

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Theses, Student Research, and Creative Activity:
Department of Teaching, Learning and Teacher
Education

Department of Teaching, Learning and Teacher
Education

7-2018

Effects of Structural Flaws on the Psychometric Properties of Multiple-Choice Questions

Sarah B. McBrien

University of Nebraska-Lincoln, sbmcbrien@gmail.com

Follow this and additional works at: <http://digitalcommons.unl.edu/teachlearnstudent>



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

McBrien, Sarah B., "Effects of Structural Flaws on the Psychometric Properties of Multiple-Choice Questions" (2018). *Theses, Student Research, and Creative Activity: Department of Teaching, Learning and Teacher Education*. 93.
<http://digitalcommons.unl.edu/teachlearnstudent/93>

This Article is brought to you for free and open access by the Department of Teaching, Learning and Teacher Education at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Theses, Student Research, and Creative Activity: Department of Teaching, Learning and Teacher Education by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

EFFECTS OF STRUCTURAL FLAWS ON THE
PSYCHOMETRIC PROPERTIES OF MULTIPLE-CHOICE QUESTIONS

by

Sarah B. McBrien

A DISSERTATION

Presented to the Faculty of
the Graduate College at the University of Nebraska
in Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy

Major: Educational Studies

Under the Supervision of Professor Allen Steckelberg

Lincoln, Nebraska

July 2018

EFFECTS OF STRUCTURAL FLAWS ON THE
PSYCHOMETRIC PROPERTIES OF MULTIPLE-CHOICE QUESTIONS

Sarah B. McBrien, Ph.D.

University of Nebraska, 2018

Adviser: Allen Steckelberg

The sentiment that there is more work to be done than there is time is pervasive among faculty members at most academic institutions. At health science centers, faculty members often balancing teaching responsibilities, clinical loads, and research endeavors. Creative use of educational support staff may provide institutions an avenue for accomplishing goals related to quality improvement, curriculum revision, and accreditation tasks. One such task is the maintenance of a bank of multiple-choice examination items that are free of structural flaws. This study measured the effects of a systematic approach to revising structural flaws in multiple-choice questions on the psychometric properties of the items. Structural flaws were identified by educational support staff instead of the faculty experts who authored the items and were responsible for teaching the content knowledge the items were intended to assess. Two-way ANOVA was used to measure the outcome of the revision project and structural flaw type on the psychometric qualities of existing conventional multiple-choice examination items. Neither variable had a statistically significant effect on the psychometric qualities of the items. Nonetheless, efforts to remove structural flaws from multiple-choice items

may lead to stronger reliability estimates, enhanced validity evidence, and an improved test-taking experience for students.

Acknowledgements

Gratitude goes first and foremost to my husband, Bob Marshall -- you have been steadfast in your support of me professionally and personally. You are my leader on the dance floor and in life.

To my big, blended family -- your love and support are the foundation on which I stand. Not a day goes by that I am grateful I get to share this life with you. Denny McBrien, Chris and Steve Pickett, Bob and Sylvia Marshall, Lisa and Todd Davis, Kelly McBrien, and Ruth and John McEntire -- your names especially deserve to be lit up on marquee, but this black and white page will have to do.

My best friends -- Nichole Christie, Kelly Stapleton, and Erin Wiesen. We have been through some of the greatest joys and the worst heartbreaks life has to offer. Your senses of humor, words of encouragement, and well-timed greeting cards do not go overlooked.

Thank you, Hugh Stoddard -- my biggest cheerleader. I am grateful that early on you saw in me what I couldn't; but most of all, I am grateful for your friendship and mentorship.

Gary Beck Dallaghan -- you have been instrumental in my development as an educator and researcher. I look forward to hearing your laughter when we next cross paths.

Special thanks to my colleagues at UNMC in the College of Medicine, College of Allied Health Professions, and beyond. Melissa Diers, Faye Haggar, Teri Hartman, Linda Love, Kyle Meyer, Jay Moore, Peggy Moore, and many others -- you deserve a round of applause for the support you provided over the last few years.

To the team at ExamSoft -- thank you for letting me share my work with the world, literally.

Last, but certainly not least, I must express my gratitude for my dissertation committee -- Al Steckelberg, Anthony Albano, Doug Golick, Guy Trainin, and Hugh Stoddard. Thank you for working over the summer to see me through this process.

Table of Contents

Chapter 1: Introduction	1
Background	1
Statement of the Problem	4
Purpose of the Study	5
Research Questions	5
Hypotheses	6
Definition of Terms	6
Significance of Study	8
Chapter 2: Literature Review	11
Best Practices in Conventional Multiple-choice Item Construction	11
All of the Above	18
None of the Above	18
Negative Phrasing	19
Unfocused Stem	20
Reliability	21
Random Error	22
Systematic Error	22
Validity	23

Content Validity.....	24
Criterion Validity.....	24
Construct Validity.....	25
Unified View of Validity.....	25
Testwiseness	26
Item-Level Indices	27
Difficulty Index.....	28
Discrimination Index.....	29
Average Answer Time.....	30
Distractor Analysis.....	30
Contribution to the Literature	31
Chapter 3: Methodology	33
Research Questions.....	33
Setting	33
Student Cohorts.....	39
Cohort Comparison.....	40
Study Variables	42
Independent Variables	42
Dependent Variables.....	42
Analysis of Variance.....	43

Assumptions.....	44
Confidentiality	46
Chapter 4: Results	47
Cohort Comparison.....	47
CUM GPA	47
MCAT PERC.....	47
Descriptive Statistics: The Items	48
Analysis of Variance.....	51
Difficulty Index.....	52
Discrimination Index	52
Average Answer Time	53
Statistical Power.....	54
Summary	55
Hypothesis 1.....	55
Hypothesis 2.....	55
Chapter 5: Discussion	57
Overview.....	57
Inclusion Criteria	57
Post-Administration Scoring.....	60
Reliability & Validity	61

Student Perception 62

Faculty Interactions..... 63

Limitations & Implications 64

 Limitations 64

 Delimitations..... 65

 Future Directions 66

Conclusions..... 67

References.....69

Appendix.....86

 Appendix A: Examples of Structural Flaws 86

 Appendix B: ANOVA Results Tables 87

List of Tables

Table 1. Number of Items by Factor	44
Table 2. Structural Flaws by Course.....	49
Table 3. Group Means: Difficulty Index.....	52
Table 4. Group Means: Discrimination Index	53
Table 5. Group Means: Average Answer Time	54

List of Figures

Figure 1. Summary of Guidelines	17
Figure 2. Procedural Timeline	38
Figure 3. Distribution of CUM GPA Scores by Year	47
Figure 4. Distribution of MCAT PERC Scores by Year	48
Figure 5. Mean Difficulty Index by Academic Year, Regardless of Flaw Type	50
Figure 6. Mean Discrimination Index by Academic Year, Regardless of Flaw Type	50
Figure 7. Mean Average Answer Time by Academic Year, Regardless of Flaw Type	51

CHAPTER 1: INTRODUCTION

Background

The Doctor of Medicine (MD) curriculum is characterized by heavy study loads, a fast-paced curriculum, and frequent, high-stakes examinations. The same high-stakes nature of medical practice makes it incumbent upon faculty and administrators in medical schools to be sure their assessment practices lead to sound decision-making about which students possess the requisite medical knowledge, technical skills, and attitudes necessary to progress through the program and ultimately enter the medical profession.

While the exact timing and sequence of events vary by program, traditional medical students experience medical school in similar fashion. The pre-clinical phase focuses on acquiring medical knowledge through introductory doctoring courses and basic science courses such as anatomy, biochemistry, physiology, pathology, and pharmacology. During the clinical phase, students move into clinical training as they work alongside practitioners in hospital and clinic settings during required and elective clerkships. Upon successful completion of the undergraduate medical education program (pre-clinical phase and clinical phase), the MD degree is conferred. Most graduates of an MD move into the graduate medical education phase of training, residency. Successful completion of licensure examinations is required at critical points along the way. Some students choose to augment their studies by earning an additional degree in business, public health, or a basic science. These variations on the traditional path to the MD can alter the sequence of events, but the major components of undergraduate medical schooling are fairly consistent for students who enroll in an allopathic school of medicine.

Assessment in the pre-clinical phase of the medical school program at the University of Nebraska College of Medicine (UNCOM) consists of primarily high-stakes summative examinations that are made up of primarily multiple-choice items but also included short answer and essay items. The multiple-choice exam is useful for faculty in many academic programs because it is designed to sample examinee knowledge in a target domain and provides objective score data for a large a number of items and a large number of examinees rather efficiently (Epstein, 2007; Kane, 2006).

In many undergraduate medical education programs, one instructor is not solely responsible for the delivery of content in a singular course. One faculty member might be identified as the course director, but he or she often engages experts who are clinicians and basic scientists with expertise in specific areas of medicine to deliver course topics to medical students through lectures, small group sessions, laboratory experiences, and simulated or real patient encounters. These same experts often take part in developing assessment items but may not be traditionally trained as educators or as multiple-choice item writers. While some physicians do engage in writing items for licensure and maintenance of certification exams for various certifying bodies at some point in their career, the level of training offered by those accrediting bodies varies widely. This variation in the depth and breadth of training for faculty members who deliver content, and therefore, construct multiple-choice items contributes to disparities in how well or how poorly individual items are written (ABEM, 2018; ABIM, 2018; ABPS, 2018; AOBFP, 2018; MRCPUK, 2018). Further, basic scientists who teach concepts like genetics, biochemistry, neuroscience, or pharmacology do not necessarily have licensure

or certification exams, and therefore, opportunities to engage in the training to write multiple-choice items for that purpose are fewer.

Updated in 2016, the National Board of Medical Examiners (NBME) provides the most trusted resource on best practices in item writing for health science instructors and test administrators (Case & Swanson, 2001; Paniagua & Swygert, 2016). Other guides for test and item development are available, and fortunately, the guidelines for writing conventional multiple-choice questions (MCQs) in which only one option is correct are fairly consistent across disciplines. In the most basic terms: each item should stand on its own; each item should measure knowledge acquisition related to one topic or idea; only one correct answer should be included; and distractors should be plausible, clearly incorrect, match the correct choice in structure and style, and avoid cues to the correct choice.

Additionally, common flaws related to the structure of multiple-choice item prompts and distractors, the focus of this study, are: avoid all of the above and none of the above options; avoid negative phrasing in the stem; avoid item structures that are unfocused in which the examinee cannot determine the correct answer without reading all of the options (Frey, Petersen, Edwards, Pedrotti, & Peyton, 2005; Haladyna & Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Kar, Lakshminarayanan, & T, 2015; Moreno, Martinez, & Muniz, 2006, 2015; Thissen, Steinberg, & Fitzpatrick, 1989).

The use of flawed multiple-choice items may contribute to increased confusion and frustration for students during testing. Equally as important, the decisions made by instructors and administrators about the academic progression of students in medical schools is ideally based on examination scores that are reliable and valid. The individual

building blocks of those exams, the individual items, must be properly designed to achieve such a goal.

Statement of the Problem

The utility of review teams for ensuring the quality of examination items to increase reliability and validity evidence is well documented (Abozaid, Park, & Tekian, 2017; Downing & Haladyna, 1997; Herndon, 2006; Weiner, 2005), but review panels typically include subject matter experts who evaluated item content alongside item structure. The human resources and infrastructure necessary to complete a comprehensive review of items are an important consideration in undertaking such a process. Faculty instructors in MD programs typically carry patient care and/or research loads in addition to their teaching responsibilities and may not be able to devote work hours to independently reviewing hundreds of multiple-choice test items for flaws related to content or structure. Gathering a group of physicians, basic scientists, and educators in the same room at the same time to review items synchronously would likely be even more difficult to achieve.

In light of faculty schedules and time constraints, educational support staff in UNCOM's Office of Medical Education (OME) developed a systematic process for identifying structural flaws in multiple-choice test items. Items with structural flaws were marked as flawed using ExamSoft™ item banking software, and faculty owners of the items were tasked with editing the items to repair the flaw(s) before the items were reused in subsequent examinations. Utilizing educational support staff in this manner may be an approach other institutions could consider for implementation.

Purpose of the Study

The purpose of this investigation was to measure the effects of structural flaws on the psychometric qualities of conventional multiple-choice examination items. Analysis focused on three generally accepted rules for multiple-choice item construction: avoid “all of the above” and “none of the above” (AOTA/NOTA) as choices; avoid negative phrasing, such as a stem like “All of the following are common symptoms of otitis media, *except*”; and avoid the use of unfocused stems, which is defined as stems that do not allow the examinee to answer the prompt without looking at each of the provided choices in order to compare them to one another. For example, “Which of the following is true of psoriasis?” These three flaw types do not represent the entire set of best practices for writing multiple-choice examination items, rather those that were identifiable by educational support staff who did not have subject matter expertise.

Research Questions

1. Does an item revision project focused on structural flaws have an effect on the psychometric qualities of multiple-choice items?
 - 1a. Does the mean difficulty index change after revision focused on item-writing best practices?
 - 1b. Does the mean discrimination index change after revision focused on item-writing best practices?
 - 1c. Does the average answer time change after revision focused on item-writing best practices?
2. How is the type of flaw identified during review associated with changes in psychometric qualities of multiple-choice items?

- 2a. Does the change in psychometric qualities of multiple-choice items differ by flaw type?

Hypotheses

1. The psychometric qualities of individual multiple-choice examination items will change as a result of item revision that is focused on structural flaws.

- 1a. Difficulty index will increase with improved item structure.

- 1b. Discrimination index will increase with improved item structure.

- 1c. Average answer time will decrease with improved item structure.

2. One or more flaw type will be associated with a change in psychometric qualities before and after revision.

- 2a. A statistical difference in the psychometric characteristics of items grouped by flaw type will exist with improved item structure.

Definition of Terms

Conventional Multiple-choice Item (MCQ): a multiple-choice examination item that consists of a stem, lead-in, and two or more distractors; the examinee is expected to respond by choosing a single correct option; the item is scored dichotomously – correct or incorrect (Case & Swanson, 2001; Paniagua & Swygert, 2016).

All of the Above: a MCQ that includes two or more distractors followed by a distractor that includes “all of the above” or slight variations on the same (“all of the above are true”; “a, b, and c are correct”) (Haladyna et al., 2002).

None of the Above: a MCQ that includes two or more distractors followed by a distractor that includes “none of the above” or slight variations of the same (“none of the above are correct”; “all of the above are incorrect”) (Haladyna et al., 2002).

Negative Phrasing: a MCQ that includes negative phrasing in the stem, such as “which of the following is not...?”; “all of the following are true, except?”; “which of the following is false?” (Haladyna et al., 2002).

