI 0.6 Reworded for clarity. Avoid trick questions combined with this guideline for clarity and to convey the intent.

Comment [NH14]: Example changed based on feedback/suggestion.

Comment [NH15]: Feedback advised revision – gender may be necessary to test some nursing concepts (OB, etc.), but should not be used unnecessarily.

Comments Enlarged

Comment [NH1]: I-CVI 0.6

Feedback indicated this guideline was unclear. Extraneous words was added to clarify (as suggested) with of the following as an example.

Comment [NH2]: I-CVI 0.4

Feedback indicated that this guideline was inconsistent with standardized tests and NCLEX and may create confusion and inconsistency when used in other ways.

Literature analysis revealed that this guideline appeared in literature by Boshier and Boshier & Bowles related to ELL population. Other experts recommend bolding and capitalizing only if negative expressions are used, which is addressed in the previous guideline.

This guideline was eliminated for lack of support.

Comment [NH3]: This guideline was moved to structural bias for clarity – see comment.

Comment [NH4]: I-CVI 0.6

Feedback supported using active verbs and present tense – comments related to use of should vs. would.

Literature analysis revealed that use of should as a conditional expression was mentioned by Boshier and Boshier & Bowles. Other nursing literature supported using should and not would. Guideline revised to address verb tense.

Comment [NH5]: Expert comment recommended clarifying this term.

Comment [NH6]: This phrase was added based on feedback to provide an example of a way to avoid testing student opinion.

Comment [NH7]: These two guidelines overlapped somewhat so were combined for simplicity.

Comment [NH8]: One expert recommended adding a guideline about number of options. Empirical literature suggests that 3-4 options is ideal. Three options is preferable to creating a fourth option that is not good quality. However, standardized tests and NCLEX still use four options. Rather than creating a new guideline, it was decided to add the suggestion of three options in this guideline. Number of options will be revisited if changes are made in practice.

Comment [NH9]: This guideline was moved for clarity – it logically follows the previous guideline.

Comment [NH10]: Categories combined because there are overlapping guidelines (feedback from pilot). Guidelines reordered to flow logically.

Comment [nah11]: Rationale: Absolute terms need to be avoided in both the stem and options. While it is a testwise cue, use of absolute terms also contributes to linguistic complexity.

Comment [NH12]: I-CVI 0.6

Reworded for clarity. Avoid trick questions combined with this guideline for clarity and to convey the intent.

Comment [NH13]: I-CVI 0.6

Reworded for clarity. Avoid trick questions combined with this guideline for clarity and to convey the intent.

Comment [NH14]: Example changed based on feedback/suggestion.

Comment [NH15]: Feedback advised revision – gender may be necessary to test some nursing concepts (OB, etc.) but should not be used unnecessarily.

APPENDIX S
DEMOGRAPHIC CHARACTERISTICS

Table S.1

General Demographic Characteristics of Sample Population

		Survey					
		% (n)					
Characteristic		COMP	S	O	L-S	C	Total
Gender							
	Male	6.8% (7)	7.4% (7)	10.6% (11)	3.3% (3)	8.2% (8)	7.4% (36)
	Female	93.2% (96)	92.6% (87)	89.4% (93)	96.7% (93)	91.8% (89)	92.6% (452)
Age							
	≤ 30	0	1.1% (1)	1% (1)	0	3.1% (3)	1% (5)
	31-45	20.4% (21)	19.1% (18)	21.2% (22)	24.4% (22)	19.8% (19)	20.9% (102)
	46-60	58.3% (60)	60.6% (57)	55.8% (58)	52.2% (47)	50% (48)	55.4% (270)
	Over 60	21.4% (22)	19.1% (18)	22.1% (23)	23.3% (21)	27.1% (26)	22.6% (110)
Ethnicity							
	Not Hispanic	99% (102)	100% (94)	97.1% (101)	98.9% (89)	94.8% (92)	98% (478)
	Hisp. /Latino	1% (1)	0	2.9% (3)	1.1% (1)	5.2% (5)	2% (10)
Race							
	White	95.1% (98)	94.7% (89)	97.1% (101)	96.7% (87)	92.8% (90)	95.3% (465)
	Black	3.9% (4)	5.3% (5)	1.9% (2)	2.2% (2)	6.2% (6)	3.9% (19)
	Am. Indian	1% (1)	0	1% (1)	1.1% (1)	1% (1)	0.8% (4)
Other Race							
	Black	0	1.1% (1)	0	0	0	0.2% (1)
	Am. /Native	1% (1)	2.2% (2)	1% (1)	0	1% (1)	1% (5)
	Asian	0	1.1% (1)	1% (1)	0	0	0.4% (2)
	Hawaii/Pac.	0	0	0	0	0	0
Region							
	Northeast	12.6% (13)	11.7% (11)	16.3% (17)	12.2% (11)	12.4% (12)	13.1% (64)
	Midwest	31.1% (32)	34% (32)	30.8% (32)	36.7% (33)	28.9% (28)	32.2% (157)
	South	25.2% (26)	28.7% (27)	26% (27)	20% (18)	26.8% (26)	25.4% (124)
	Southwest	16.5% (17)	6.4% (6)	16.3% (17)	14.4% (13)	11.3% (11)	13.1% (64)
	West	14.6% (15)	19.1% (18)	10.6% (11)	16.7% (15)	20.6% (20)	16.2% (79)

Note. (COMP) = Comprehensive; (S) = Stem; (O) = Options; (L-S) = Linguistic-Structural; (C) = Cultural; Hisp. = Hispanic; Am. /Native = American Indian/Alaska Native; Hawaii/Pac. = Hawaiian Native/ Pacific Islander.

Table S.2

Demographic Characteristics of Sample – Education and Experience

Characteristic	Survey % (n)					Total
	COMP	S	O	L-S	C	
Highest Degree						
Master's	34% (35)	40.4% (38)	32.7% (34)	33.3% (30)	37.1% (36)	35.5% (173)
Doctorate	66% (68)	59.6% (56)	67.3% (70)	66.7% (60)	62.9% (61)	64.5% (315)
Degree Year						
Before 1980	1% (1)	1.1% (1)	1% (1)	1.1% (1)	3.1% (3)	1.4% (7)
1980-1999	24.3% (25)	18.1% (17)	21.2% (22)	20% (18)	12.4% (12)	19.3% (94)
After 1999	74.8% (77)	80.9% (76)	77.9% (81)	78.9% (71)	84.5% (82)	79.3% (387)
Clinical FTE						
< 10 Years	11.7% (12)	12.8% (12)	9.6% (10)	13.3% (12)	12.4% (12)	11.9% (58)
10-19 Years	29.1% (30)	22.3% (21)	33.7% (35)	35.6% (32)	19.6% (19)	28.1% (137)
20+ Years	59.2% (61)	64.9% (61)	56.7% (59)	51.1% (46)	68% (66)	60% (293)
Specialty						
Critical Care	27.2% (28)	21.3% (20)	14.4% (15)	21.1% (19)	23.7% (23)	21.5% (105)
Med-Surg	28.2% (29)	26.6% (25)	31.7% (33)	28.9% (26)	21.6% (21)	27.5% (134)
Mental Health	4.9% (5)	9.6% (9)	5.8% (6)	6.7% (6)	9.3% (9)	7.2% (35)
Community	6.8% (7)	6.4% (6)	16.3% (17)*	4.4% (4)	7.2% (7)	8.4% (41)
Family	27.2% (28)	25.5% (24)	26% (27)	26% (27)	26.8% (26)	26.4% (129)
Gerontology	1.9% (2)	5.3% (5)	2.9% (3)	3.3% (3)	6.2% (6)	3.9% (19)
Other	3.9% (4)	5.3% (5)	2.9% (3)	8.9% (8)	5.2% (5)	5.1% (25)
Academic FTE						
< 10 Years	48.5% (50)	51.1% (48)	51% (53)	50% (45)	52.6% (51)	50.6% (247)
10-19 Years	26.2% (27)	30.9% (29)	28.8% (30)	26.7% (24)	27.8% (27)	28.1% (137)
20+ Years	25.2% (26)	18.1% (17)	20.2% (21)	23.3% (21)	19.6% (19)	21.3% (104)

Note. (COMP) = Comprehensive; (S) = Stem; (O) = Options; (L-S) = Linguistic-Structural; (C) = Cultural; FTE = Full-Time Equivalent; Critical Care includes Emergency, Perioperative, and Anesthesia; Med-Surg = Adult, Medical-Surgical, Oncology; Community includes Public Health, Home Care, and Palliative Care; Family = Women's Health, Family Health, Obstetrics, Midwifery, Maternal-Child, Pediatrics.