Unfocused Stem: a MCQ that requires the examinee to consider all of the provided options before responding; the central idea is not included in the stem; the item does not adhere to the guideline that a test-taker could cover the options while reading the stem and be able to successfully answer the prompt (Case & Swanson, 2001; Haladyna et al., 2002; Paniagua & Swygert, 2016).

Difficulty Index: the percentage of examinees who answered the item correctly; expressed in values from 0 to 1.00 (Case & Swanson, 2001; Paniagua & Swygert, 2016).

Discrimination Index: a measure of how well an individual item differentiates between students who did well on the entire test and students who did not; expressed in values from -1.00 to 1.00 (Case & Swanson, 2001; Paniagua & Swygert, 2016).

Average Answer Time: the mean amount of time examinees used to respond to an individual item; expressed in a minutes and seconds format in ExamSoft™ and translated to a total seconds format for analysis.

Modified: an item that was recorded as containing one or more structural flaw by support staff; the item was subsequently revised by the subject matter expert to repair the flaw before reintroducing it to students in the second year of this study.

Deferred: an item that was recorded as containing one or more structural flaw by support staff; the item was not revised by the subject matter expert before reintroducing it to students during the second year of this study.

Archived: an item that was recorded as having one or more structural flaw by support staff; the subject matter expert did not repair or defer the item, rather it was removed from the bank of available items and no longer available for use in examinations.

Cumulative Grade Point Average (CUM GPA): verified by final college transcripts, the incoming grade point average for a participant's undergraduate studies.

MCAT Percentile Score (MCAT PERC): reported directly via the application system, the percentile score on the Medical College Admission Test.

Significance of Study

As noted, many of the faculty members who are responsible for authoring items for inclusion in medical school assessments do not have formal training as item writers. They are experts in their fields as geneticists, anatomists, pathologists, or oncologists, but they are not necessarily trained as educators in general or specifically on the task of crafting well-structured MCQs. The interplay between the subject matter expert and educational support personnel who have the time, experience, and resources to facilitate a review of all MCQs is an important part of this study. That is, if the psychometric characteristics of the items do improve after a deliberate, systematic process was instituted, it might be worthwhile for other institutions to consider the resources available to facilitate an ongoing and systematic process for ensuring only MCQs that are structured according to best practices are introduced to students.

The item flaws that were identified and subsequently measured in this study were identified by personnel in the Office of Medical Education (OME). Neither of the individuals were faculty members of the College at the time of review and did not have

expertise in the content included in the individual items. Rather, their expertise resided in their training and experience related to test development and administration. Their ability to examine all of the multiple-choice items being used in the first-year curriculum to identify any of those items that had structural flaws was not based on subject matter expertise, rather expertise with the item banking software, experience working with assessments in medical school, and for one of the individuals, in her direct training as part of graduate coursework in assessment. The demand on faculty members' time in mind, institutions may benefit from creative approaches to implementing a quality assurance program that utilizes educational support staff working alongside subject matter experts to ensure that all examination items meet standards for format and structure.

The opportunity to enhance the validity of examinations by identifying and repairing individually flawed test items is an important consideration for administrators in all types of curricula, not just medical school programs. Validating the investment of time and resources at UNCOM may aid administrators who wish to convince others of the necessity of personnel or technology resources that facilitate such a process in their setting. Engaging educational support staff in maintaining a bank of examination items that is free of structurally-flawed items offers opportunities to reduce strain on faculty members' time while increasing communication between those same subject matter experts and the educational support staff who have experience editing examination items and interpreting psychometric analysis of test items. A continuous quality improvement process for maintaining a bank of high-quality multiple-choice items may also enhance

the assessment experience for students, which in turn, could lead to increased student trust in the testing program.

CHAPTER 2: LITERATURE REVIEW

The goal of this literature review is to make connections between best practices in conventional multiple-choice item construction and psychometric theory and practice. It begins by reviewing best practices in the construction of multiple-choice test items, focusing on support and criticism for the three types of item flaws being investigated: all of the above/none of the above, negative phrasing, and unfocused stem. Next, psychometric analysis of tests and items is discussed in terms of reliability, validity, and item-level psychometric indices. Throughout the literature review, evidence about the effect of structural flaws on test performance is included and discussed.

Best Practices in Conventional Multiple-choice Item Construction

An entire body of work related to proper construction of conventional multiple-choice test questions exists in scholarly journals from the fields of psychology, education, and measurement. Generally speaking, the guidelines for writing multiple-choice questions are the same across those fields (Frey et al., 2005; Haladyna & Downing, 1989; Haladyna et al., 2002; Kar et al., 2015; Moreno et al., 2006, 2015; Thissen et al., 1989). Where inconclusive evidence about a guideline exists, investigations intended to settle the debate are plenty (Bishara & Lanzo, 2015; P. H. Harasym, Doran, Brant, & Lorscheider, 1993; P.H. Harasym, Price, Brant, Violato, & Lorscheider, 1992; Laprise, 2012; Odegard & Kown, 2007). These best practices were the foundation for the revision project undertaken at UNCOM.

The most trusted item writing guide for medical educators in the United States, *Constructing Written Test Questions for the Basic and Clinical Sciences*, is published by the National Board of Medical Examiners (2016) and is based on the earlier work of Case

and Swanson (2001). The NBME guide focuses on multiple types of examination items; but our focus is on the conventional multiple-choice (MCQ) item, an A-type item, in which the stem prompts examinees to choose a single best answer. The following general rules for the construction of the conventional multiple-choice item (MCQ) are cited by the NBME in their 2016 guide:

- Stem and options should include clear language and avoid vague terms such as “may” and “usually”.
- The stem should be focused. The examinee should be able to answer the question posed without looking at the provided options.
- Each of the options should stand on their own so they can be judged as entirely correct or entirely incorrect.
- Distractors (incorrect options) can be partially incorrect or entirely incorrect.

The NBME guide includes detailed descriptions of acceptable item formats and suggestions for typical item flaws to avoid, accompanied by examples of each (Paniagua & Swygert, 2016).

Specific guidance from the National Board of Medical Examiners (2016) on the three flaws being investigated for this study is included in their section titled “flaws related to irrelevant difficulty.” The NBME encourages item writers to replace “none of the above” with a plausible, specific option because including “none of the above” as an option requires the examinee to treat each of the presented options as separate true-false questions. Use of “all of the above” is not included as a technical flaw in the NBME guide (Paniagua & Swygert, 2016).

Negative phrasing in the stem of an item is listed as a technical flaw in the NBME guide, pointing out that the examinee must find the most false or most incorrect option while most of the other items included in the assessment are likely to ask them to identify the most correct option. Including a negatively worded stem carries with it the risk that the student will misunderstand the intent of the item, regardless of attempts to bold, highlight, or underline words meant to signal the student to identify the least correct option (Paniagua & Swygert, 2016).

Last, the NBME refers to an item that includes an unfocused stem as an item that violates the “cover the options” rule. They contend that if the stem and lead-in to the options are properly constructed the examinee would be able to answer the item without looking at the provided options (Paniagua & Swygert, 2016).

In their 2002 work, Haladyna, Downing, and Rodriguez identified 31 item-writing rules for classroom assessment. The rules were identified by consulting 27 textbooks and 27 research studies and reviews focused on educational testing. The 31 guidelines are divided into five areas: content concerns, formatting concerns, style concerns, writing the stem, and writing the choices. The guidelines presented by Haladyna, Downing, and Rodriguez is comprehensive in that it provides guidelines for test construction as a broader concept but also focuses on specific guidelines for the development of multiple-choice items. The “content concerns” section includes guidelines such as avoiding items based in opinion or those that are tricky or assess knowledge of trivial information. The “formatting concerns” section suggests that item authors should format items vertically, not horizontally; and the “writing the choices” section includes guidelines such as placing

choices in a logical order, avoiding cueing to the correct answer, and varying the location of the correct answer (Haladyna et al., 2002).

The 31 item writing rules developed by Haladyna, Downing, and Rodriguez (2002) either directly or indirectly address the three technical flaws investigated in this study:

- *All of the Above/None of the Above (AOTA/NOTA)*: “None-of-the-above should be used carefully”; “Avoid All-of-the-above.”
- *Negative Phrasing*: “Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.”
- *Unfocused Stem*: “Ensure that the directions in the stem are very clear”; “Include the central idea in the stem instead of the choices.”

This 2002 publication has been the foundation for follow-up studies and suggestions about best practices in conventional multiple-choice item writing (Haladyna et al., 2002).

Frey, Petersen, Edwards, Pedrotti, and Peyton (2005) employed a similar method as Haladyna, Downing, and Rodriguez by analyzing 20 classroom assessment textbooks. The authors identified 40 item-writing rules, most of which were specific to multiple-choice items. Related to the study at hand, the authors found that writers of the consulted texts agreed that “all of the above” and “none of the above” should be avoided (80% and 75% agreement, respectively). Eleven of 20 text writers (55%) agreed that negative wording should be avoided. Avoidance of an unfocused stem did not emerge in the analysis, but similar suggestions did: stems should clearly state the problem (10, 50%), distractors should not be longer than the stem (8, 40%), options should be independent of

each other (5, 25%), and complex item formats that require the examinee to determine that a and b are correct, but c is not, for example, should be avoided (3, 15%). The authors found “persuasive empirical evidence” for four guidelines: avoiding “all of the above”, ordering answer options logically, the inclusion of three to five answer choices, and avoiding complex item formats (Frey et al., 2005).

Moreno, Martinez, and Muniz (2006) aimed to validate a streamlined form of the guidelines established by Haladyna, Downing, and Rodriguez (2002 and 1989). Through a set of questionnaires sent to measurement professionals, they gained consensus that their set of 12 guidelines successfully synthesized earlier lists of guidelines and finalized the list in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). Again, the authors comment either directly or indirectly on the three structural flaws included in this investigation.

Moreno, Martinez, and Muniz (2006) directly identified items including AOTA/NOTA options as potentially problematic. They describe the inclusion of these answer options as violating the general rule that each option should be independent of one another. Moreno, Martinez, and Muniz note that using all of the above is likely to introduce difficulty to examinees with low levels of knowledge about the domain because it requires the examinee to know that at least two of the presented options are correct in order to conclude that “all of the above” is the best choice. They go on to explain that using “none of the above” as an option introduces negative logic, which is more complicated than referring to ideas in positive terms. Next, they maintain that “...it is preferable to use affirmative or clearly interrogative expressions rather than negative

ones, which tend to be more difficult to understand” (Moreno et al., 2006). Finally, the authors do not directly address use of an unfocused stem, but their guidelines underscore the importance of answer options being independent of one another and of keeping the options as short as possible. They include in their guidelines that options should be presented in a logical order, that the student should not have to place them in order; and while this guideline does not directly relate to the use of an unfocused stem, it speaks to the additional strain placed on an examinee who must read each of the options and interact with them in the context of the stem and of each other in order to identify the most correct option. (Moreno et al., 2006)

Gierl and Lai (2013) compared multiple-choice items generated under three conditions: by content specialists using traditional test development methods, by a second set of unique content specialists using traditional methods, and by automatic item generation. Automatic item generation (AIG) employs computer algorithms to create multiple-choice items with cognitive and psychometric theory as guides. The authors found that the items generated by content specialists and those created by AIG were similar in item quality on seven of eight indicators.

The eight indicators measured in Gierl and Lai’s study underscore the guidelines for multiple-choice item-writing best practices included thus far. In summary, the eight guidelines evaluated include: the question measured specific content as outlined in the test blueprint; the question is based on important topics in the domain; the question is free of grammatical errors; the central idea is presented in the stem; the stem is worded positively; the item includes only one correct option; that correct option is not cued by portions of the stem or other options; and all of the distracting options are plausible. The

eight guidelines were based on frequently cited manuscripts and directly validate two of the three item-writing principles being explored in this investigation; use of “all of the above” and “none of the above” as options was not mentioned, but a focused stem and positive wording in the stem were included as principles for item writers (Gierl & Lai, 2013).

A summary of recommendations from these five sources regarding the three flaw types studied in this investigation is presented in Figure 1.

	All of the Above/None of the Above (AOTA/NOTA)	Negative Phrasing	Unfocused Stem
NBME (2016)	Use NOTA carefully; Avoid AOTA	Word stem positively; Capitalize and boldface negative cues if must be used	Clear directions in the stem; Central idea in stem
Haladyna, Downing, & Rodriguez (2002)	Avoid AOTA; Use NOTA carefully	Use positives, not negatives; Avoid “NOT” in choices	Central idea in stem
Frey, Petersen, Edwards, Pedrotti, & Peyton (2005)	AOTA should not be used; NOTA should not be used	“Negative wording should not be used”	Stems should clearly state the problem, answer options shorter than stem, options independent of each other, avoid complex item formats
Moreno, Martinez, & Muniz (2006)	Use of AOTA/NOTA may increase difficulty	Use affirmative or interrogative terms	Options independent of one another
Gierl, Lai, & Hollis (2013)	Not included	Stem worded positively	Central idea in stem, not options

Figure 1. *Summary of Guidelines*

All of the Above. Differences of opinion about the use of “all of the above” (AOTA) and “none of the above” (NOTA) item types emerge in the literature, with those related to the use of AOTA being less contentious (Downing, 2005). Proponents of AOTA, especially when it is the correct response, contend that exposure to a list of correct options will have a positive lasting memory effect on examinees (Bishara & Lanzo, 2015). On the other hand, criticism of AOTA is based on two major considerations. First, the use of AOTA as the final distractor cues the savvy student to choose that option whether he actually knows it is the best choice or not. Such a student may earn a test score that is inflated beyond his actual mastery of the content (Bishara & Lanzo, 2015).

In some instances, an instructor is interested in assessing an examinee’s ability to identify multiple correct responses. For instance, an instructor who wants to assess his students’ ability to identify amoxicillin, cefdinir, and doxycycline as first-line pharmaceutical options for sinusitis might include all three drugs and “all of the above” as options (Aring & Chan, 2011). In a conventional multiple-choice item where one option is intended to be the best answer, it is expected that a master student will choose AOTA. Still, students who choose any of the other options are also partially correct. Since a conventional multiple-choice item is meant to be scored dichotomously, the student who is partially correct will earn zero points. Use of the “select all that apply” item type is suggested for items like that described (Bishara & Lanzo, 2015; Downing, 2005).

None of the Above. Investigators who have studied the impact of NOTA option choices encourage test developers to use “none of the above” carefully. Most prevalent is

the concern that one must include a list of exclusively incorrect option choices in order to lure students to choose “none of the above” as the best choice. We know that multiple-choice testing has a memory effect on the examinee, and introducing only incorrect options may hinder a student’s ability to recall the truly correct answer (the negative testing effect) (Brown, Schilling, & Hockensmith, 1999). When NOTA is not the best choice the negative testing effect is of less concern because the examinee ultimately identifies the most correct option from the provided list (P. H. Harasym et al., 1993; P.H. Harasym et al., 1992).

In some instances, an instructor may be interested in assessing students’ capacity to recognize that all of the provided options are incorrect. In medical education this type of test item may be beneficial because students are just as often required to know when something is not indicated as often as they are to know when something is. For instance, a medical student may be expected to know which types of over-the-counter pain relievers are indicated for a pregnant woman, but more important might be the list that is contraindicated because they are unsafe to the mother and/or fetus.

Negative Phrasing. Criticism of negative phrasing comes from the notion that a student who successfully chose the answer option that is “not correct” or “not true” does not necessarily know what *is* true about the content being assessed. The item structure prevents an examinee from expressing knowledge in a positive fashion. Empirical studies show that examinees take longer to answer negatively worded items, possibly because they are asked to deviate from the typically required frame of mind that requires them to find the correct response (Chiavaroli, 2017; P. H. Harasym et al., 1993; P.H. Harasym et al., 1992).

Unfocused Stem. An item that is structured around an unfocused stem is sometimes referred to as a multiple true-false question because the structure requires the test-taker to read each of the provided options, compare them to one another, and choose the correct answer. This violates what the NBME refers to as the “cover the option” rule, a principle of conventional MCQ writing that says examinees should be able to read the stem and respond to the prompt without looking at the options (Case & Swanson, 2001; Paniagua & Swygert, 2016).

An important distinction between a conventional multiple-choice item including an unfocused stem and a multiple true-false item is important in this study. The multiple true-false item is meant to include a list of three or more mutually exclusive statements that are entirely true or entirely false, and most importantly, the item is not scored dichotomously. The examinee earns a point or portion of a point for each of the correct options he identifies. The conventional MCQ is meant to be scored as either correct or incorrect; a multiple true-false item is meant to be scored polytomously. Haladyna and Rodriguez (2013) refer to the multiple true-false structure as a potential replacement for complex item types (those that include options like “a and b, but not c”). Haladyna and Rodriguez further submit that the multiple true/false structure is efficient, yields strong reliability estimates, is preferred by students, and should be used in place of the complex MCQ (Albanese & Sabersm, 1988; Downing, Baranowski, Grosso, & Norcini, 1995; Frisbie, 1992; Haladyna & Rodriguez, 2013).

To be clear, the type of flaw this study is focused on is the unfocused stem in which a student is asked to respond to a question that asks them to read all of the options in order to make a judgment. For instance: “Which of the following is true of manic

depression?” or “Which of the following statements about the medulla is true?” In some cases, the unfocused stem is also negatively worded: “Which of statements below is incorrect?” or “Which of the following disorders is not common for geriatric patients?”

Reliability

Classical test theory is founded on the basic premise that an examinee’s total score, or observed score, is made up of his or her true score and some variable amount of error. The true score is the score we’d expect an examinee to earn on repeated attempts of the same assessment. It is meant to represent the examinee’s score in the absence of any error. Because it is not feasible to obtain scores from each student on an infinite number of exam administrations to remove the effect of error associated with each administration, the total score is an approximation of the examinee’s true capacity to perform related to the domain being evaluated. Error contributes to variability in scores and therefore affects the reliability of an examination (De Champlain, 2010; Haladyna & Downing, 2004).

Reliability is the precision with which an educational measurement produces reproducible results. That is, how accurately would scores from repeated attempts on the same assessment be replicated in the absence of changing motivation, skill or knowledge acquisition, or practice effects? Scores from repeated administrations closer to one another indicate an assessment that is more reliable than an assessment with scores that vary further from one another. Measurement error accounts for these differences and can be either random or systematic (Schaughency, Smith, van der Meer, & Berg, 2012; Thorndike & Thorndike-Christ, 2010).

Random Error. Random error is inconsistent and can be based on factors internal or external to the examinee and might include fatigue, physical or emotional well-being, motivation, or the physical environment in which the test is administered. Random error affects each individual differently (AERA, APA, & NCME, 2014; Haladyna & Downing, 2004; Thorndike & Thorndike-Christ, 2010).

Systematic Error. Systematic error is present when one or more factors extraneous to the test construct increase or decrease scores for all examinees. Also called construct-irrelevant variance (CIV), systematic error may be introduced by aspects of the test administration that are consistent for all test-takers and includes factors such as inadequate instructional materials for test administrators or examinees, complexity in the language or presentation of test items that is unrelated to the construct being measured, or a mis-keyed item. CIV is systematic because it negatively or positively affects individual examinees and groups of examinees similarly. Systematic error affects the average score and item-level indices consistently for all examinees (AERA et al., 2014; Haladyna & Downing, 2004; Thorndike & Thorndike-Christ, 2010).

Error introduced by structural flaws in the multiple-choice items being analyzed in the present study is systematic in nature because the manner in which the flaws affect examinee scores is likely to be similar for all test-takers. Increased item difficulty as a result of negative phrasing, for example, is likely to decrease the average exam score because all examinees experience the item flaw similarly. Individual differences among examinees in their ability to manage the use of negative phrasing in exam items may exist based on their reading ability, first language, or previous experience with multiple-choice

items, contributing additionally to random error (Haladyna & Rodriguez, 2013; Thorndike & Thorndike-Christ, 2010).

Error as it affects group and individual test scores is of concern to medical educators because increased error decreases reliability coefficients, making it more difficult to make decisions about exam performance for a cohort of students or for individual students. Downing (2002) investigated the effect of item flaws on construct-irrelevant variance by analyzing 33 flawed MCQs from a basic science course. He found that the flawed test items were responsible for increasing item difficulty of those items by seven percentage points over standard items (about half a standard deviation for that examination). The increased difficulty of those flawed items introduced CIV, interfering with accurate interpretation of student scores. Downing suggested enhanced faculty development in the area of multiple-choice item-writing to aid item authors in eliminating flawed items from their examinations (Downing, 2002).

Increased item difficulty as a result of structural flaws may introduce construct-irrelevant variance, resulting in depressed scores and, therefore, faulty decision-making by administrators about student progression and retention. The reliability of an assessment is intimately related to validity of the same because evidence that the test adequately measures what it is, in fact, intended to measure is the foundation of validity (Haladyna & Rodriguez, 2013).

Validity

A test can be reliable in that it consistently measures the target domain, but reliability does not beget validity. Validity asks whether the test “measures what we want to measure, all of what we want to measure, and nothing but what we want to

measure” (Thorndike & Thorndike-Christ, 2010). Validity is the degree to which evidence supports our interpretations of tests scores (Haladyna & Rodriguez, 2013; Kane, 2006). Test validation is the process of assessing of a measurement tool’s validity through content validity evidence, criterion validity evidence, and construct validity evidence. Messick (1980) reconceptualized the three types of validity evidence in a unified view of validity, noting that content validity evidence and criterion validity evidence cannot stand on their own, rather they contribute to construct validity evidence (Messick, 1980).

Content Validity. Validity evidence that evaluates the appropriateness of content included on an examination is content-related. Processes for establishing content validity evidence include explicitly describing what the test is intended to measure by considering the knowledge and skills that represent the test construct. A thorough test blueprint is developed and evaluated by subject matter experts outlining the specifications for the test. The blueprint serves as a guide for test developers and should include level of cognition (typically according to Bloom’s taxonomy) learning objectives, content areas, relative weighting of items, and preferred item types. Threats to content-related validity include underrepresentation of the domain if a valuable component of the domain is missing or misrepresentation of the domain if items that measure something other than what is defined as the test construct are included (Geisinger, Shaw, & McCormick, 2012; Kane, 2006; Knupp & Harris, 2012; Thorndike & Thorndike-Christ, 2010).

Criterion Validity. Validity evidence that is focused on how well the score from an assessment correlates with scores on another measure is criterion-related. Predictive validity is a type of criterion validity that is interested in projecting scores on one variable

based on scores from the test. Scores may be correlated with later performance on a job or success in an educational program, for example. When we are examining whether scores on a measurement tool correlate with another measure at essentially the same time, concurrent validity is the appropriate term. Collection of concurrent validity evidence is useful for determining how well a new measurement is associated with an existing, validated measurement. Criterion-related validity is more concerned with how well the measure of interest correlates with other variables than it is with the content of the measurement (Geisinger et al., 2012; Thorndike & Thorndike-Christ, 2010).

Construct Validity. A construct is a characteristic that is not observable, rather it is made up of “measurable skills, traits, or attributes” (Geisinger et al., 2012). Examples include sociability, anxiety, self-esteem, and critical reasoning. Construct validity is focused on gathering evidence that a measure is positively correlated with other measures of the target trait and negatively correlated with those measures that are outside of the defined construct (Geisinger et al., 2012; Thorndike & Thorndike-Christ, 2010).

Unified View of Validity. The unified view of validity asserts that all validity is construct validity because content-related validity evidence and criterion-related validity evidence contribute to evidence that a tool measures what the construct it is intended to measure. The unified view of validity considers test validation a process of gathering evidence to support the use of the scores for their intended purpose (Geisinger et al., 2012; Thorndike & Thorndike-Christ, 2010).

This is a good point to return to the concept of construct-irrelevant variance in testing, as it is a threat to establishing validity. If the target domain of all of the tests given as part of this study is general medical knowledge, inadvertently introducing

constructs outside of that threatens the interpretations made from exam scores. If the individual items that make up the tests given to first-year medical students are fraught with structural flaws that introduce some other construct, administrators may make faulty decisions about groups of students or individual students. Testwiseness is one such construct that may be introduced in the presence of structural flaws.

Testwiseness

Testwiseness is defined as a set of skills that allows an examinee to respond correctly to test items without actually knowing the content. Using cues embedded in the item itself, a student may be able to eliminate some options in a MCQ in order to significantly increase the odds of answering an item correctly even when she is not entirely sure of the correct answer. Testwiseness is a skill that examinees who are exposed to frequent selected-response item types may acquire over time, but it is also a skill that can be taught. In fact, some test preparation courses include skills related to testwiseness in their agenda, effectively coaching students how to choose a correct answer by using the test and the items within it to their advantage. Paying attention to the length of options, grammar in the stem and options, and other unintentional cues in the item itself may decrease examinees' ability to apply testwiseness to individual items, therefore ensuring that decisions made about earned scores are based in knowledge acquisition instead of exposure and experience completing multiple-choice examinations (Millman & Bishop, 1965)

Specifically, "all of the above" and "none of the above" options are more likely than the distractors that proceed it to be correct; and examinees understand this (Thorndike & Thorndike-Christ, 2010). They can quickly scan a list of distractors, realize

that more than one option is correct, and automatically choose “all of the above” even if they aren’t certain about the other distractors. In the following example item, a student who is certain that numbness and sharp pains in one’s extremities are common symptoms of neuropathy can easily choose choice d as the correct option even if he doesn’t know for certain that heat intolerance is a symptom of neuropathy.

Neuropathy is a common symptom of diabetes mellitus. What sensation is a patient with uncontrolled diabetes likely to experience?

- a. numbness in fingers and toes
- b. sharp pains in extremities
- c. heat intolerance
- d. all of the above (Mayo Clinic, 2018)

Faulty composition of multiple-choice items may introduce construct-irrelevant variance to the error already inherent in the testing process (Downing, 2002; Haladyna & Downing, 2004). The high stakes nature of medical school necessitates administrators’ ability to identify students who are mastering material and students who are not. Toward that aim, written examinations used to test medical knowledge during Doctor of Medicine (MD) courses must be continuously reviewed for reliability and validity evidence to ensure that administrators are making well-founded decisions about which students move forward in the curriculum and which students receive remediation.

Item-Level Indices

Since it is not necessarily the goal of an educational program to generalize beyond the institution’s students, the indices born from classical test theory are adequate and are typically available via commercial item banking software (De Champlain, 2010). Both

exam-level and item-level psychometric indices based in classical test theory are easy to calculate and used in educational programs because their assumptions can be met with modest sample sizes. Exam-level indices include values such as the mean and median score and measures of internal consistency. Item-level statistics like the difficulty index, discrimination index, point biserial correlation, average answer time, and distractor analysis are available for specific items within a test and are easily calculated within item banking software.

Difficulty Index. The difficulty index is the percent of examinees who answered an item correctly and is usually represented in decimal format. Also commonly referred to as the p-value, the difficulty index can range from 0.00 to 1.00. When 90% of the examinees answered an item correctly, the difficulty index is 0.90 (De Champlain, 2010; Haladyna & Rodriguez, 2013; Livingston, 2006; Thorndike & Thorndike-Christ, 2010).

Tracking changes in the difficulty index of an item over time may alert instructors and administrators to changes in student performance on an individual item and, therefore, the content it measures. The features of commercial software packages such as ExamSoft™ that allow institutions to categorize their items furthers analysis because data can be mined about a group of items and subsequently monitored for changes. For instance, changes in student performance on one item focused on the benefits of increasing iron in one's diet may not be entirely helpful, but grouping items that assess knowledge about nutrition and monitoring student performance on that set of items can provide valuable data about how curriculum changes, personnel changes, or other contextual factors may be affecting student performance on assessment items.

It is important to point out that the difficulty index for an item is dependent on the testing population for each administration. A set of students with stronger understanding of the domain of reference will undoubtedly earn higher percentages than a group of students with less experience and knowledge in the domain of reference (De Champlain, 2010). For the study at hand, two groups of medical students will be compared and are anecdotally similar on demographic and cognitive variables. Even so, it will be important to establish that the two student cohorts are statistically similar in order to compare their performance from one academic year to the next.

Discrimination Index. The discrimination index indicates how well an item differentiates between examinees who performed well on the assessment and examinees who performed poorly on the assessment. The discrimination index has values between -1.00 and 1.00 with values closer to 0 indicating lack of discrimination between high-performing students and low-performing students. Higher values indicate a strong correlation between students who did well on the individual item and students who did well on the assessment as a whole. An item with a 0.20 discrimination index differentiates between high- and low-performing examinees better than an item with a 0.05 discrimination index. An item with a negative discrimination index typically indicates a flaw in the item because students who performed poorly on the assessment overall answered the item correctly more often than students who performed well on the assessment (De Champlain, 2010; Haladyna & Rodriguez, 2013; Livingston, 2006; Thorndike & Thorndike-Christ, 2010).

The discrimination index and the difficulty index work together to create a picture of how a group of students performed on an individual item. As the difficulty index

approaches 1.00 and all or most examinees answer an item correctly, the discrimination index naturally decreases because there is less opportunity for differentiation amongst examinees. Likewise, an item that is extremely difficult for all exam takers will have a low discrimination index.

Average Answer Time. The average answer time for an individual item may reveal nuances in examinee behavior that the difficulty index and discrimination index cannot. That is, a statistically significant change in the average number of seconds required to answer an item might indicate a difference in students' ability to interpret an item that the difficulty index cannot reveal. Students may ultimately be able to answer an item correctly equally as often before and after revisions, but the difficulty index will not reveal nuances in the effort exerted by students to achieve this.