* $p < .05$

Table S.3

Demographic Characteristics of Sample – Faculty Status and Expertise

Characteristic	Survey % (n)					Total
	COMP	S	O	L-S	C	
Faculty Status						
Staff/Admin.	1.9% (2)	2.1% (2)	1% (1)	2.2% (2)	1% (1)	1.6% (8)
Faculty	98.1% (101)	97.9% (92)	99% (104)	97.8% (88)	99% (96)	98.4% (481)
FT/PT Status						
FT	94.9% (93)	94.3% (83)	99% (100)	95% (76)	94.8% (92)	95.7% (444)
PT	5.1% (5)	5.7% (5)	1% (1)	5% (4)	5.2% (5)	4.3% (20)
Tenure Status						
Not Tenured	54.9% (50)	61.6% (45)	64.4% (67)	61% (47)	54.5% (42)	59.5% (251)
Tenured/Track	45.1% (41)	38.4% (28)	35.6% (37)	39% (30)	45.5% (35)	40.5% (171)
Rank						
Lecturer	1.9% (2)	6.4% (6)	1.9% (2)	5.6% (5)	3.1% (3)	3.7% (18)
Instructor	21.4% (22)	19.1% (18)	16.3% (17)	21.1% (19)	13.4% (13)	18.2% (89)
Assistant Prof.	46.6% (48)	42.6% (40)	51.9% (54)	37.8% (34)	57.7% (56)	47.5% (232)
Associate Prof.	18.4% (19)	24.5% (23)	21.2% (22)	24.4% (22)	21.6% (21)	21.9% (107)
Professor	10.7% (11)	7.4% (7)	7.7% (8)	10% (9)	4.1% (4)	8% (39)
Other	1% (1)	0	1% (1)	1.1% (1)	0	0.6% (3)
CNE	20.4% (21)	13.8% (13)	19.2% (20)	22.2% (20)	24% (23)	19.9% (97)
Expertise – Teaching						
Novice	1% (1)	1.1% (1)	5.8% (6)	1.1% (1)	1.1% (1)	2.1% (10)
Adv. Beginner	7.9% (8)	9.6% (9)	4.9% (5)	9% (8)	8.4% (8)	7.9% (38)
Competent	19.8% (20)	30.9% (29)	22.3% (23)	20.2% (18)	22.1% (21)	23% (111)
Proficient	43.6% (44)	31.9% (30)	44.7% (46)	42.7% (38)	42.1% (40)	41.1% (198)
Expert	27.7% (28)	26.6% (25)	22.3% (23)	27% (24)	26.3% (25)	25.9% (125)
Expertise – Item Writing						
Novice	7.9% (8)	8.5% (8)	9.7% (10)	10.6% (10)	6.7% (6)	8.6% (42)
Adv. Beginner	12.9% (13)	26.6% (25)	20.4% (21)	16% (15)	16.9% (15)	18.2% (89)
Competent	34.7% (35)	36.2% (34)	35.9% (37)	31.9% (30)	51.7% (46)	37.2% (182)
Proficient	42.6% (43)	24.5% (23)	30.1% (31)	34% (32)	19.1% (17)	29.9% (146)
Expert	2% (2)	4.3% (4)	3.9% (4)	7.4% (7)	5.6% (5)	4.5% (22)
Total	103	94	104	90	97	488

Note. (COMP) = Comprehensive; (S) = Stem; (O) = Options; (L-S) = Linguistic-Structural; (C) = Cultural; FT = Full-Time; PT = Part Time; CNE= Certified Nurse Educator.

APPENDIX T
DESCRIPTIVE STATISTICS

Table T.1

Descriptive Statistics – Test Item B-1

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-1	ES1	66 _(COMP)	1	65	ES1	83 _(STEM)	4	79				
	ES3		6	60	ES3		9	74				
	ES5		2	64	ES5		8	75				
	ES6		46	20	ES6		61	22				
	ES7		49	17	ES7		60	23				
	ES8		32	34	ES8		47	36				
	ES10		9	57	ES10		16	67				
	ES2		55	11	ES2		63	20				
	ES4		1	65	ES4		0	83				
	ES9		10	56	ES9		6	77				
	EO11	66 _(COMP)	3	63	LS22	66 _(COMP)	9	57	C31	66 _(COMP)	1	65
	EO18		15	51	LS26		18	48	C32		0	66
	EO19		9	57	LS27		10	56	C33		2	64
	EO21		33	33	LS28		4	62	C34		7	59
	EO12		1	65	LS29		6	60	C35		1	65
	EO13		0	66	LS30		1	65	C36		3	63
	EO14		0	66	LS23		7	59	C37		0	66
	EO15		3	63	LS24		11	55	C38		4	62
	EO16		5	61	LS25		1	65				
	EO17		0	66								
	EO20		9	57								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem.

Table T.2

Descriptive Statistics – Test Items B-2, B-3

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-2	EO11	85 _(O)	13	72	C31	87 _(C)	3	84				
	EO18		6	79	C32		0	87				
	EO19		13	72	C33		0	87				
	EO21		25	60	C34		2	85				
	EO12		1	84	C35		0	87				
	EO13		2	83	C36		5	82				
	EO14		9	76	C37		38	49				
	EO15 [^]		42	43	C38		58	29				
	EO16		9	76								
	EO17		3	82								
	EO20		33	52								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-3	ES1	79 _(STEM)	72	7	EO11	83 _(O)	36	47				
	ES3		38	41	EO18		34	49				
	ES5		9	70	EO19		20	63				
	ES6		54	25	EO21		45	38				
	ES7		62	17	EO12		1	82				
	ES8		5	74	EO13 [^]		59	24				
	ES10		11	68	EO14		8	75				
	ES2		19	60	EO15		19	64				
	ES4		0	79	EO16		5	78				
	ES9		0	79	EO17		1	82				
					EO20		3	80				

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem; (O) = Options; (C) = Cultural; [^] = indicates a guideline that contains known bias for this test item.

Table T.3

Descriptive Statistics – Test Items B-5, B-10

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-5	ES1	80 _(STEM)	71	9								
	ES3		53	27								
	ES5		12	68								
	ES6		62	18								
	ES7		52	28								
	ES8		3	77								
	ES10		21	59								
	ES2		61	19								
	ES4 [^]		75	5								
	ES9		4	76								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-10	EO11	83 _(O)	38	45	LS22	79 _(L-S)	30	49	C31	87 _(C)	1	86
	EO18		14	69	LS26		39	40	C32		0	87
	EO19		27	56	LS27		26	53	C33		8	79
	EO21		30	53	LS28		9	70	C34 [^]		23	64
	EO12 [^]		80	3	LS29		9	70	C35		0	87
	EO13		1	82	LS30		4	75	C36		0	87
	EO14		31	52	LS23 [^]		11	68	C37		0	87
	EO15		1	82	LS24		18	61	C38		4	83
	EO16		8	75	LS25		21	58				
	EO17		1	82								
	EO20		44	39								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem; (O) = Options; (L-S) = Linguistic-Structural; (C) = Cultural; [^] = indicates a guideline that contains known bias for this test item.