Distractor Analysis. Distractor analysis refers to examining which of the distractors students chose. Presented in raw values or as a percentage, most commercial software packages report the number of students who chose each of the options in a MCQ. Attention to these values can aid an instructor who seeks further information about why examinees answered an item incorrectly (Livingston, 2006). Many commercial software packages are capable of reporting all of these values by assessment and over time, which can provide powerful analysis for instructors and administrators who are interested in building a bank of items for use in an academic program. A single item can be tracked over time to watch for changes in how students respond to the item based on changes in instructional techniques or broader curriculum changes. For instance, a medical school program interested in improving scores related to genetics material may be able to analyze student responses on individual items related to genetics

throughout their academic program in order to identify common misconceptions made by students during assessments.

Contribution to the Literature

Multiple studies exist that measure the effect of structural item flaws on examinee performance. This study builds on that canon of research by examining MCQs that were identified as flawed by educational support personnel, corrected by the subject matter expert, and re-introduced to students during the next academic year. The opportunity to investigate the effect of the work done by support staff and faculty members to identify and repair flawed items is unique because we can compare student performance pre-revision and post-revision to identify correlations between certain types of flaws on a number of indices: the difficulty index, discrimination index, and average answer time. The results of this investigation may reveal valuable metrics about which item flaws to repair and which flaws can wait to be revised without affecting the decisions made about students based on their examination scores. When we consider the demands on medical school faculty members' clinical duties, research interests, and teaching responsibilities, it is prudent to consider those activities that yield the highest return on investment of time and energy.

Faculty time constraints in mind, two such staff members did the work of examining nearly 1000 multiple-choice items for structural flaws. Because neither of them was subject matter experts, they could not identify flaws in content and were, therefore, limited to identifying flaws that were identifiable based on structure alone. This type of model, in which staff members do the initial work of identifying flaws, may

be useful to other institutions who recognize a need for “cleaning up” their item bank but are not certain they have the resources to do so.

Web-based item banking and exam delivery software allows test developers to leverage the organizational structure of the item bank and the automatic computation of psychometric indices in undertaking a review of MCQs. The purpose of this study is to investigate how an item review process focused on structural flaws is associated with the psychometric qualities of those items. Specifically, this investigation focuses on how the modification of items containing widely accepted multiple-choice item flaws is related to changes in the psychometric qualities of the same items.

CHAPTER 3: METHODOLOGY

Research Questions

1. Does an item revision project focused on structural flaws have an effect on the psychometric qualities of multiple-choice items?
 - 1a. Does the mean difficulty index change after revision focused on item-writing best practices?
 - 1b. Does the mean discrimination index change after revision focused on item-writing best practices?
 - 1c. Does the average answer time change after revision focused on item-writing best practices?
2. How is the type of flaw identified during review associated with changes in psychometric qualities of multiple-choice items?
 - 2a. Does the change in psychometric qualities of multiple-choice items by flaw type?

Setting

Summative examinations in the pre-clinical curriculum at the University of Nebraska College of Medicine (UNCOM) consisted primarily of conventional multiple-choice items. Examinations were taken using ExamSoft™'s secure, offline test delivery platform, Exemplify®. Examinations were timed and delivered to students on their own mobile devices. As a general rule, total test time was calculated by allowing approximately 1.3 minutes per MCQ, a guideline that was based on the timing guideline used by the National Board of Medical Examiners for the United States Medical Licensing Examination Step series (United States Medical Licensing Examination, 2018).

Rounding to the nearest quarter of an hour was typical practice to facilitate logistics of test administration.

Items were authored by faculty members who presented content to students during lecture presentations, small group activities, and laboratory sessions. Items were collected using the administrator portal, a web-based item banking software solution. Support for ExamSoft™ was provided by the Office of Medical Education (OME). The course director had oversight of the types and number of items included on examinations and provided guidance to individuals contributing to the exam blueprint. Approval of examination items for adherence to formatting and content guidelines was completed by the course director and curriculum specialist.

Many institutions or independent departments develop further style guidelines and rules for their item writers. Over time, a core set of faculty members who contributed to the development of assessments in the medical school curriculum at UNCOM developed general content and style guidelines. The style guidelines relate to punctuation and grammar usage in the items themselves and are not in conflict with established best practices for multiple-choice item-writing. Content guidelines include limitations on the type of content included. For instance, first-year medical students at UNCOM are not expected to diagnose patient problems on summative assessments. In the second year of the program, students may be asked to diagnose patient problems during summative assessments but are not expected to be proficient in proper dosing of medications.

The computer-based testing program at the UNCOM began in 2010. At that time, commercial software packages did not appeal to administrators tasked with identifying a solution for delivering examinations to first- and second-year medical students during

their pre-clinical coursework. Instead, a computer-based testing coordinator was hired to connect multiple software solutions already in use on campus. Using a robust Microsoft Access database, the learning management system, a secure browser, and an online scoring system managed by the Information Technology Services (ITS) department, summative examinations were delivered electronically to first-year medical students beginning in the fall of 2010. The system accomplished the most basic goals of the computer-based testing program (secure delivery, systematic item banking, student feedback reports), but it came with challenges that plagued the program due to financial, logistical, and spatial limitations. For instance, the management of multiple software programs required personnel to interact with student performance data for each exam administration across multiple software platforms, increasing opportunity for human error and monopolizing the testing coordinator's time (Dolan & Burling, 2012; Vale, 2006).

In 2015, the decision to move to a commercial software package was motivated by changes in the physical space available for testing and because of changes in personnel available to support the Microsoft Access database used for item banking. The College chose ExamSoft Worldwide™ because of the company's growth in the medical education community and based on recommendations from other medical center administrators and faculty who had been using ExamSoft™'s products with success. ExamSoft™'s platform consists of two main software components.

Examinations are delivered to students using an offline, secure testing platform called Exemplify® (formerly SofTest®). The platform includes a host of options, such as the use of a calculator, a timer with built-in and customizable reminders, spell check, and the ability to view images and video embedded in the test questions. Students

download Examplify® on their own device or use an institution's laboratory equipment to complete exams. Most conventional item types such as multiple-choice, fill-in-the-blank, true-false, essay, and matching are supported.

The second component of ExamSoft™'s software is the web-based item authoring and banking portal. Instructors and administrators use this tool to organize items into folders for item banking and can categorize items on multiple factors (item writer, topic, cognitive level, accreditation standards, learning outcomes). The administrator portal keeps current and historical records of student performance on individual items and by assessment. Each item can be inspected to identify detailed logs of when the item was created, who created it, how many times it has been modified, when it was last used in an examination, how many students have answered it correctly over time and by examination, and the categories to which it is assigned. Users can generate student performance reports based on individual assessments or based on the categories to which items are assigned.

It is ExamSoft™'s web-based administrator portal that facilitated a systematic review of all MCQs slated for use in the first year curriculum at UNCOM. During the summer of 2016, the Office of Medical Education (OME) lead a project to review each MCQ for flaws in the structure of the multiple-choice items already existing in the item bank hosted by ExamSoft™. Using the administrator portal, each item was reviewed by support staff in the OME, and items containing flaws were marked as such. The individuals assigned to this task had familiarity with NBME item writing guidelines and the College's preferences for authoring multiple-choice items. Both individuals also

possessed extensive experience interpreting item analysis and counseling faculty members in the College related to those analyses.

The MCQs included in this study were items used in the basic science courses that made up the first-year curriculum at UNCOM: anatomy, biochemistry, physiology, and neurosciences. Items used as part of the clinical doctoring course were not included in analysis even though they were included in the revision process because the content and structure of those questions were vastly different than those used in basic science course exams.

Items were identified by support personnel as “requiring revision” if they did not conform to one or more established formatting guidelines. Using ExamSoft™’s item banking platform, individual items were marked for revision. Items that had more than one violation in formatting were noted to address each of those areas of improvement. Only those items focused on structural flaws were marked for revision because the item reviewers lacked the content knowledge necessary to identify flaws related to subject matter.

Identification of the items that required review was completed during the first portion of summer break in 2016, and course directors were provided instructions for revising items shortly after with the intention of completing the revision process before the beginning of the new academic year. During this process, course directors had access to the reason for an item’s assignment to the revision list and student performance data from previous test administrations. Each course director had purview to assign the task to subject matter experts who taught in his/her course or to complete the revision process independently.

Psychometric qualities of these items pre-revision (2015-16 academic year) and post-revision (2016-17 academic year) was collected as part of the day-to-day business of the Office of Medical Education and were used to measure changes in psychometric qualities of the items pre- and post-revision. A visual depiction of the timeline for collecting student performance metrics that were analyzed for this study is presented in Figure 2.

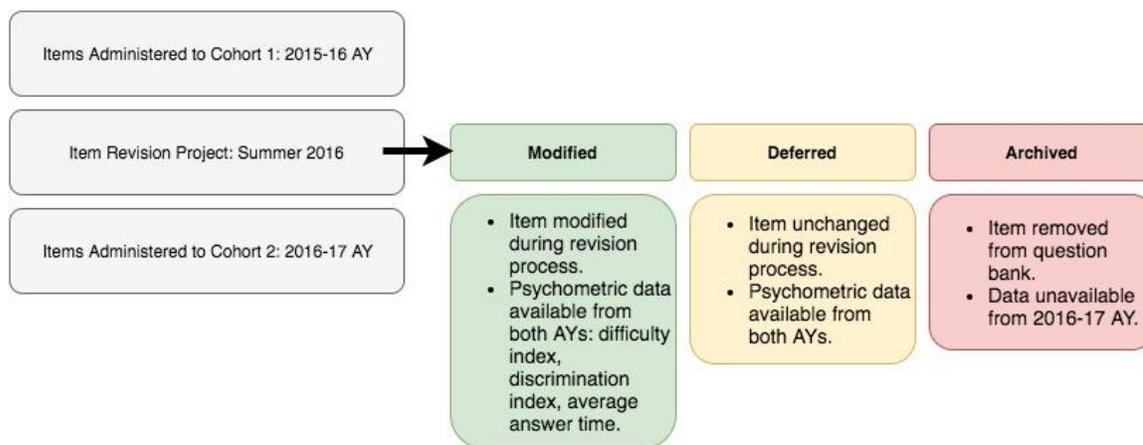


Figure 2. *Procedural Timeline*

Based on widely accepted guidelines for the structure of conventional multiple-choice items, items were classified according to three following areas of concern: all of the above/none of the above, negative phrasing, and unfocused stem (Frey et al.; Gierl & Lai; Kar et al.; Moreno et al.; Paniagua & Swygert). Items that included “all of the above” or “none of the above” in the answer choices, whether it was the intended response or not, were marked as such. Items containing phrases like “all of the following, except:” or “which is not?” were marked for negative phrasing. Items with an unfocused stem (also known as multiple true-false question) contained phrasing like “which of the following is true of...?” or “which statement best characterizes...?”. Items written with an unfocused stem require students to read each of the provided options in

order to choose the best option instead of mentally answering the question posed in the stem and subsequently finding the correct response among the answer options.

Student Cohorts

Students enrolled in the first year of the Doctor of Medicine program at UNCOM during the 2015-16 and 2016-17 academic years were the study population (2015-16 N=132, 2016-17 N=132). An identical sample of students was unachievable in this context because of the natural progression of first-year students (M1s) in 2015-16 to the second year of the curriculum in the 2016-17 academic year. The items presented to students during summative exams were presented once per academic year typically, making an opportunity to collect pre-modification and post-modification data from the same set of students impossible.

The two cohorts' exam performance was compared for this study because the groups of medical students who made up the class of first-year medical students in 2015-16 and in 2016-17 were similar in nature and completed coursework that was nearly identical in the manner of teaching, the chronological order in which it was presented, and the depth and breadth with which topics were covered. It is a generally accepted amongst medical educators at UNCOM that the characteristics of a medical school cohort do not change much from year to year, but statistical analysis was necessary to confirm this anecdotal observation. Two variables were used to confirm academic likeness: undergraduate cumulative grade point average (CUM GPA) and Medical College Admission Test percent score (MCAT PERC).

The undergraduate cumulative GPA of each student who matriculated to UNCOM was verified by official transcripts from the student's undergraduate institution and was calculated for all coursework completed during the student's baccalaureate program.

The MCAT score of each student was reported to UNCOM directly via the American Medical College Application System (AMCAS). The Association of American Medical Colleges (AAMC), authors of the MCAT, revised the format of the examination in 2015. As a result, scores were reported to institutions using a different scale depending on the timeframe in which the student completed the exam. Students who matriculated to UNCOM during the 2015-16 academic year had scores from the old version of the MCAT. Matriculants into the 2016-17 cohort had scores from both the old and the new versions. Conversion tables provided by AAMC were used to identify a percentile score for each student. This percentile score was used to compare the cohorts' MCAT scores (Association of American Medical Colleges, 2015c, 2018).

Cohort Comparison. The Mann-Whitney U test was used to confirm that any differences between the two student cohorts were not related to systematic differences between the two groups. The Mann-Whitney U test is employed when one wishes to compare the means of two groups that are not normally distributed. The distribution of CUM GPA and MCAT PERC were both left skewed with values stacked up close to 4.0 on the CUM GPA scale and the 80th percentile for MCAT PERC scores. The CUM GPA and MCAT PERC datasets met all assumptions required for the Mann-Whitney U test: the dependent variables were continuous variables; the independent variables were categorical; observations were independent of one another; and the distributions of CUM

GPA and MCAT PERC scores were the same shape for the two cohorts (Keppel & Wickens, 2004).

Results that are not statistically significant were desired in this case, as non-significant results allowed for retention of the null hypothesis that the 2015-16 cohort and the 2016-17 cohort were not statistically different from one another in their academic performance. The Mann-Whitney U test was computed separately for CUM GPA and MCAT PERC scores. The probability of rejecting the null hypothesis when it is true was set at $p < 0.05$. Setting the p-value at 0.05 ensured a 95% chance that any differences between student cohorts did not occur by chance (Cohen, 1992).

Because the number of participants in this study was limited by the cohort size of each first-year class of medical students, a priori power analysis was conducted to ensure a difference between the two groups would be detected if one existed. G*Power (Erdfelder, Faul, & Buchner, 1996) was used to conduct the analysis with alpha set at .05 and a medium effect size of 0.50 (Cohen, 1992). The necessary sample size for detecting a difference if one existed was 110 students per cohort. The cohort size of 132 students per academic year was sufficiently sized that had a difference in undergraduate cumulative GPA or MCAT percentile existed, it would have been detected.

The Kolmogorov-Smirnov (K-S) Goodness of Fit test was used to verify that the distribution of scores in each student cohort was similar for CUM GPA and MCAT PERC scores. The K-S is a nonparametric test that can be used to compare two known distributions. Conducting this test augmented the Mann-Whitney U in that it examined the distribution of scores instead of relying solely on identifying any differences in the means of the two cohorts. The K-S test was computed separately for CUM GPA and

MCAT PERC scores (Arnold & Emerson, 2011; Drezner, Turek, & Zerom, 2010; Massey, 1951).

Study Variables

Independent Variables. The first independent variable was the manner in which each flawed item was handled during the revision process. Three categories existed: modified, deferred, archived. An item was assigned to the modified category if it was perceived to be repaired by the subject matter expert (SME) and subsequently verified by the OME as such. Deferred items referred to items that the SME chose not to repair before re-use during the 2016-17 academic year. Items that were identified by SMEs as being irreparable and were removed from the item bank entirely were considered archived.

Secondly, the structural flaw(s) associated with each of the items under review was an independent variable. Each item was categorized in the ExamSoft™ administrator portal according to the structural flaws identified by staff in the OME: all of the above/none of the above, negative phrasing, and unfocused stem.

Dependent Variables. The psychometric properties of the multiple-choice items served as the dependent variables in this study: difficulty index, discrimination index, average answer time. To measure change in the indices from the 2015-16 to the 2016-17 school year, the 2015-16 value was subtracted from the 2016-17 value for each item. Hypothesis testing was carried out separately for each index.

The difficulty index is the percent of examinees who answered an item correctly and is represented in decimal format. A test item with a difficulty index of 0.75 indicates 75% of the students answered that item correctly. The difficulty index can be averaged

across all items in an examination or group of examination items to arrive at the average percent score for that group of items.

The discrimination index indicates how adequately an individual item stratifies examinees who performed well on the assessment and examinees who performed poorly on the assessment. The discrimination index has potential values between -1.00 and 1.00 with values closer to 0 indicating lack of discrimination between high-performing students and low-performing students.

The average answer time for each item served as the third dependent variable in this study and was transformed from the minutes and seconds format provided by ExamSoft™ to a total seconds format for analysis.

Analysis of Variance

To investigate the hypothesis that an item revision project led by non-subject matter experts and focused on structural flaws had an effect on the psychometric qualities of the items, two-way analysis of variance (ANOVA) was employed to inspect the effect of the item review project and the type of structural flaws on psychometric qualities. Two-way ANOVA was an ideal test statistic because it is used to identify an interaction between two independent variables and one dependent variable (Keppel & Wickens, 2004).

Three separate two-way ANOVAs were conducted, one for each dependent variable (difficulty index, discrimination index, average answer time). The mean change in each of the three dependent variables served as the unit of measure, as the change in indices from 2015-16 to 2016-17 was the value of interest.

Independent Variables (Categorical):

Result of Item Review (DISPOSITION): Modified, Deferred

Type of flaw (FLAW): All of the Above/None of the Above, Negative Phrasing, Unfocused Stem

Dependent Variables (Ratio):

Psychometric Indices: Difficulty Index, Discrimination Index, Average Answer Time

Two hundred twenty-one (221) items were identified during the item revision project as structurally flawed. Of those, 116 items were either modified or deferred. Sixty-five of those items were introduced to students during examinations given in both the 2015-16 AY and the 2016-17 AY.

Items containing “all of the above” (AOTA) or “none of the above” (NOTA) as an answer option were removed as a level under Factor B, as only one item was available for analysis in each of the modified and deferred levels and, therefore, could not be included in the ANOVA model.

Table 1. *Number of Items by Factor*

		Flaw Type (factor B)	
		<u>Negative Phrasing</u>	<u>Unfocused Stem</u>
Disposition (factor A)	Modified	31	6
	Deferred	6	20
	Total	37	26

Assumptions. Several assumptions must be met to appropriately employ ANOVA. Independent variables were categorical, and the dependent variable was continuous; no relationship between the observations in each group of independent variables or the groups themselves existed. The three remaining assumptions (absence of

outliers, normality in the dependent variable, and homogeneity in the dependent variable) were tested separately for each dependent variable.

Outliers. Visual inspection of the boxplots was performed to identify outliers in the dependent variables of difficulty index, discrimination index, and average answer time. Two outliers were identified in the difficulty index variable, and one outlier was identified in average answer time. The values were greater than three box lengths from the edge of the box and were removed from analysis (Keppel & Wickens, 2004; Laerd Statistics, 2015). Inspection of the items' psychometric qualities led the researcher to believe that a factor not related to the revision project was responsible for the extreme change. For example, an item from the physiology course with a difficulty index of 0.84 in 2015-16 and 0.05 in 2016-17 was likely impacted by some other contextual factor related to the instruction associated with that item or by a technical difficulty (E.g., missing image or figure) that was not related to the revision itself. The average answer time of the item removed from analysis changed from 65 seconds in 2015-16 to 131 seconds in 2016-17. Since the average change was less than one second for all other items, the investigator deemed this an anomaly in test administration and removed the data point from analysis of average answer time.

Normality. Shapiro-Wilk's test was utilized to assess whether the dependent variables were normally distributed. In all cases, the dependent variables were normally distributed: difficulty index: $p=.138$, discrimination index: $p=.230$, average answer time: $p=.306$. Visual inspection of histograms for each dependent variable confirmed results of the Shapiro-Wilk's test (Keppel & Wickens, 2004; Laerd Statistics, 2015).

Homogeneity. Last, two-way ANOVA assumes that the variance of the dependent variables is equal. For all three dependent variables, Levene's test for equality of error variances were not significant, indicating homogeneity of variance: difficulty index: $p=.167$, difficulty index: $p=.278$, average answer time: $p=.255$ (Keppel & Wickens, 2004; Laerd Statistics, 2015).

Analysis was completed using SPSS Version 25 (International Business Machines, 2017).

Confidentiality

A slight risk was present for members of the student cohorts being compared because cumulative GPA and MCAT scores for each student, provided by UNCOM's Office of Admissions and Student Affairs, were used as variables to determine likeness between the two student cohorts. Cumulative GPAs and MCAT scores for the two cohorts were provided for analysis without any identifying features to minimize the risk to individual students.

Student performance for each of the items considered in this investigation was downloaded in an aggregate fashion from ExamSoft™. Student names and other unique identifiers were not associated with student performance data from either academic year.

This study was approved by the Institutional Review Board at both the University of Nebraska-Lincoln and the University of Nebraska Medical Center under the exempt educational, behavioral, and social science research category.

CHAPTER 4: RESULTS

Cohort Comparison

CUM GPA. The Mann-Whitney U test was conducted to compare undergraduate cumulative grade point averages (CUM GPA) for the 2015-16 and 2016-17 student cohorts. There was not a significant difference in CUM GPA between the 2015-16 cohort ($Mdn=3.82$) and the 2016-17 cohort ($Mdn=3.82$), $U=8500$, $z=-.342$, $p=.73$.

The Kolmogorov-Smirnov distributions of CUM GPA for the 2015-16 and 2016-17 cohorts were not statistically different from each other, $D=.083$, $p=.75$. The distribution of undergraduate cumulative GPA for each academic year cohort are shown in Figure 3.

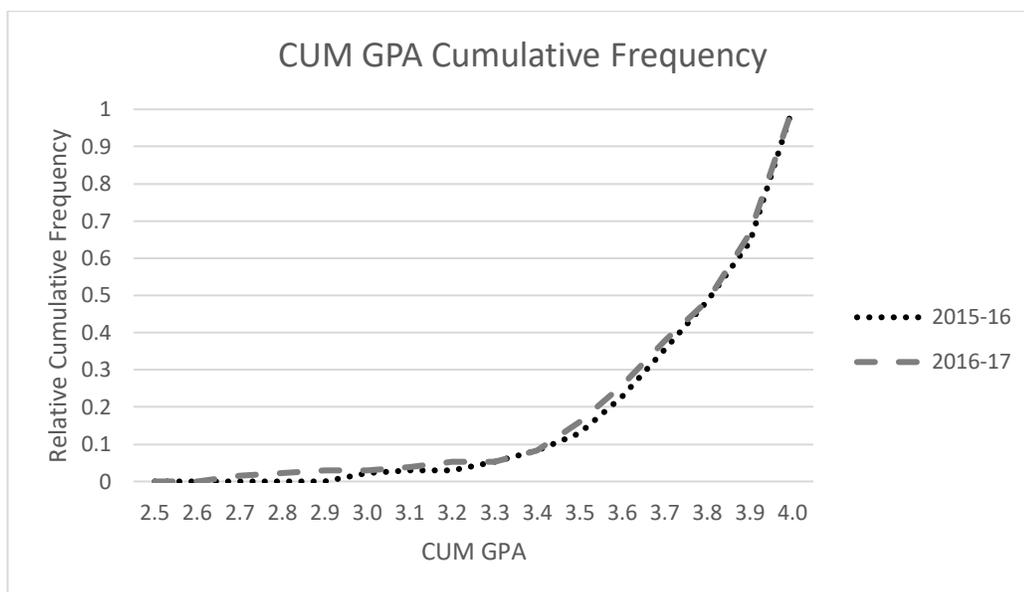


Figure 3. *Distribution of CUM GPA Scores by Year*

MCAT PERC. The Mann-Whitney U test was conducted to compare Medical College Admission Test percentile scores (MCAT) for the 2015-16 and 2016-17 student cohorts. There was not a significant difference in MCAT PERC between the 2015-16 cohort ($Mdn=83$) and the 2016-17 cohort ($Mdn=79$), $U=7657$, $z=-1.71$, $p=.09$.

The Kolmogorov-Smirnov distributions of MCAT PERC for the 2015-16 and 2016-17 cohorts were not statistically different from each other, $D=.152$, $p=.10$. The distributions of MCAT percentile scores for each academic year cohort are shown in Figure 4.

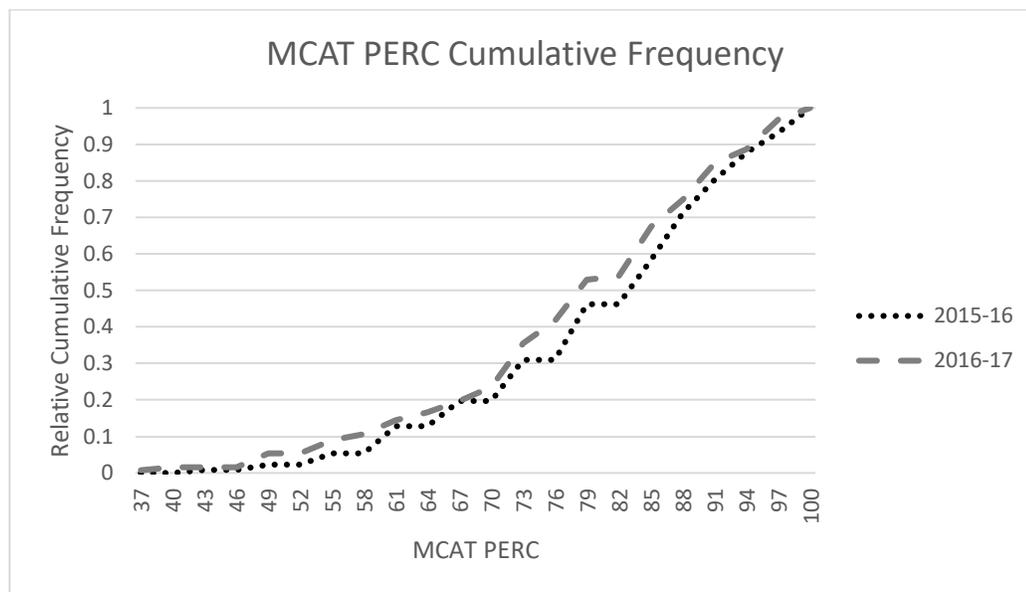


Figure 4. *Distribution of MCAT PERC Scores by Year*

Results of the Mann-Whitney U and Kolmogorov-Smirnov tests confirmed that the two student cohorts included in this investigation were not statistically different from one another in regard to academic ability. This confirmation of the academic likeness between the two groups provided evidence for comparing the psychometric properties of multiple-choice examination items administered to the two cohorts.

Descriptive Statistics: The Items

Sixty-three unique multiple-choice items were eligible for inclusion in this study. They were items that were administered to both the 2015-16 and 2016-17 cohorts and were either modified or deferred during the summer break between the academic years

included. Items that were archived or were not administered to students in both years were not eligible for inclusion.

The 63 items used for analysis came from three courses offered to medical students in the first year of the program. No items from the anatomy course were included. Fewer multiple-choice items were included in anatomy exams, and all of the items that were identified as flawed were archived by the course director.

Table 2. *Structural Flaws by Course*

		<u>Deferred</u>	<u>Modified</u>	<u>Total</u>
Negative Phrasing	Biochemistry	2	8	10
	Physiology	4	20	24
	Neurosciences	0	3	3
	Total	6	31	37
Unfocused Stem	Biochemistry	6	3	9
	Physiology	13	2	15
	Neurosciences	1	1	2
	Total	20	6	26
Grand Total		26	37	63

The mean difficulty index for 61 items included in analysis increased slightly from the 2015-16 academic year ($M=0.82$, $SD=0.13$) to the 2016-17 academic year ($M=0.85$, $SD=0.12$). The mean difficulty index for the 2015-16 school year and the 2016-17 school year are depicted by disposition, regardless of flaw type, in Figure 5.

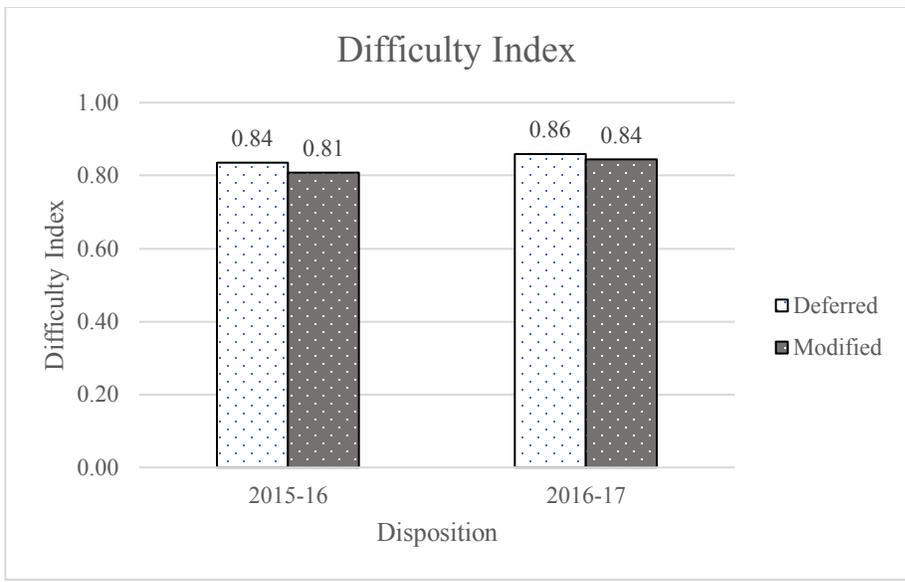


Figure 5. Mean Difficulty Index by Academic Year, Regardless of Flaw Type

The mean discrimination index for 63 items decreased slightly from the 2015-16 academic year ($M=0.23, SD=0.15$) to the 2016-17 academic year ($M=0.22, SD=0.15$).