Table T.4

Descriptive Statistics – Test Item B-11

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-11	ES1	75 _(COMP)	12	64	LS22	75 _(COMP)	42	33	C31	75 _(COMP)	5	70
	ES3		26	50	LS26		30	45	C32		0	75
	ES5 [^]		33	43	LS27		53	22	C33		4	71
	ES6		44	32	LS28		6	69	C34		2	73
	ES7		44	32	LS29		23	52	C35		0	75
	ES8		7	69	LS30		4	71	C36		2	73
	ES10		7	69	LS23		16	59	C37		0	75
	ES2		63	13	LS24		21	54	C38		2	73
	ES4 [^]		64	12	LS25		1	74				
	ES9		16	60								
	EO11	75 _(COMP)	45	30	EO11	82 _(O)	40	42				
	EO18		16	59	EO18		20	62				
	EO19		23	52	EO19		16	66				
	EO21		55	20	EO21		50	32				
	EO12		2	73	EO12		2	80				
	EO13		11	64	EO13		13	69				
	EO14 [^]		48	27	EO14 [^]		47	35				
	EO15		10	65	EO15		10	72				
	EO16		10	65	EO16		7	75				
	EO17		4	71	EO17		0	82				
	EO20		9	66	EO20		4	78				

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (O) = Options; [^] = indicates a guideline that contains known bias for this test item.

Table T.5

Descriptive Statistics – Test Items B-12, B-14

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-12	C31	87 _(C)	1	86								
	C32		75	12								
	C33		20	67								
	C34		21	66								
	C35		10	77								
	C36		0	87								
	C37		47	40								
	C38		2	85								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-14	EO11 [^]	81 _(O)	45	36								
	EO18		21	60								
	EO19		18	63								
	EO21		81	0								
	EO12		0	81								
	EO13		0	81								
	EO14		0	81								
	EO15		2	79								
	EO16		1	80								
	EO17		3	78								
	EO20		1	80								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; EO = guideline in the dimension of bias in the options; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (O) = Options; (C) = Cultural; [^] = indicates a guideline that contains known bias for this test item.

Table T.6

Descriptive Statistics – Test Item B-13

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-13	ES1	69 _(COMP)	65	8	ES1	78 _(STEM)	65	13	EO11	69 _(COMP)	60	12
	ES3 [^]		40	33	ES3 [^]		24	54	EO18		21	51
	ES5		7	66	ES5		3	75	EO19		19	53
	ES6		26	47	ES6		24	54	EO21		38	34
	ES7		34	39	ES7		37	41	EO12		0	72
	ES8		1	72	ES8		2	76	EO13		0	72
	ES10		6	67	ES10		3	75	EO14		9	63
	ES2		16	57	ES2		11	67	EO15		27	45
	ES4		0	73	ES4		0	78	EO16		19	53
	ES9		5	68	ES9		11	67	EO17		2	70
									EO20		25	47
	LS22	69 _(COMP)	37	34	LS22	76 _(L-S)	49	27				
	LS26		18	53	LS26		21	55				
	LS27		18	53	LS27		24	52				
	LS28		7	64	LS28		6	70				
	LS29 [^]		16	55	LS29 [^]		15	61				
	LS30		5	66	LS30		8	68				
	LS23 [^]		13	58	LS23 [^]		12	64				
	LS24		6	65	LS23 [^]		9	67				
	LS25		8	63	LS25		9	67				
	C31	69 _(COMP)	3	66	C31	86 _(C)	3	83				
	C32		0	69	C32		0	86				
	C33		3	66	C33		2	84				
	C34		3	66	C34		3	83				
	C35		1	68	C35		1	85				
	C36		6	63	C36		8	78				
	C37		40	29	C37		59	27				
	C38		7	62	C38		3	83				

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem; (O) = Options; (L-S) = Linguistic-Structural; (C) = Cultural; [^] = indicates a guideline that contains known bias for this test item.

Table T.7

Descriptive Statistics – Test Items B-15, B-16, B-20

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-15	C31 [^]	86 _(C)	23	63								
	C32		0	86								
	C33		8	78								
	C34		7	79								
	C35		0	86								
	C36		17	69								
	C37		17	69								
	C38		1	85								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-16	C31 [^]	86 _(C)	75	11								
	C32		74	12								
	C33		81	5								
	C34		44	42								
	C35		12	74								
	C36		44	42								
	C37		19	67								
	C38		11	75								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-20	LS22	75 _(L-S)	18	57								
	LS26 [^]		21	54								
	LS27		15	60								
	LS28		7	68								
	LS29		7	68								
	LS30		6	69								
	LS23		8	67								
	LS24		8	67								
	LS25		8	67								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (L-S) = Linguistic-Structural; (C) = Cultural; [^] = indicates a guideline that contains known bias for this test item.

Table T.8

Descriptive Statistics – Test Item B-18

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-18	ES1	66 _(COMP)	1	65	ES1	78 _(S)	2	76	EO11	66 _(COMP)	13	53
	ES3		30	36	ES3		28	50	EO18		26	40
	ES5 [^]		18	48	ES5 [^]		24	54	EO19		25	41
	ES6		10	56	ES6		11	67	EO21		36	30
	ES7		16	50	ES7		16	62	EO12 [^]		58	8
	ES8		8	58	ES8		1	77	EO13		1	65
	ES10		6	60	ES10		3	75	EO14 [^]		37	29
	ES2		57	9	ES2		61	17	EO15		3	63
	ES4		17	49	ES4		8	70	EO16		18	48
	ES9		6	60	ES9		9	69	EO17		3	63
									EO20		10	56
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
	LS22 [^]	66 _(COMP)	45	21	LS22 [^]	72 _(L-S)	46	26	C31	66 _(COMP)	4	62
	LS26		35	31	LS26		39	33	C32		0	66
	LS27		48	18	LS27		49	23	C33		4	62
	LS28		9	57	LS28		11	61	C34		4	62
	LS29		19	47	LS29		18	54	C35		5	61
	LS30		5	61	LS30		8	64	C36		0	66
	LS23		22	44	LS23		16	56	C37		1	65
	LS24		6	60	LS24		10	62	C38		4	62
	LS25		8	58	LS25		22	50				

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem; (L-S) = Linguistic-Structural; [^] = indicates a guideline that contains known bias for this test item.

Table T.9

Descriptive Statistics – Test Items B-21, B-22, B-23

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-21	LS22	73 _(L-S)	13	60								
	LS26		13	60								
	LS27		23	50								
	LS28		4	69								
	LS29		4	69								
	LS30		1	72								
	LS23		10	63								
	LS24		11	62								
	LS25 [^]		55	18								
B-22	LS22	73 _(L-S)	15	58								
	LS26		26	47								
	LS27		29	44								
	LS28		12	61								
	LS29		30	43								
	LS30		13	60								
	LS23		23	50								
	LS24 [^]		53	20								
	LS25		11	62								
B-23	LS22	73 _(L-S)	14	59								
	LS26		29	44								
	LS27		54	19								
	LS28		36	37								
	LS29		29	44								
	LS30 [^]		17	56								
	LS23		9	64								
	LS24		21	52								
	LS25		5	68								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; LS = guideline in the dimension of linguistic-structural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (L-S) = Linguistic-Structural; ^ = indicates a guideline that contains known bias for this test item.

Table T.10

Descriptive Statistics – Test Items B-25, B-27, B-28

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-25	LS22	72 _(L-S)	25	47								
	LS26		23	49								
	LS27 [^]		44	28								
	LS28		12	60								
	LS29		30	42								
	LS30		10	62								
	LS23		6	66								
	LS24		6	66								
	LS25		14	58								
B-27	ES1 [^]	78 _(STEM)	67	11								
	ES3 [^]		36	42								
	ES5		8	70								
	ES6		29	49								
	ES7		37	41								
	ES8		7	71								
	ES10		8	70								
	ES2		9	69								
	ES4		1	77								
	ES9		21	57								
B-28	C31	86 _(C)	0	86								
	C32		0	86								
	C33		0	86								
	C34		1	85								
	C35		0	86								
	C36		2	84								
	C37		2	84								
	C38 [^]		52	34								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (S) = Stem; (L-S) = Linguistic-Structural; (C) = Cultural; [^] = indicates a guideline that contains known bias for this test item.