Figure 6 shows the average discrimination index for the 2015-16 school year and the 2016-17 school year by disposition, regardless of flaw type.

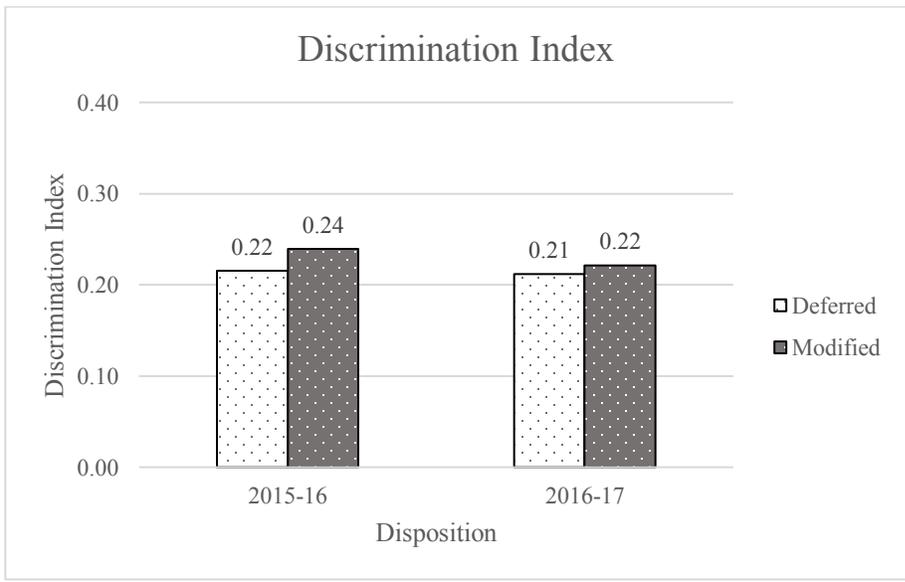


Figure 6. Mean Discrimination Index by Academic Year, Regardless of Flaw Type

Last, the mean average answer time for 62 items included in analysis increased by nearly one second from the 2015-16 academic year ($M=67.98$, $SD=12.23$) to the 2016-17 academic year ($M=68.87$, $SD=10.65$). In Figure 7, the mean average answer time for the 2015-16 school year and the 2016-17 school year is shown by disposition, regardless of flaw type.

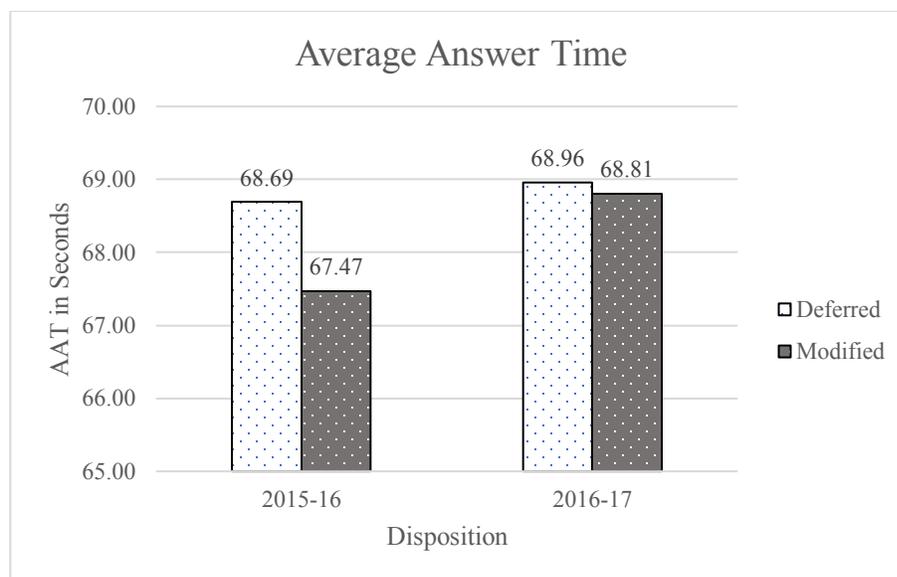


Figure 7. Mean Average Answer Time by Academic Year, Regardless of Flaw Type

Analysis of Variance

Two-way ANOVA was conducted to examine the effect of the flaw type and project disposition on the psychometric qualities of multiple-choice items. Flaw type included two levels (negative phrasing and unfocused stem), as did project disposition (deferred and modified). The three dependent variables (difficulty index, discrimination index, average answer time) were tested separately. Change in the psychometric indices (difficulty index, discrimination index, and average answer time) served as the unit of measure. The change value was achieved by subtracting the 2015-16 value from the 2016-17 value for each item.

Difficulty Index. The interaction effect between flaw type and disposition on difficulty index was not statistically significant at the .05 alpha level, $F(1, 57)=.903$, $p=.346$, partial $\eta^2=.016$. Analysis of the main effects was not significant for flaw type, $F(1, 57)=.057$, $p=.812$, partial $\eta^2=.001$, or for disposition, $F(1, 57)<.001$, $p=.995$, partial $\eta^2<.001$.

The mean change in difficulty index of all 61 items included in analysis from the 2015-16 school year to the 2016-17 school year was less than one percentage point ($M=0.029$, $SD=0.109$). The largest mean change in difficulty index existed in items that originally included negative phrasing in the stem but were revised during the revision process ($M=0.039$, $SD=0.126$).

Table 3. *Group Means: Difficulty Index*

<u>Flaw</u>	<u>Disposition</u>	<u>Mean</u>	<u>Std. Deviation</u>	<u>N</u>
Negative Phrasing (NP)	Deferred	0.005	0.076	6
	Modified	0.039	0.126	29
	Total	0.033	0.119	35
Unfocused Stem (US)	Deferred	0.031	0.080	20
	Modified	-0.003	0.144	6
	Total	0.023	0.096	26
Total	Deferred	0.025	0.079	26
	Modified	0.032	0.128	35
	Total	0.029	0.109	61

Discrimination Index. There was not a statistically significant interaction effect between flaw type and disposition on discrimination index at the .05 alpha level, $F(1, 59)=.603$, $p=.441$, partial $\eta^2=.010$. Further inspection of the main effects of flaw type, $F(1, 59)=.004$, $p=.949$, partial $\eta^2<.001$, and disposition, $F(1, 59)=.064$, $p=.801$, partial $\eta^2=.001$, on discrimination index also were not significant. On average, the

discrimination index decreased slightly from the 2015-16 to the 2016-17 academic year ($M=-0.011$, $SD=0.144$).

Table 4. *Group Means: Discrimination Index*

<u>Flaw</u>	<u>Disposition</u>	<u>Mean</u>	<u>Std. Deviation</u>	<u>N</u>
Negative Phrasing (NP)	Deferred	0.027	0.064	6
	Modified	-0.022	0.168	31
	Total	-0.014	0.156	37
Unfocused Stem (US)	Deferred	-0.013	0.125	20
	Modified	0.012	0.139	6
	Total	-0.007	0.126	26
Total	Deferred	-0.004	0.114	26
	Modified	-0.017	0.163	27
	Total	-0.011	0.144	63

Average Answer Time. The interaction effect between flaw type and disposition on average answer time was not statistically significant at the .05 alpha level, $F(1, 58) < .001$, $p = .989$, partial $\eta^2 < .001$. Analysis of the main effects was not significant for flaw type, $F(1, 58) < .001$, $p = .989$, partial $\eta^2 < .001$, or for disposition, $F(1, 58) = .052$, $p = .820$, partial $\eta^2 = .001$.

The mean of the average answer time for the 62 items included in analysis increased by less than a second ($M = .89$, $SD = 14.53$). Items that were modified during the revision process yielded an increase in average answer time of nearly one and one-half seconds ($M = 1.33$, $SD = 16.07$).

Table 5. *Group Means: Average Answer Time*

<u>Flaw</u>	<u>Disposition</u>	<u>Mean</u>	<u>Std. Deviation</u>	<u>N</u>
Negative Phrasing (NP)	Deferred	0.17	5.57	6
	Modified	1.33	15.93	30
	Total	1.14	14.65	36
Unfocused Stem (US)	Deferred	0.30	13.89	20
	Modified	1.33	18.35	6
	Total	0.54	14.63	26
Total	Deferred	0.27	12.36	26
	Modified	1.33	16.07	36
	Total	0.89	14.53	62

Statistical Power. Power is the probability of rejecting the null hypothesis when it is false. That is, how likely is it that a difference would be detected if one actually existed (Cohen, 1992)? In this study, power reflects the probability of finding a difference in the psychometric properties of the multiple-choice items after the revision project based on flaw type, disposition, and the interaction between flaw type and disposition if a difference, in fact, existed.

Power was calculated post hoc because the sample size in this study was bounded by two contextual factors that made establishing a larger set of items to include in the study impossible. Observed power of the interaction for each of the dependent variables were below the widely accepted .80 (Cohen, 1992): difficulty index: .154, discrimination index: .119, average answer time: .050. Observed power of the main effects of flaw and disposition for each of the dependent variables were also below the accepted value, ranging from .050 to .057.

Sample size for the interaction effect was below the accepted standard of $N=30$, but it met or nearly met that standard for testing the main effects of flaw and disposition

for each of the three dependent variables (Keppel & Wickens, 2004). See Tables 3, 4, and 5 for detailed group sizes.

Summary

Hypothesis 1. The psychometric qualities of multiple-choice examination items will change as a result of item revision that is focused on structural flaws.

Hypothesis 1a. Difficulty index will increase with improved item structure.

Hypothesis 1b. Discrimination index will increase with improved item structure.

Hypothesis 1c. Average answer time will decrease with improved item structure.

Results of the two-way ANOVA were not statistically significant when inspecting the effect of the revision process on any of the three dependent variables, so the null hypothesis that no change in difficulty index, discrimination index, and average answer time would occur with improved item structure was retained. Only slight differences in item-level indices from the 2015-16 academic year to the 2016-17 academic year were present and were not statistically significant for items that were modified to eliminate structural flaws.

Hypothesis 2. One or more flaw type will be associated with a change in psychometric qualities before and after revision.

Hypothesis 2. A statistical difference in the psychometric characteristics of items grouped by flaw type will exist with improved item structure.

Regardless of their disposition after the revision project (modified or deferred), no statistical difference between the negative phrasing and unfocused stem groups existed in regard to difficulty index, discrimination index, and average answer time. The null hypothesis that no difference in the psychometric characteristics for items in the negative phrasing group and the unfocused stem group would exist with improved item structure must be retained.

CHAPTER 5: DISCUSSION

Overview

The purpose of this study was to investigate the effect of a revision project focused on structural flaws in multiple-choice examination items. Initiated and led by the Office of Medical Education at UNCOM, two educational support staff members identified flaws based on widely accepted guidelines for multiple-choice item-writing (Case & Swanson, 2001; Frey et al., 2005; Gierl & Lai, 2013; Haladyna et al., 2002; Moreno et al., 2006; Paniagua & Swygert, 2016) and engaged subject matter experts in the process of repairing those flaws. While the results of the analysis did not yield statistically significant results, other factors not revealed by statistical analysis are worth consideration in the context of the effects of such a process on the assessment program.

Inclusion Criteria

Sixty-five (65) items were included for analysis in this study based on the following criteria: the item was delivered on an examination to first-year MD students during the 2015-16 academic year; the item was either modified or deferred during the revision project that took place over the summer of 2016; and the item was delivered on an examination to first-year MD students during the 2016-17 academic year. Nearly 1000 items were reviewed by educational support staff. Two hundred twenty-one (221) items were marked for review during the revision project, and only 116 of those items were either modified or deferred. The remaining 105 items were archived in the item banking software during the summer of 2016, no longer available for use in examinations. This decision to archive flawed test items could be considered a positive outcome of the revision project. The decision to archive an item was typically

accompanied with an explanation from the subject matter expert that the item was so poorly structured that it was beyond repair.

Fifty-one items (51) items were modified or deferred during the revision project but were not delivered to students during both academic years. Four (4) of these items were modified but were not used on tests in both school years. Forty-seven (47), then, were deferred during the revision project but were not used on tests in both school years. It is the 43 items that were used during the 2015-16 school year, deferred during the revision project, and not used during the 2016-17 school year that may be of interest in the context of this study. While the items were technically deferred and available for re-use during examinations after the revision project, they were not used during the 2016-17 academic year. Processes at UNCOM meant to ensure adequate assessment of the topics delivered to students would have prevented course directors and instructors from leaving the content assessed by these items out of assessments entirely. Unless the items were not included because they were deemed inappropriate based on their content, it is reasonable to assume that new items were written, or existing items without flaws were used in place of these items that retained the structural flaws. Perhaps the decision not to choose these items from the item bank reflects the advancement of the underlying purpose of the revision project, to remove structural flaws from multiple-choice questions. In the absence of statistically significant results, anecdotal evidence such as this should be considered by other institutions interested in implementing a process for eliminating flawed multiple-choice items from their existing bank.

None of the items from the anatomy course that were indicated for revision were treated in a way that made including them in analysis possible. Twelve of the 221 flawed

items came from the anatomy course, and the course director chose to archive all 12 items instead of considering them for revision. This low number is due to the nature of examinations in the anatomy course. Anatomy examinations at UNCOM were comprised of 40-60 multiple-choice items, 10-15 short answer items, two essays, and a laboratory practical component worth 40-60 points. The laboratory component requires students to identify anatomical structures in the gross anatomy lab and is short answer in nature. The three courses that were included in analysis typically included examinations with over 100 multiple-choice items. The lower number of multiple-choice items used during anatomy exams contributed to the low number of items identified as needing repair.

Further, the structure of anatomy test items did not leave much room for structural flaws. Anatomists are basic scientists trained in a specific manner to teach anatomy and conduct research related to human anatomy. Their perspective is often that medical students should be familiar with the intricate details of human anatomy – every bone, muscle, and nerve. To that end, students often find themselves memorizing lists of terms and labeled diagrams of the human body. As a result, there was less room for variation in the way examination items were presented. For example, students were instructed to “identify the radial nerve” or answer a question like, “which muscle in the upper limb crosses two joints?”. The straightforwardness with which many of the examination items were presented to students made structural flaws less likely.

Attempts to enhance MCQs used to assess anatomy knowledge by moving them to higher levels of Bloom’s Taxonomy may result in an item writer inadvertently introduce structural flaws (Bloom & Krathwohl, 1956). As anatomy instructors are encouraged to design multiple-choice items that assess higher-order thinking by

introducing clinical scenarios, for example, close attention should be paid to the structure of those items to avoid the introduction of structural flaws. Use of a review process for continuous quality assurance, including the use of educational support personnel to identify structural flaws in newly submitted test items, should be implemented to avoid the admission of structurally flawed test items.

Post-Administration Scoring

After a UNCOM cohort completes an examination and before grades are released to students, the psychometric qualities of individual items are reviewed by the course director and a member of the education team to identify items that are problematic for including in final exam grades. Causes for removing an item from scoring typically boiled down to two reasons. First, an item could be removed from scoring due to human or computer error that made responding to the item impossible. Second, an item could be removed from scoring if the psychometric properties were not acceptable.

A general rule of thumb was used at UNCOM in which the difficulty index and discrimination index were added together. If the resulting value was above 0.80, the item was generally included for scoring as is. Items with a value between 0.70 and 0.79 were inspected closely to verify that the question performed as the item writer intended. Items with values below 0.69 were considered unacceptable, and only in rare cases were they retained in scoring. For example, an item with a difficulty index of 0.55 and discrimination index of 0.18 would result in a value of 0.73. If that item was intended to be a difficult item and no apparent flaws were present, it may have been retained for final scoring.

The practice of reviewing psychometric characteristics before finalizing examination scores should stay in place, but avoiding obvious flaws before introducing items to students is ideal. If proper attention is not paid to avoiding structural flaws that lead to decreased psychometric values, flawed items may be removed from scoring after an examination is already administered to students. Doing so alters the construction of the examination and could result in underrepresentation of a domain, an important part of validity evidence. At the least, institutions should consider workflows that include a review of all new items before they are included in an exam. UNCOM was in a unique position during the revision project to leverage the availability of two support staff members who had direct training and experience in the development of multiple-choice items and review of the psychometric properties of items. Even in the absence of this luxury, institutions should consider steps toward reducing structural flaws in multiple-choice items.

Reliability & Validity

The results of this study were not what the investigator expected, but another view of the results exists related to the reliability and validity of the scores gathered from these assessments. The lack of statistical significance in change of psychometric properties of the items from the 2015-16 school year to the 2016-17 school year might reflect success in delivering a set of examination items that were modified to remove structural flaws without altering the overall test scores. Much like a parallel form of a test, the lack of significant change in the psychometric indices pre- and post-revision could be used to affirm the success of a revision project without altering student scores from one cohort to the next. Even though students at UNCOM are only compared to their cohort peers to

generate student rankings, consistency in scores over time provides evidence for the reliability of the examinations (Thorndike & Thorndike-Christ, 2010).

Attempts to limit the ability of students to use testwiseness their advantage by eliminating structural flaws that led them to the correct answer without actually mastering the content allows an institution to create confidence around the scores generated from examinations. Increased confidence in the scores based on the tests' ability to measure what they are intended to measure instead of some other construct augments validity evidence already established through the test construction and administration process (Thorndike & Thorndike-Christ, 2010).

The scores generated by multiple-choice examinations may be most important for students whose scores place them just above or just below the established passing score. An institution may be at risk for making faulty judgements about student readiness for promotion to the next level of training if the decision relies almost entirely on scores from multiple-choice examinations comprised of any number of flawed items. A student who has not mastered content but is testwise may be able to enhance his examination score just enough to move above the passing score while another student who has mastered content but is affected negatively by flawed test items may fall below the passing score. The ramifications for either situation are severe, so test administrators and faculty members in medical schools must be confident that test scores represent student knowledge and nothing else.

Student Perception

Student perceptions of the testing program may also be impacted by including structurally flawed multiple-choice items in examinations. Medical students often enter

the MD program with strong analytical skills and notice when items are removed from scoring during the post-administration scoring period at a rate higher than what is perceived to be acceptable. They are also savvy test takers and are aware of flaws in multiple-choice items and the effect those flaws have on the test-taking experience. Reducing obvious flaws may contribute to increased confidence in the item-writing skills of instructors and, more generally, in the overall effectiveness of the assessment program at an institution. If the student feels that her test scores do not adequately reflect her knowledge, confidence in the institution's ability to make appropriate decisions about student progression and retention may decline.

Student preferences toward multiple-choice items are typically based on avoiding having to construct a response to a prompt and that guessing is an available option when one is unsure of the correct answer (Gellman & Berkowitz, 1993). Reducing opportunities for guesswork in multiple-choice examinations ensures that the scores from examinations are valid and can be used confidently to make decisions about student promotion, retention, and remediation (McCoubrie, 2004; Ware & Torstein, 2009).

Faculty Interactions

A byproduct of the revision project undertaken by the Office of Medical Education (OME) at UNCOM was increased opportunity for interaction between faculty members and the support staff individuals who identified flawed items and managed the revision project. Faculty members who teach just a few topics throughout the school year were not as "in touch" with the OME as faculty members who were regular fixtures in the curriculum. This project provided reason for some of these faculty members to schedule appointments and make phone calls to talk through the revision process and potential

impacts of flawed items on students, on course directors (especially during the post-administration scoring phase), and on reliability estimates and validity evidence. In the absence of statistically significant results, this increased engagement between faculty staff around best practices in item writing should be considered a positive result of the revision project.

Alamoudi, El-Deek, Park, Shawwa, and Tekian (2017) found that faculty members who attended faculty development programs focused on item multiple-choice question properties and item analysis had significantly greater knowledge of MCQ item analysis and were more likely to conduct item analysis after administering an examination. Further, the authors contend that ongoing departmental support may lead to long-term changes in faculty members' behaviors and attitudes toward the use of psychometric analysis to improve the quality of their MCQs. Crisp and Palmer (2007) also suggest that academic units provide easily understood tools and frameworks that facilitate faculty members' understanding of principles related to assessment and student learning. Though the interactions between SMEs and educational support staff that resulted from the revision project studied here were informal in nature, the impact of such engagement of faculty members with item-writing principles and psychometric analysis could lead to long-term changes in approaches to academic units working together to use psychometric data for the purpose of improving multiple-choice items (Abozaid et al., 2017; Alamoudi, El-Deek, Park, Al Shawwa, & Tekian, 2017; Crisp & Palmer, 2007).

Limitations & Implications

Limitations. This study failed to reach desired levels for effect size and statistical power, due in large part to the limited sample size. Even if a statistically

significant difference existed for the interaction between disposition and flaw type in the entire population of items, it would not have been detected here (Cohen, 1992; Cook & Hatala, 2015). Of specific concern are two groups that contained six items: deferred negative phrasing and modified unfocused stem. The modified negative phrasing and deferred unfocused stem groups approached or reached the generally accepted sample size of 30 per group (Keppel & Wickens, 2004).

Secondly, the inability to directly compare student responses to these items pre- and post-revision is certainly a limitation. A repeated measures design in which student performance on multiple-choice items before and after being revised to remove structural flaws that are correlated by student was not possible. Multiple-choice items are typically only delivered once to students during their progression through the MD program at UNCOM, making repeated exposure impossible. Even if a repeated measures design were possible, additional consideration would have to be given to how exposure to the item on more than one occasion may alter student performance.

Delimitations. Under ideal circumstances, the sample of items available for inclusion in analysis would have been expanded to include academic years prior to 2015-16 and after 2016-17. Since UNCOM uses an item banking approach, a certain percentage of items on an exam are cycled in and out of the test blueprint each academic year. Expanding the sample beyond the two academic years included in this study may have increased sample size to well over 200 items; however, contextual factors bounded this research study to using data from the 2015-16 and 2016-17 school years and only for multiple-choice questions from the first year of the MD program.

First, the implementation of ExamSoft™ occurred at UNCOM over the summer of 2015. Psychometric characteristics of the questions included in this study were available from the technology solution used previously, but the transition from that institutionally-developed system to ExamSoft™'s online platform made including psychometric data from academic years prior to 2015-16 problematic.

Second, UNCOM underwent curriculum revision and implemented a new curriculum in the fall of 2017. The major goal of the new curriculum included transitioning to a systems-based approach in which the anatomy, physiology, and pathology related to a system of the body are taught during one course instead of teaching the same domains for the entire body at once before moving to the next domain. Secondary goals included increasing active learning opportunities, incorporating e-learning methods, and identifying opportunities for basic scientists and clinicians to co-teach concepts. This major curriculum change made comparing student performance metrics from the 2017-18 academic year to earlier years impossible because an underlying assumption of this study was that the learning experience for students in both school years was essentially unchanged in scope and sequence.

Third, multiple-choice items delivered to second-year medical students did not undergo the same revision process until the summer of 2017, making data collection from the 2017-18 academic year a complicating factor for the timeline of this investigation. For these reasons, analysis focused on psychometric qualities of exam items used in the first year of the 2015-16 and 2016-17 years only.

Future Directions. Results of this study must be considered in relation to effect size and statistical power, both of which were below desired levels. Increasing the

sample size of items included in the revision project is a natural next step in furthering research in this area. By broadening the sample to include test items from the second year of the program, the sample would likely increase at least twofold and include items from a broader set of course topics such as pathology and pharmacology, which were taught in the second year of UNCOM's curriculum prior to the curriculum revision.

Mixed methods approaches could reveal rich data about the engagement of staff members, course directors, regular instructors, and occasional instructors in best practices for item writing, about how the project could be improved, and about faculty and staff development needs.

Conclusions

As academic institutions continue to “do more with less,” creative use of resources is imperative. To implement the structured revision project studied here, two educational support staff members combed through nearly 1000 multiple-choice items to identify structural flaws. The amount of time required to complete this task was estimated at a combined 25-30 working hours. A faculty member who must juggle teaching duties, clinical loads, and research projects may not have been able to devote the hours required to complete the task, but support staff members who saw a slowdown in the pace of the work day over the summer were able to devote the time necessary to identify flawed multiple-choice items. This study did not show a statistical difference between items that were modified and items that were deferred (thus retaining flaws), nor between items with negative phrasing and items with an unfocused stem; but factors like the number of items that were archived and the number of items that were deferred during the revision period and were not subsequently used on an exam provide some

promise that the revision project was, indeed, successful. Further, the interactions between educational support staff and faculty members during the project and the positive impact of removing obvious flaws from multiple-choice items on the test-taking experience for students are positive effects that cannot be overlooked.

REFERENCES

- Abozaid, H., Park, Y. S., & Tekian, A. (2017). Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher, 39*(sup1), S50-S54.
- Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163-181. doi:10.1037/a0015719
- AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Alamoudi, A. A., El-Deek, B. S., Park, Y. S., Al Shawwa, L. A., & Tekian, A. (2017). Evaluating the long-term impact of faculty development programs on MCQ item analysis. *Medical Teacher, 39*(sup1), S45-S49.
- Albanese, M. A. (1993). Type K and other complex multiple-choice items: an analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28-33.
- Albanese, M. A., & Sabersm, D. L. (1988). Multiple true-false items: a study of interitem correlations, scoring alternatives, and reliability estimation. *Journal of Educational Measurement, 25*, 111-124.
- Ali, S. H., & Ruit, K. G. (2015). The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspectives on Medical Education, 4*(5), 244-251.

American Board of Emergency Medicine. (2018). Becoming an ABEM Item Writer.

Retrieved from <https://www.abem.org/public/general-information/who-is-abem-/becoming-an-abem-item-writer>

American Board of Internal Medicine. (2018). How Exams are Developed. Retrieved

from <http://www.abim.org/about/exam-information/exam-development.aspx>

American Board of Physician Specialties. (2018). About the American Board of

Physician Specialties. Retrieved from <http://www.abpsus.org/about-abps>

American Osteopathic Board of Family Physicians. (2018). Item Writers. Retrieved from

<http://www.aobfp.org/about/item-writers/>

Aring, A. M., & Chan, M. M. (2011). Acute Rhinosinusitis in Adults. Retrieved from

<https://www.aafp.org/afp/2011/0501/p1057.html>

Arnold, T. B., & Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete

null distributions. *The R Journal*, 3(2), 34-39.

Association of American Medical Colleges. (2015a). *Percentile Ranks for MCAT Total*

and Section Scores for Exams Administered from January 2012 through

September 2014. Retrieved from <https://aamc->

[orange.global.ssl.fastly.net/production/media/filer_public/5f/16/5f169a91-12b7-](https://aamc-orange.global.ssl.fastly.net/production/media/filer_public/5f/16/5f169a91-12b7-42e0-8749-a17f3bebe7a4/finalpercentileranksfortheoldmcatexam.pdf)

[42e0-8749-a17f3bebe7a4/finalpercentileranksfortheoldmcatexam.pdf](https://aamc-orange.global.ssl.fastly.net/production/media/filer_public/5f/16/5f169a91-12b7-42e0-8749-a17f3bebe7a4/finalpercentileranksfortheoldmcatexam.pdf)

Association of American Medical Colleges. (2015b). *Summary of Total and*

Section Scores from the MCAT Exam Based on Results for Tests Administered

in April and May 2015 Retrieved from <https://aamc->

[orange.global.ssl.fastly.net/production/media/filer_public/43/5a/435aff7b-3db8-](https://aamc-orange.global.ssl.fastly.net/production/media/filer_public/43/5a/435aff7b-3db8-4aa0-90cb-70184be8c8b8/percentilenevmcat.pdf)

[4aa0-90cb-70184be8c8b8/percentilenevmcat.pdf](https://aamc-orange.global.ssl.fastly.net/production/media/filer_public/43/5a/435aff7b-3db8-4aa0-90cb-70184be8c8b8/percentilenevmcat.pdf)

- Association of American Medical Colleges. (2015c). *Using MCAT Data in 2016 Medical Student Selection*. Retrieved from <https://www.aamc.org/download/434596/data/usingmcatdata2016.pdf>
- Association of American Medical Colleges. (2017). *Summary of MCAT Total and Section Scores*. Retrieved from https://aamc-orange.global.ssl.fastly.net/production/media/filer_public/43/5a/435aff7b-3db8-4aa0-90cb-70184be8c8b8/percentilenewmcat.pdf
- Association of American Medical Colleges. (2018). *The MCAT Essentials*. Retrieved from https://aamc-orange.global.ssl.fastly.net/production/media/filer_public/c3/de/c3de3536-7c7b-4318-99cc-431376ec23dc/essentials_2018_final_feb2018.pdf
- Bandaranayake, R. C. (2008). Setting and maintaining standards in multiple choice examinations: AMEE guide no. 37. *Medical Teacher*, 30(9-10), 836-845.
- Baranowski, R. A. (2006). Item editing and editorial review. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 349-358). New Jersey: Lawrence Erlbaum Associates.
- Bishara, A. J., & Lanzo, L. A. (2015). All of the above: when multiple correct response options enhance the testing effect. *Memory*, 23(7), 1013-1028.
- Bloom, B. S., & Krathwohl, D. R. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: Longman.
- Brown, A. S., Schilling, H. E., & Hockensmith, M. L. (1999). The negative suggestion effect: pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, 91(4), 756-764.