Table T.11

Descriptive Statistics – Test Items B-29, B-30

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-29	EO11	79 _(O)	9	70								
	EO18		8	71								
	EO19^		34	45								
	EO21^		5	74								
	EO12		0	79								
	EO13		1	78								
	EO14		2	77								
	EO15		4	75								
	EO16		4	75								
	EO17		0	79								
	EO20		20	59								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-30	ES1	78 _(STEM)	69	9								
	ES3		34	44								
	ES5		18	60								
	ES6		24	54								
	ES7		29	49								
	ES8		5	73								
	ES10^		12	66								
	ES2		46	32								
	ES4^		74	4								
	ES9		10	68								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (S) = Stem; (O) = Options; ^ = indicates a guideline that contains known bias for this test item.

Table T.12

Descriptive Statistics – Test Items B-31, B-32, B-33

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-31	EO11	78 _(O)	38	40	C31	86 _(C)	5	81				
	EO18		25	53	C32^		83	3				
	EO19		29	49	C33		11	75				
	EO21		15	63	C34		5	81				
	EO12		1	77	C35^		41	45				
	EO13		2	76	C36		12	74				
	EO14		8	70	C37		4	82				
	EO15		7	71	C38		4	82				
	EO16^		36	42								
	EO17		5	73								
	EO20^		67	11								
B-32	LS22	72 _(L-S)	18	54								
	LS26		9	63								
	LS27		4	68								
	LS28		2	70								
	LS29		5	67								
	LS30		2	70								
	LS23		5	67								
	LS24		18	54								
	LS25		2	70								
B-33	LS22	72 _(L-S)	22	50								
	LS26		20	52								
	LS27		34	38								
	LS28^		65	7								
	LS29		18	54								
	LS30		2	70								
	LS23		20	52								
	LS24		3	69								
	LS25		1	71								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (O) = Options; (L-S) = Linguistic-Structural; (C) = Cultural; ^ = indicates a guideline that contains known bias for this test item.

Table T.13

Descriptive Statistics – Test Items B-34, B-36

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-34	EO11	78 _(O)	27	51								
	EO18		23	55								
	EO19		22	56								
	EO21		9	69								
	EO12		3	75								
	EO13		9	69								
	EO14		2	76								
	EO15		5	73								
	EO16		10	68								
	EO17 [^]		50	28								
	EO20		20	58								
Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-36	ES1	78 _(STEM)	3	75	EO11	80 _(O)	37	43				
	ES3		3	75	EO18		18	62				
	ES5		8	70	EO19		15	65				
	ES6 [^]		18	60	EO21 [^]		10	70				
	ES7 [^]		29	49	EO12		0	80				
	ES8		0	78	EO13		0	80				
	ES10		1	77	EO14		20	60				
	ES2		4	74	EO15		1	79				
	ES4		0	78	EO16		7	73				
	ES9		15	63	EO17		3	77				
					EO20		3	77				

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (S) = Stem; (O) = Options; [^] = indicates a guideline that contains known bias for this test item.

Table T.14

Descriptive Statistics – Test Item B-35

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
B-35	ES1	65 _(COMP)	57	9	ES1	78 _(STEM)	67	11				
	ES3		31	35	ES3		43	35				
	ES5		28	38	ES5		29	49				
	ES6		51	15	ES6		64	14				
	ES7		54	12	ES7		71	7				
	ES8		20	46	ES8		35	43				
	ES10		44	22	ES10		66	12				
	ES2		4	62	ES2		4	74				
	ES4		1	65	ES4		2	76				
	ES9		50	16	ES9		64	14				
	EO11	65 _(COMP)	24	41	LS22	65 _(COMP)	2	63				
	EO18		35	30	LS26		10	55				
	EO19		18	47	LS27		6	59				
	EO21		49	16	LS28		5	60				
	EO12		1	64	LS29		4	61				
	EO13		38	27	LS30		3	62				
	EO14		27	38	LS23		13	52				
	EO15		17	48	LS24		16	49				
	EO16		2	63	LS25		8	57				
	EO17		0	65								
	EO20		14	51								
	C31	65 _(COMP)	28	37	C31	86 _(C)	33	53				
	C32		1	64	C32		0	86				
	C33		3	62	C33		3	83				
	C34		3	62	C34		7	79				
	C35		22	43	C35		33	53				
	C36		47	18	C36		61	25				
	C37		31	34	C37		39	47				
	C38		2	63	C38		2	84				

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; B = designates test item that is biased; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem; (C) = Cultural; ^ = indicates a guideline that contains known bias for this test item.

Table T.15

Descriptive Statistics – Test Item F-10

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
F-10	ES1	66 _(COMP)	0	66	ES1	78 _(STEM)	1	77				
	ES3		9	57	ES3		10	68				
	ES5		2	64	ES5		1	77				
	ES6		1	65	ES6		1	77				
	ES7		5	61	ES7		7	71				
	ES8		1	65	ES8		0	78				
	ES10		1	65	ES10		1	77				
	ES2		7	59	ES2		13	65				
	ES4		0	66	ES4		1	77				
	ES9		0	66	ES9		4	74				
	EO11	66 _(COMP)	0	66	EO11	78 _(O)	9	69				
	EO18		3	63	EO18		12	66				
	EO19		11	55	EO19		13	65				
	EO21		10	56	EO21		11	67				
	EO12		1	65	EO12		0	78				
	EO13		0	66	EO13		4	74				
	EO14		0	66	EO14		22	56				
	EO15		23	43	EO15		39	39				
	EO16		0	66	EO16		23	55				
	EO17		0	66	EO17		8	70				
	EO20		17	49	EO20		29	49				
	LS22	66 _(COMP)	30	36	C31	66 _(COMP)	0	66				
	LS26		14	52	C32		0	66				
	LS27		20	46	C33		0	66				
	LS28		7	59	C34		0	66				
	LS29		4	62	C35		0	66				
	LS30		1	65	C36		0	66				
	LS23		11	55	C37		0	66				
	LS24		0	66	C38		4	62				
	LS25		0	66								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; F = designates test item that is fair; ES = guideline in the dimension of bias in the stem; EO = guideline in the dimension of bias in the options; LS = guideline in the dimension of linguistic-structural bias; C = guideline in the dimension of cultural bias; T_P = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (COMP) = Comprehensive; (S) = Stem; (O) = Options; ^ = indicates a guideline that contains known bias for this test item.

Table T.16

Descriptive Statistics – Test Item F-8

Item	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N	GL	T _{P(TOOL)}	T _G	T _N
F-8	EO11	78 _(O)	9	69	C31	86 _(C)	0	86				
	EO18		12	66	C32		0	86				
	EO19		13	65	C33		1	85				
	EO21		11	67	C34		12	74				
	EO12		0	78	C35		0	86				
	EO13		4	74	C36		0	86				
	EO14		22	56	C37		0	86				
	EO15		39	39	C38		0	86				
	EO16		23	55								
	EO17		8	70								
	EO20		29	49								

Note. GL = FITr guideline – guidelines are listed in the order they appear on the FITr; F = designates test item that is fair; EO = guideline in the dimension of bias in the options; C = guideline in the dimension of cultural bias; T_p = number of participants evaluating each test item; T_G = total number of times each guideline is selected as violated for each test item; T_N = total number times each guideline is selected as not violated for each test item; (O) = Options; (C) = Cultural; ^ = indicates a guideline that contains known bias for this test item.

Table T.17

Descriptive Statistics by Dimension for the Items on the Comprehensive Survey

Test Item (Survey)	T _P	T _{K(B/F)}	B _{STEM}	B _{OPTIONS}	B _{L-S}	B _C
B-1 (COMP)	66	--	211	78	67	18
B-11 (COMP)	75	7	316	233	196	15
B-13 (COMP)	69	11	200	220	128	63
B-18 (COMP)	66	11	169	230	197	22
B-35 (COMP)	65	12	340	225	67	137
F-10 (COMP)	66	0	26	65	87	4

Note. (COMP) = Comprehensive; T_P = number of participants evaluating each test item; T_{KB} = total score assigned to each test item known to be biased (unfair); T_{KF} = total score assigned to each test item known to be fair (unbiased); -- = item not designated as known through review of literature; B_{STEM} = total number of guidelines selected for bias in the stem for each test item; B_{OPTIONS} = total number of guidelines selected for bias in the options for each test item; B_{L-S} = total number of guidelines selected for linguistic-structural bias for each test item; B_C = total number of guidelines selected for cultural bias for each test item.

Table T.18

Descriptive Statistics by Dimension for the Items on the Stem Survey

Test Item (Survey)	T _P	T _{K(B/F)}	B _{STEM}	B _{OPTIONS}	B _{L-S}	B _C
B-1 (S)	83	--	274			
B-3 (S)	79	2	270			
B-5 (S)	80	5	414			
B-13 (S)	78	2	180			
B-18 (S)	78	2	163			
B-27 (S)	78	2	223			
B-30 (S)	78	4	321			
B-35 (S)	78	6	445			
B-36 (S)	78	4	81			
F-10 (S)	78	0	39			

Note. (S) = Stem; T_P = number of participants evaluating each test item; T_{KB} = total score assigned to each test item known to be biased (unfair); T_{KF} = total score assigned to each test item known to be fair (unbiased); -- = item not designated as known through review of literature; B_{STEM} = total number of guidelines selected for bias in the stem for each test item; B_{OPTIONS} = total number of guidelines selected for bias in the options for each test item; B_{L-S} = total number of guidelines selected for linguistic-structural bias for each test item; B_C = total number of guidelines selected for cultural bias for each test item.

Table T.19

Descriptive Statistics by Dimension for the Items on the Options Survey

Test Item (Survey)	T _P	T _{K(B/F)}	B _{STEM}	B _{OPTIONS}	B _{L-S}	B _C
B-2 (O)	85	2		156		
B-3 (O)	83	3		231		
B-10 (O)	83	2		275		
B-11 (O)	82	1		209		
B-14 (O)	81	1		172		
B-29 (O)	79	3		87		
B-31 (O)	78	3		233		
B-34 (O)	78	2		180		
B-36 (O)	80	1		114		
F-8 (O)	78	0		170		
F-10 (O)	78	0		170		

Note. (O) = Options; T_P = number of participants evaluating each test item; T_{KB} = total score assigned to each test item known to be biased (unfair); T_{KF} = total score assigned to each test item known to be fair (unbiased); B_{STEM} = total number of guidelines selected for bias in the stem for each test item; B_{OPTIONS} = total number of guidelines selected for bias in the options for each test item; B_{L-S} = total number of guidelines selected for linguistic-structural bias for each test item; B_C = total number of guidelines selected for cultural bias for each test item.

Table T.20

Descriptive Statistics by Dimension for the Items on the Linguistic-Structural Survey

Test Item (Survey)	T _P	T _{K(B/F)}	B _{STEM}	B _{OPTIONS}	B _{L-S}	B _C
B-10 (L-S)	79	2			167	
B-13 (L-S)	76	4			153	
B-18 (L-S)	72	4			219	
B-20 (L-S)	75	1			98	
B-21 (L-S)	73	1			134	
B-22 (L-S)	73	2			212	
B-23 (L-S)	73	1			214	
B-25 (L-S)	72	1			170	
B-33 (L-S)	72	1			185	
B-32 (L-S)	72	0			65	

Note. (L-S) = Linguistic-Structural; T_P = number of participants evaluating each test item; T_{KB} = total score assigned to each test item known to be biased (unfair); T_{KF} = total score assigned to each test item known to be fair (unbiased); B_{STEM} = total number of guidelines selected for bias in the stem for each test item; B_{OPTIONS} = total number of guidelines selected for bias in the options for each test item; B_{L-S} = total number of guidelines selected for linguistic-structural bias for each test item; B_C = total number of guidelines selected for cultural bias for each test item.

Table T.21

Descriptive Statistics by Dimension for the Items on the Cultural Survey

Test Item (Survey)	T _P	T _{K(B/F)}	B _{STEM}	B _{OPTIONS}	B _{L-S}	B _C
B-2 (C)	87	2				106
B-10 (C)	87	1				36
B-12 (C)	87	2				176
B-13 (C)	86	2				79
B-15 (C)	86	1				73
B-16 (C)	86	4				360
B-28 (C)	86	1				57
B-31 (C)	86	2				165
B-35 (C)	86	1				178
F-8 (C)	86	0				13

Note. (C) = Cultural; T_P = number of participants evaluating each test item; T_{KB} = total score assigned to each test item known to be biased (unfair); T_{KF} = total score assigned to each test item known to be fair (unbiased); B_{STEM} = total number of guidelines selected for bias in the stem for each test item; B_{OPTIONS} = total number of guidelines selected for bias in the options for each test item; B_{L-S} = total number of guidelines selected for linguistic-structural bias for each test item; B_C = total number of guidelines selected for cultural bias for each test item.

APPENDIX U

COMPARISONS OF MEANS FOR GUIDELINE
AND DIMENSION

Table U.1

Known Groups Comparison: Difference of Means of Guideline Scores – Evaluate the Stem (ES)

FITr Guideline	Biased Item	T _{GB} (Mean)	Fair Item	T _{GF} (Mean)	<i>p</i>
ES-1	B-27	67 (.86)	F-10	1 (.007)	+
ES-3	B-11	29 (.35)	F-10	19 (.13)	+
	B-13	64 (.42)	F-10	19 (.13)	+
	B-27	36 (.46)	F-10	19 (.13)	+
ES-4	B-5	75 (.94)	F-10	1 (.007)	+
	B-11	68 (.84)	F-10	1 (.007)	+
	B-30	74 (.95)	F-10	1 (.007)	+
ES-5	B-11	37 (.45)	F-10	3 (.02)	+
	B-18	42 (.95)	F-10	3 (.02)	+
ES-6	B-36	18 (.23)	F-10	2 (.014)	+
ES-7	B-36	29 (.37)	F-10	12 (.08)	+
ES-10	B-30	12 (.15)	F-10	1 (.014)	+

+ $p < .05$

Table U.2

Known Groups Comparison: Difference of Means of Guideline Score – Evaluate the Options (EO)

FITr Guideline	Biased Item	T _{GB} (Mean)	Fair Item	T _{GF} (Mean)	<i>p</i>
EO11	B-14	45 (.56)	F-8	9 (.114)	+
	B-14	45 (.56)	F-10	0 (0)	+
EO12	B-18	59 (.88)	F-8	0 (0)	+
	B-10	80 (.96)	F-10	1 (.15)	+
EO13	B-3	59 (.71)	F-8	4 (.05)	+
	B-3	59 (.71)	F-10	0 (0)	+
EO14	B-11	88 (.58)	F-8	22 (.28)	+
	B-18	37 (.55)	F-10	0 (0)	+
EO15	B-2	42 (.49)	F-8	39 (.5)	
	B-2	42 (.49)	F-10	23 (.34)	+
EO16	B-31	36 (.46)	F-8	23 (.295)	+
	B-31	36 (.46)	F-10	0 (0)	+
EO17	B-34	50 (.64)	F-8	8 (.1)	+
	B-34	50 (.64)	F-10	0 (0)	+
EO19	B-29	34 (.43)	F-8	13 (.165)	+
	B-29	34 (.43)	F-10	11 (.164)	+
EO20	B-13	25 (.34)	F-10	17 (.25)	+
	B-31	67 (.86)	F-8	29 (.37)	+
EO21	B-36	10 (.125)	F-8	11 (.14)	
	B-36	10 (.125)	F-10	10 (.15)	

+ $p < .05$

Table U.3

Known Groups Comparison: Difference of Means of Guideline Score – Linguistic-Structural Bias (LS)

FITr Guideline	Biased Item	T _{GB} (Mean)	Fair Item	T _{GF} (Mean)	<i>p</i>
LS22	B-18	92 (.66)	F-10	30 (.45)	+
LS23	B-10	11 (.14)	F-10	11 (.16)	
	B-18	22 (.33)	F-10	11 (.16)	+
LS24	B-22	53 (.73)	F-10	0 (0)	+
LS25	B-21	55 (.75)	F-10	0 (0)	+
LS26	B-20	21 (.28)	F-10	14 (.209)	<i>p</i> = .16
LS27	B-25	44 (.61)	F-10	20 (.299)	+
LS28	B-33	65 (.9)	F-10	7 (.1)	+

+ *p* < .05

Table U.4

Known Groups Comparison: Difference of Means of Guideline Score – Cultural Bias (C)

FITr Guideline	Biased Item	T _{GB} (Mean)	Fair Item	T _{GF} (Mean)	<i>p</i>
C31	B-15	23 (.27)	F-8	0 (0)	+
	B-16	75 (.87)	F-10	0 (0)	+
C32	B-31	83 (.97)	F-8	0 (0)	+
	B-31	83 (.97)	F-10	0 (0)	+
C33	B-16	81 (.94)	F-8	1 (.01)	+
	B-16	81 (.94)	F-10	0 (0)	+
C35	B-31	41 (.48)	F-8	0 (0)	+
	B-31	41 (.48)	F-10	0 (0)	+
C38	B-28	52 (.6)	F-8	0 (0)	+
	B-28	52 (.6)	F-10	4 (.06)	+

+ $p < .05$

Table U.5

Known Groups Comparison: Difference of Means of Dimension Scores

Dimension	Biased Item	ST B _{DIM-B} (μ)	Fair Item	ST B _{DIM-F} (μ)	p
Stem	B-5	41.4 (.52)	F-10	6.5 (.04)	+
	B-11	60.9 (.37)	F-10	6.5 (.04)	+
	B-13	38.4 (.25)	F-10	6.5 (.04)	+
	B-18	33.6 (.23)	F-10	6.5 (.04)	+
	B-27	22.3 (.29)	F-10	6.5 (.04)	+
	B-30	32.1 (.41)	F-10	6.5 (.04)	+
	B-36	8.1 (.10)	F-10	6.5 (.04)	+
Options	B-2	21.5 (.25)	F-10	5.9 (.09)	+
	B-3	25.5 (.31)	F-8	21.5 (.27)	+
	B-10	31.3 (.38)	F-8	21.5 (.27)	+
	B-11	46.8 (.29)	F-8	21.5 (.27)	+
	B-13	20.3 (.28)	F-8	21.5 (.27)	+
	B-14	21.1 (.26)	F-10	5.9 (.09)	+
	B-18	21.2 (.32)	F-8	21.5 (.27)	+
	B-29	14.4 (.18)	F-8	21.5 (.27)	+
	B-31	26 (.33)	F-8	21.5 (.27)	+
	B-34	21.4 (.2739)	F-8	21.5 (.2727)	+
	B-36	16 (.2)	F-10	5.9 (.09)	+
Linguistic- Structural	B-11	23.1 (.29)	F-10	9.7 (.14)	+
	B-18	46.6 (.33)	F-10	5.9 (.14)	+
	B-20	10.9 (.145)	F-10	5.9 (.144)	$p = .488$
	B-21	14.9 (.20)	F-10	5.9 (.14)	+
	B-22	23.6 (.32)	F-10	5.9 (.14)	+
	B-25	18.9 (.26)	F-10	5.9 (.14)	+
	B-33	20.6 (.285)	F-10	5.9 (.14)	+
Cultural	B-15	9.1 (.106)	F-10	0.5 (.007)	+
	B-16	45 (.52)	F-8	1.6 (.019)	+
	B-28	7.1 (.083)	F-8	1.6 (.019)	+
	B-31	20.6 (.24)	F-8	1.6 (.019)	+

+ $p < .05$

APPENDIX V
RESULTS OF TESTS OF INDEPENDENCE
OF SCORES

Table V.1

Independence of Cultural Bias and Demographic Variables

Variable	Test Item									
	B-2	B-10	B-12	B-13	B-15	B-16	B-28	B-31	B-35	F-8
Gender	--	--	--	--	--	--	--	--	--	.049
Age	--	--	--	--	--	--	--	--	--	--
Ethnicity	--	--	--	--	--	--	.026	.048	--	--
Race	--	--	--	--	--	--	--	.037	--	--
Region	--	--	--	--	--	--	--	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	.002
Clinical FTE	--	--	--	--	--	.042	--	--	--	.025
Academic	--	--	--	--	--	--	--	--	--	--
Specialty	--	--	--	--	--	.030	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	--	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	--	--	--	--	--	--
CNE	--	--	--	--	--	--	--	--	--	--
Teaching Expertise	--	--	--	--	--	--	--	--	--	--
Item Writing	--	--	--	--	--	--	--	--	--	.038

Note: B = designates test item that is biased; F = designates test item that is fair; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator; -- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 103$

Table V.2

Independence of Linguistic-Structural Bias and Demographic Variables

Variable	Test Item									
	B-10	B-13	B-18	B-20	B-21	B-22	B-23	B-25	B-32	B-33
Gender	.027	.006	.013	--	--	.009	.003	.009	--	--
Age	--	--	--	--	--	--	--	--	--	--
Ethnicity	.025	--	.028	.027	.027	--	.027	.028	.028	.028
Race	--	--	--	--	--	--	--	--	--	--
Region	--	--	--	--	--	--	.033	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	--
Clinical	--	--	--	--	--	--	--	--	--	--
FTE	--	--	--	--	--	--	--	--	--	--
Academic	--	--	--	--	--	--	--	--	--	--
Specialty	--	--	--	--	--	--	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	.025	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	--	--	--	--	--	--
CNE	--	--	--	--	--	--	--	--	--	--
Teaching Expertise	--	--	--	--	--	--	--	--	--	--
Item Writing	--	--	--	.022	--	--	--	--	--	--

Note: B = designates test item that is biased; F = designates test item that is fair; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator; -- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 103$

Table V.3

Independence of Bias in the Options and Demographic Variables

Variable	Test Item									
	B-2	B-3	B-10	B-11	B-14	B-29	B-31	B-34	B-36	F-8
Gender	--	--	--	--	--	--	--	--	--	--
Age	--	--	--	--	--	--	--	--	--	--
Ethnicity	--	--	--	.048	--	--	--	--	--	--
Race	--	--	--	--	--	--	--	--	--	--
Region	--	--	--	--	--	--	--	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	--
Clinical FTE	--	--	--	--	--	--	--	--	--	--
Academic	--	--	--	--	--	--	--	--	--	--
Specialty	--	--	--	--	--	--	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	--	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	--	--	--	--	--	--
CNE	--	--	.034	--	--	--	.040	--	--	--
Teaching Expertise	--	--	--	.013	--	--	--	--	--	--
Item Writing	--	--	--	--	--	--	--	--	--	--

Note: B = designates test item that is biased; F = designates test item that is fair; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator.

-- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 106$

Table V.4

Independence of Bias in the Stem and Demographic Variables

Variable	Test Item									
	B-3	B-5	B-11	B-13	B-18	B-27	B-30	B-35	B-36	F-10
Gender	--	--	--	--	--	--	--	--	--	--
Age	--	--	--	--	--	--	--	--	--	--
Ethnicity	.048	--	--	--	.008	--	--	--	--	--
Race	--	.048	--	--	--	--	.037	.023	--	--
Region	--	--	--	--	--	--	--	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	--
Clinical FTE	--	--	--	--	--	--	--	--	--	.029
Academic	.033	--	--	.028	--	--	--	--	--	--
Specialty	--	--	--	--	--	--	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	--	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	--	--	--	--	--	--
CNE	--	--	--	--	--	--	--	--	--	--
Teaching Expertise	--	--	--	--	--	--	--	--	--	--
Item Writing	--	.048	--	.001	--	--	.012	--	--	--

Note: B = designates test item that is biased; F = designates test item that is fair; T_S = Total score assigned to a test item; S = Bias in the stem; O = Bias in the options; L-S = Linguistic-structural bias; C = Cultural bias; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator.

-- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 103$

Table V.5

Independence of Comprehensive Items and Demographic Variables – B1, B-11

Variable	Test Item									
	B-1					B-11				
	Dimension of Bias									
Variable	T _S	S	O	L-S	C	T _S	S	O	L-S	C
Gender	--	--	--	--	--	--	--	.029	--	--
Age	--	--	--	--	--	.049	.007	--	--	--
Ethnicity	--	--	--	--	--	--	--	--	--	--
Race	--	--	--	--	--	--	--	--	--	--
Region	--	--	--	--	--	--	--	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	--
Clinical FTE	--	--	--	--	--	--	--	--	--	--
Academic	--	--	--	--	--	--	.019	--	--	--
Specialty	--	--	--	--	--	--	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	--	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	--	--	--	--	--	--
CNE	--	--	--	--	--	--	--	--	--	--
Teaching Expertise	--	--	--	--	--	--	--	--	--	--
Item Writing	--	--	--	--	.036	--	--	--	--	--

Note: B = designates test item that is biased; F = designates test item that is fair; T_S = Total score assigned to a test item; S = Bias in the stem; O = Bias in the options; L-S = Linguistic-structural bias; C = Cultural bias; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator.

-- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 103$

Table V.6

Independence of Comprehensive Items and Demographic Variables – B-13, B-18

Variable	Test Item									
	B-13					B-18				
	Dimension of Bias									
Variable	T _S	S	O	L-S	C	T _S	S	O	L-S	C
Gender	--	--	--	--	--	--	--	--	--	--
Age	--	--	--	--	--	--	--	--	--	.004
Ethnicity	--	--	--	--	--	--	--	--	--	--
Race	--	--	--	--	--	--	--	--	--	--
Region	--	--	--	--	--	--	--	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	--
Clinical FTE	--	--	--	--	--	--	--	--	--	--
Academic	--	--	--	--	--	--	--	--	--	--
Specialty	--	--	--	--	--	--	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	--	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	--	--	--	--	--	--
CNE	--	--	--	--	--	--	--	.012	--	.032
Teaching Expertise	--	--	--	--	--	--	--	--	--	--
Item Writing	--	--	--	--	--	--	--	--	--	--

Note: B = designates test item that is biased; F = designates test item that is fair; T_S = Total score assigned to a test item; S = Bias in the stem; O = Bias in the options; L-S = Linguistic-structural bias; C = Cultural bias; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator.

-- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 103$

Table V.7

Independence of Comprehensive Items and Demographic Variables – B-35, F-10

Variable	Test Item									
	B-35					F-10				
	Dimension of Bias									
Variable	T _S	S	O	L-S	C	T _S	S	O	L-S	C
Gender	--	--	--	--	--	--	--	--	--	--
Age	.045	--	--	--	--	--	.045	--	--	--
Ethnicity	--	--	--	--	--	--	--	--	--	--
Race	--	--	--	--	--	--	--	--	--	--
Region	--	--	--	--	--	--	--	--	--	--
Highest Degree	--	--	--	--	--	--	--	--	--	--
Year	--	--	--	--	--	--	--	--	--	--
Clinical FTE	--	--	--	--	--	--	--	--	--	--
Academic	--	--	--	--	--	.048	--	--	--	--
Specialty	--	--	--	--	--	--	--	--	--	--
Faculty	--	--	--	--	--	--	--	--	--	--
FT/PT Status	--	--	--	--	--	--	--	--	--	--
Tenure	--	--	--	--	--	--	--	--	--	--
Rank	--	--	--	--	.014	--	--	--	--	--
CNE	--	--	--	--	--	--	--	--	--	--
Teaching Expertise	--	--	--	--	--	--	--	--	--	--
Item Writing	--	--	--	--	--	--	--	--	--	--

Note: B = designates test item that is biased; F = designates test item that is fair; T_S = Total score assigned to a test item; S = Bias in the stem; O = Bias in the options; L-S = Linguistic-structural bias; C = Cultural bias; FTE = years of full-time equivalent experience; FT/PT = full-time/part-time; CNE = Certified Nurse Educator.

-- indicates $p > .05$, $CI = 95$; measures of independence are listed for $p < .05$, $CI 95$; $N = 103$

APPENDIX W
AGREEMENT INDICES

Table W.1

Agreement Indices – Guideline Level – Bias in the Stem (ES)

Perfect (.9-1)		Excellent (.8-.89)	Very Good (.7-.79)	Good (.6-.69)	Fair (.5-.59)	Poor (< .5)
B3-ES4	B35-ES2	B13-ES9	B1-ES2	B30-ES6	B18-ES3	
B3-ES9	B13-ES10	B1-ES9	B35-ES9	B3-ES6	B30-ES2	
B13-ES4	B18-ES8	B5-ES1	B11-ES9	B13-ES6	B11-ES6	
B36-ES8	B18-ES10	B3-ES5	B3-ES7	B5-ES3	B11-ES7	
B36-ES4	B5-ES4	B27-ES2	B18-ES7	B11-ES3	B13-ES3	
B1-ES4	B3-ES8	B30-ES1	B5-ES6	B5-ES7	B11-ES5	
F10-ES1	B30-ES8	B30-ES9	B30-ES5	B27-ES6	B30-ES3	
F10-ES8	B13-ES5	B35-ES7	B36-ES6	B30-ES7	B27-ES3	
F10-ES4	B1-ES5	F10-ES3	B35-ES10	B36-ES7	B1-ES8	
B27-ES4	F10-ES7	B35-ES1	B5-ES2	B35-ES8	B13-ES7	
B36-ES10	B3-ES1	F10-ES2	B3-ES2	B35-ES5	B27-ES7	
F10-ES6	B27-ES8	B13-ES1	B5-ES10		B3-ES3	
F10-ES10	B11-ES8	B3-ES10	B1-ES7		B35-ES3	
B13-ES8	B11-ES10	B27-ES1	B27-ES9			
B18-ES1	B1-ES3	B18-ES6	B1-ES6			
F10-ES5	B18-ES9	B5-ES5	B18-ES5			
B35-ES4	B36-ES5	B30-ES10				
F10-ES9	B27-ES10	B11-ES1				
B1-ES1	B27-ES5	B11-ES4				
B5-ES8		B1-ES10				
B36-ES1		B11-ES2				
B36-ES3		B18-ES4				
B5-ES9		B13-ES2				
B30-ES4		B18-ES2				
B36-ES2		B36-ES9				
		B35-ES6				
Total	44	26	16	11	13	0

Note. Items are listed in descending order in each column by MC test item and guideline. B = Biased item; F = Fair item.

Table W.2

Agreement Indices – Guideline Level – Bias in the Options (EO)

Perfect (.9-1)	Excellent (.8-.89)	Very Good (.7-.79)	Good (.6-.69)	Fair (.5-.59)	Poor (< .5)
B1-EO13	B14-EO15	B2-EO14	B35-EO20	B31-EO18	B3-EO18
B1-EO14	B11-EO12	B2-EO16	B14-EO19	B10-EO19	B35-EO13
B1-EO17	B11-EO17	B11-EO16	B36-EO18	B11-EO21	B35-EO14
B13-EO12	B29-EO14	B29-EO11	B1-EO18	B34-EO11	B29-EO19
B13-EO13	B31-EO13	B34-EO21	B11-EO18	B13-EO20	B3-EO11
B14-EO21	B34-EO14	B34-EO13	B3-EO15	F10-EO15	B18-EO14
B14-EO12	B13-EO17	F8-EO11	B3-EO19	B34-EO17	B14-EO11
B14-EO13	B35-EO16	B18-EO12	B11-EO19	B10-EO21	B11-EO11
B14-EO14	B2-EO17	B36-EO21	B35-EO21	B35-EO11	B18-EO21
B29-EO12	B3-EO20	B13-EO14	B36-EO14	B31-EO19	B3-EO21
B29-EO17	B10-EO12	B11-EO15	B29-EO20	F8-EO20	B10-EO11
B35-EO17	B14-EO17	B34-EO16	B34-EO20	B10-EO14	B31-EO16
B36-EO12	B36-EO17	B1-EO19	F10-EO20	B13-EO15	B36-EO11
B36-EO13	B36-EO20	B1-EO20	B14-EO18	B18-EO19	B35-EO18
F8-EO12	B34-EO12	B31-EO20	B35-EO15	B2-EO20	B10-EO20
F10-EO11	B1-EO11	F8-EO21	B13-EO19	B11-EO14	B13-EO21
F10-EO13	B1-EO15	B11-EO13	B13-EO16	B18-EO18	B31-EO11
F10-EO14	B18-EO15	B18-EO20	B18-EO16		B2-EO15
F10-EO16	B18-EO17	F10-EO21	B35-EO19		
F10-EO17	F10-EO18	B2-EO11	B34-EO19		
B2-EO12	B29-EO15	B2-EO19	F8-EO14		
B3-EO12	B29-EO16	F8-EO18	B3-EO13		
B3-EO17	F8-EO13	F8-EO19	B13-EO18		
B10-EO13	B3-EO16	B13-EO11	B2-EO21		
B10-EO15	B29-EO21	F10-EO19	B34-EO18		
B10-EO17	B31-EO17	B10-EO18	F8-EO16		
B14-EO16	B34-EO15	B36-EO19			
B14-EO20	B2-EO18	B31-EO21			
B36-EO15	B1-EO16	B18-EO11			
B29-EO13	B11-EO20				
B31-EO12	B36-EO16				
B1-EO12	B31-EO15				
B18-EO13	B3-EO14				
F10-EO12	B10-EO16				
B35-EO12	B29-EO18				
B2-EO13	B31-EO14				
	F8-EO17				
Total	73	29	26	17	18
					2

Note. Items are listed in descending order in each column by MC test item and guideline. B = Biased item; F = Fair item.

Table W.3

Agreement Indices – Guideline Level – Linguistic-Structural Bias (L-S)

Perfect (.9-1)	Excellent (.8-.89)	Very Good (.7-.79)	Good (.6-.69)	Fair (.5-.59)	Poor (< .5)
F10-LS24 B25-LS24	B1-LS23	B22-LS22	F10-LS27	B11-LS26	
F10-LS25 B13-LS28	F10-LS28	B13-LS29	B33-LS22	B22-LS29	
B11-LS25 B13-LS30	B20-LS23	F10-LS26	B11-LS29	B13-LS22	
B21-LS30 B1-LS29	B20-LS24	B11-LS23	B21-LS27	B25-LS29	
B33-LS25 B35-LS27	B20-LS25	B18-LS25	B22-LS23	B11-LS22	
B1-LS30 B20-LS28	B10-LS28	B10-LS24	B25-LS26	F10-LS22	
B1-LS25 B20-LS29	B10-LS29	B23-LS30	B10-LS27	B18-LS26	
F10-LS30 B18-LS30	B13-LS25	B20-LS22	B18-LS22	B33-LS27	
B32-LS28 B33-LS28	B18-LS24	B21-LS25	B25-LS22	B10-LS26	
B32-LS30 B13-LS24	B23-LS23	B35-LS24	B22-LS26	B23-LS28	
B32-LS25	B35-LS25	B32-LS22	B10-LS22		
B33-LS30	B32-LS26	B32-LS24	B25-LS27		
B35-LS22	B1-LS22	B33-LS29	B22-LS27		
B33-LS24	B21-LS23	B23-LS27	B23-LS26		
B35-LS30	B25-LS30	B13-LS26	B23-LS29		
B10-LS30	B10-LS23	B10-LS25			
B11-LS30	B18-LS28	B18-LS29			
B21-LS28	B21-LS24	B1-LS26			
B21-LS29	B22-LS25	B18-LS23			
B32-LS27	B1-LS27	B22-LS24			
B1-LS28	B35-LS26	B33-LS26			
F10-LS29	B22-LS28	B33-LS23			
B35-LS29	B25-LS28	B11-LS24			
B23-LS25	B1-LS24	B20-LS26			
B32-LS29	F10-LS23	B13-LS27			
B32-LS23	B13-LS23	B23-LS24			
B35-LS28	B21-LS22	B11-LS27			
B11-LS28	B21-LS26	B18-LS27			
B20-LS30	B22-LS30				
B25-LS23	B23-LS22				
	B25-LS25				
	B20-LS27				
	B35-LS23				
Total	40	33	28	15	10
					0

Note. Items are listed in descending order in each column by MC test item and guideline. B = Biased item; F = Fair item.

Table W.4

Agreement Indices – Guideline Level – Cultural Bias (C)

Perfect (.9-1)		Excellent (.8-.89)	Very Good (.7-.79)	Good (.6-.69)	Fair (.5-.59)	Poor (< .5)
B1-C32	B12-C31	B1-C34	B16-C37	B2-C38	B35-C31	
B1-C37	B15-C38	B12-C35	B12-C33	B13-C37	B2-C37	
B2-C32	B28-C34	B16-C31	B12-C34	B35-C35	B12-C37	
B2-C33	B13-C35	B16-C38	B10-C34	B28-C38	B35-C37	
B2-C35	B1-C31	B31-C33	B15-C31		B31-C35	
B10-C32	B1-C35	B12-C32	B35-C36		B16-C34	
B10-C35	B18-C37	B16-C32	B16-C36			
B10-C36	B2-C34	B16-C35				
B10-C37	B12-C38	B31-C36				
B11-C32	B28-C36	F8-C34				
B11-C35	B28-C37	B15-C36				
B11-C37	B35-C38	B15-C37				
B12-C36	B11-C34					
B13-C32	B11-C36					
B15-C32	B11-C38					
B15-C35	B1-C33					
B18-C32	B13-C33					
B18-C36	B2-C31					
B28-C31	B31-C32					
B28-C32	B13-C31					
B28-C33	B13-C34					
B28-C35	B35-C33					
F8-C31	B1-C36					
F8-C32	B10-C38					
F8-C35	B31-C37					
F8-C36	B10-C38					
F8-C37	B31-C38					
F8-C38	B11-C33					
F10-C31	B2-C36	F10-C38				
F10-C32	B16-C33	B13-C38				
F10-C33	B31-C31	B35-C34				
F10-C34	B31-C34	B11-C31				
F10-C35	B1-C38	B18-C35				
F10-C36	B18-C31	B15-C34				
F10-C37	B18-C33	B13-C36				
B35-C32	B18-C34	B10-C33				
B10-C31	B18-C38	B15-C33				
Total	83	12	6	4	7	0

Note. Items are listed in descending order in each column by MC test item and guideline. B = Biased item; F = Fair item.