- Bruning, R. H., Schraw, G. J., & Norby, M. M. (2011). *Cognitive Psychology and Instruction* (5th ed.). Boston: Pearson Education, Inc.
- Brunnquell, A., Degirmenci, U., Kreil, S., Kornhuber, J., & Weih, M. (2011). Web-based application to eliminate five contraindicated multiple-choice question practices. *Evaluation & the Health Professions, 34*(2), 226-238.
doi:10.1177/0163278710370459
- Buckendahl, C. W., & Plake, B. S. (2006). Evaluating tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 725-738). New Jersey: Lawrence Erlbaum Associates.
- Burton, R. F. (2006). Sampling knowledge and understanding: how long should a test be? *Assessment & Evaluation in Higher Education, 31*(5), 569-582.
- Bush, M. (2014). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education, 40*(2), 218-231.
- Campion, D., & Miller, S. (2006). Test production effects on validity. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 599-624). New Jersey: Lawrence Erlbaum Associates.
- Case, S. M., & Swanson, D. B. (2001). *Constructing Written Test Questions for the Basic and Clinical Sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.
- Chiavaroli, N. (2017). Negatively-worded multiple choice questions: an avoidable threat to validity. *Practical Assessment, Research & Evaluation, 22*(3), 1-14.
- Clauser, J. C., & Hambleton, R. K. (2012). Item analysis procedures for classroom assessments in higher education. In C. Secolksy & D. B. Denison (Eds.),

Handbook on Measurement, Assessment, and Evaluation in Higher Education (pp. 296-309). New York: Routledge.

- Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, 1(3), 98-101.
- Collins, J. (2006). Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, 26, 543-551.
- Cook, D. A., & Hatala, R. (2015). Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education*, 20, 73-83.
- Crisp, G. T., & Palmer, E. J. (2007). Engaging academics with a simplified analysis of their multiple-choice question (MCQ) assessment results. *Journal of University Teaching & Learning Practice*, 4(2).
- De Champlain, A. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44, 109-117.
- Delgado, A., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, 14, 197-201.
- DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, 2(2), 1-23.
- Dolan, R. P., & Burling, K. S. (2012). Computer-based testing in higher education. In C. Secolksy & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 321-335). New York: Taylor & Francis.

- Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77(10), S103.
- 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. (2006a). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook for Test Development*. New Jersey: Lawrence Erlbaum Associates.
- Downing, S. M. (2006b). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 3-26). New Jersey: Lawrence Erlbaum Associates.
- Downing, S. M. (2006c). Written tests: constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in Health Professions Education* (pp. 149-184). New York: Routledge.
- Downing, S. M., Baranowski, R., Grosso, L., & Norcini, J. (1995). Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. *Applied Measurement in Education*, 8(2), 187-197.
- Downing, S. M., & Haladyna, T. M. (1997). Test item development: validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10(1), 61-82.

- Drezner, Z., Turek, O., & Zerom, D. (2010). A modified Kolmogorov-Smirnov test for normality. *Communication in Statistics-Simulation and Computation*, 39(4), 693-704.
- Edelstein, R. A., Reid, M. H., Usatine, R., & Wilkes, M. S. (2000). A comparative study of measures to evaluate medical students' performances. *Academic Medicine*, 75(8), 825-833.
- Engelhard, G., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education*, 12(2), 199-210.
- Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine*, 356, 387-396.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28(1), 1-11.
- Fares, J., Al Tabosh, H., Saadeddin, Z., El Mouhayyar, C., & Aridi, H. (2016). Stress, burnout and coping strategies in preclinical medical students. *North American Journal of Medical Sciences*, 8(2), 75-81.
- 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.
- Frisbie, D. A. (1973). Multiple-choice versus true-false: a comparison of reliabilites and concurrent validities. *Journal of Educational Measurement*, 10, 297-304.
- Frisbie, D. A. (1992). The status of multiple true-false testing. *Educational Measurement: Issues and Practice*, 5, 21-26.

- Frisbie, D. A., & Becker, D. F. (1991). An analysis of textbook advice about true-false tests. *Applied Measurement in Education, 4*, 67-83.
- Frisbie, D. A., & Druva, C. A. (1986). Estimating the reliability of multiple true-false tests. *Journal of Educational Measurement, 23*, 99-106.
- Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19*, 29-35.
- Fuchtgott, E. (1983). Emphasizing the negative: a note on “not” in multiple-choice questions. *Teaching of Psychology, 10*(1).
- Garrison, W. M., & Coggiola, D. C. (1980). *Practical procedures for test length reduction and item selection*. Department of Health, Education, and Welfare: National Technical Institute for the Deaf.
- Geisinger, K. F., Shaw, L. H., & McCormick, C. (2012). The validation of tests in higher education. In C. Secolksy & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 194-207). New York: Taylor & Francis.
- Gellman, E. S., & Berkowitz, M. (1993). Test-item type: what students prefer and why. *College Student Journal, 27*(1), 17-26.
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: a comprehensive review. *Review of Educational Research, 87*(6), 1082-1116.
- Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education, 47*(7), 726-7333.

- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items* (3rd ed.). New Jersey: Lawrence Erlbaum Associates.
- Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook for Test Development* (pp. 392-409). New York: Routledge.
- Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 27-50.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- 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), 309-333.
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York: Taylor & Francis.
- Harasym, P. H., Doran, M. L., Brant, R., & Lorscheider, F. L. (1993). Negation in stems of single-response multiple-choice items: an overestimation of student ability. *Evaluation & the Health Professions*, 16(3), 342-357.
- Harasym, P. H., Price, P. G., Brant, R., Violato, C., & Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions*, 15(2), 198-220.
- Harvill, L., & Davis, G. (1997). Medical students' reasons for changing answers on multiple-choice tests. *Academic Medicine*, 72(10), S97-A99.

- Haynes, S., Richard, D., & Kubany, E. (1995). Content validity in psychological assessment: a functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.
- Heinrich-Heine-Universität Düsseldorf. (2017). *G*Power 3.1 Manual*. Retrieved from http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf
- Herndon, C. (2006). Peer review and organizational learning: improving the assessment of student learning. *Research & Practice in Assessment, 1*, 8-13.
- Hill, G. C., & Woods, G. T. (1974). Multiple true-false questions. *Education in Chemistry, 11*, 86-87.
- Hoepfl, M. C. (1994). Developing and evaluating multiple choice tests. *Technology Teacher, 53*(7), 25-26.
- International Business Machines. (2017). SPSS Statistics.
- Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PLoS One, 8*(8), e70270.
- Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 131-153). New Jersey: Lawrence Erlbaum Associates.
- Kane, M. (2013). Validating the interpretations and use of test scores. *Journal of Educational Measurement, 50*, 1-73.
- Kar, S. S., Lakshminarayanan, S., & T, M. (2015). Basic principles of constructing multiple choice questions. *Indian Journal of Community & Family Medicine, 1*(2), 65-69.

- Karegar Maher, M. H., Barzegar, M., & Gasempour, M. (2016). The relationship between negative stem and taxonomy of multiple-choice questions in residency pre-board and board exams. *Research and Development in Medical Education*, 5(1), 32-35.
- Keppel, G., & Wickens, T. (2004). *Design and Analysis: A Researcher's Handbook* (4 ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
- Kniveton, B. H. (1996). Student perceptions of assessment methods. *Assessment & Evaluation in Higher Education*, 21(3), 229.
- Knupp, T., & Harris, D. J. (2012). Building content and statistical test specification. In C. Secolksy & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 239-256). New York: Routledge.
- Laerd Statistics. (2015). Two-way ANOVA using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from <https://statistics.laerd.com/>
- Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A.-P., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine*, 28(2), 166-173.
- Laprise, S. L. (2012). Afraid not: student performance versus perception based on exam question format. *College Teaching*, 60, 31-36.
- Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology & Society*, 14(4), 99-110.
- Leppink, J., van Gog, T., Paas, F., & Sweller, J. (2015). Cognitive load theory: researching and planning teaching to maximise learning. In J. Cleland & S. J.

- Durning (Eds.), *Researching Medical Education* (pp. 207-218). West Sussex, UK: Wiley Blackwell.
- Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*, 14-26.
- Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 421-444). New Jersey: Lawrence Erlbaum Associates.
- Malau-Aduli, B. S., Assenheimer, D., Choi-Lundberg, D., & Zimitat, C. (2014). Using computer-based technology to improve feedback to staff and students on MCQ assessments. *Innovations in Education and Teaching International*, *51*(5), 510-522.
- Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68-78.
- Mayo Clinic. (2018, 2018). Symptoms & causes. *Peripheral Neuropathy*. Retrieved from <https://www.mayoclinic.org/diseases-conditions/peripheral-neuropathy/symptoms-causes/syc-20352061>
- McCoach, D. B., Rambo, K. E., & Welsh, M. (2012). Issues in the analysis of change. In C. Secolksy & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 229-236). New York: Taylor & Francis.
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, *26*(8), 709-712.

- Membership of the Royal Colleges of Physicians of the United Kingdom. (2018). Question Writers. Retrieved from <https://www.mrcpuk.org/get-involved-examiners/question-writers>
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.
- Millman, J., & Bishop, C. H. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707-726.
- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). New Jersey: Lawrence Erlbaum Associates.
- Moreno, R., Martinez, R. J., & Muniz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, *2*(2), 65-72.
- Moreno, R., Martinez, R. J., & Muniz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, *27*(4), 388-394.
- National Board of Medical Examiners. (2018). About the NBME. Retrieved from <http://www.nbme.org/about/index.html>
- Odegard, T. N., & Kown, 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.
- Osterlind, S. J. (1998). *Constructing Test Items: Multiple-choice, Constructed Response, Performance, and Other Formats*. Boston: Kluwer Academic.
- Osterlind, S. J., & Wang, Z. (2012). Item response theory in measurement, assessment, and evaluation for higher education. In C. Secolksy & D. B. Denison (Eds.),

Handbook on Measurement, Assessment, and Evaluation in Higher Education
(pp. 150-160). New York: Taylor & Francis.

Paniagua, M. A., & Swygert, K. A. (2016). *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners.

Pass, F., Renkl, A., & Sweller, J. (2004). Cognitive load: instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science*, 32(1/2), 1-8.

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 259-273). New York: Routledge.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": an exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, 23(4), 349-371.

Schaughency, E., Smith, J. K., van der Meer, J., & Berg, D. (2012). Classical test theory and higher education: five questions. In C. Secolksy & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 117-131). New York: Taylor & Francis.

Sievertsen, H. H., Gino, F., & Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences of the United States of America*, 113(10), 2621-2624.

doi:10.1073/pnas.1516947113

- Sim, J. H., Wen, T. T., Wei-Han, H., Vadivelu, J., & Hassan, H. (2015). Development of an instrument to measure medical students' perceptions of the assessment environment: initial validation. *Medical Education Online, 20*.
- Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher, 28*(6), 566-568.
- Stoffel, H., Raymond, M. R., Bucak, S. D., & Haist, S. A. (2014). Editorial changes and item performance: implications for calibration and pretesting. *Practical Assessment, Research & Evaluation, 19*(14), 1-11.
- Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cognitive Science, 12*, 257-285.
- Tarrant, M., Knierim, A., Hayes, S., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*, 662-671.
- 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*, 539-543.
- The Psychometric Society. (2017). What is Psychometrics? Retrieved from <https://www.psychometricsociety.org/content/what-psychometrics>
- Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: the distractors are also part of the item. *Journal of Educational Measurement, 26*(2), 161-176.

- Thompson, B., & Vacha-Haase, T. (2012). Reliability. In C. Secolksy & D. B. Denison (Eds.), *Handbook on Measurement, Assessment, and Evaluation in Higher Education* (pp. 178-193). New York: Routledge.
- Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education* (8th ed.). Boston: Pearson Education, Inc.
- Towns, M. H. (2014). Guide to developing high-quality, reliable, and valid multiple-choice assessments. *Journal of Chemical Education*, *91*(9), 1426-1431.
- United States Medical Licensing Examination. (2018). Retrieved from <http://www.usmle.org/>
- Vahalia, K. V., Subramaniam, S. C., Marks, S. C., & De Souza, E. J. (1995). The use of multiple-choice tests in anatomy: common pitfalls and how to avoid them. *Clinical Anatomy*, *8*(61), 61-65.
- Vale, C. D. (2006). Computerized item banking. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 261-286). New Jersey: Lawrence Erlbaum Associates.
- Vyas, R., & Supe, A. (2008). Multiple choice questions: a literature review on the optimal number of options. *National Medication Journal of India*, *21*, 130-133.
- Ware, J., & Torstein, V. (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.
- Weiner, W. F. (2005). Establishing a culture of assessment. *Academe*, *95*(4).
- Wise, S. J., Finney, S. J., Enders, C. K., Freeman, S. A., & Severance, D. D. (1999). Examinee judgments of changes in item difficulty: implications for item review in

computerized adaptive testing. *Applied Measurement in Education*, 12(2), 185-198.

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education*, 14, 465-473.

Zakay, D., & Glicksohn, J. (1992). Overconfidence in a multiple-choice test and its relationship to achievement. *The Psychological Record*, 42, 519-524.

APPENDIX A: EXAMPLES OF STRUCTURAL FLAWS

A defect in the spiral aorticopulmonary septum is LEAST likely to result in which of the following?

Seq	Answer Choice	Correct	Lock
a)	Ventral septal defect	<input type="checkbox"/>	<input type="checkbox"/>
b)	Transposition of great vessels	<input type="checkbox"/>	<input type="checkbox"/>
c)	Atrial septal defect	<input checked="" type="checkbox"/>	<input type="checkbox"/>
d)	Persistent truncus arteriosus	<input type="checkbox"/>	<input type="checkbox"/>
e)	Pulmonary stenosis	<input type="checkbox"/>	<input type="checkbox"/>

Which of the following is a true statement?

Seq	Answer Choice	Correct	Lock
a)	Conjugated bile salts are less water-soluble than non-conjugated bile salts.	<input type="checkbox"/>	<input type="checkbox"/>
b)	Primary bile acids are formed in the liver from cholesterol.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
c)	Secondary bile acids are formed by bacteria in the stomach.	<input type="checkbox"/>	<input type="checkbox"/>
d)	Secondary bile acids are all excreted in the feces.	<input type="checkbox"/>	<input type="checkbox"/>

APPENDIX B: ANOVA RESULTS TABLES

Tests of Between-Subjects Effects

Dependent Variable: Diff Delta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	.013 ^a	3	.004	.348	.791	.018	1.044	.114
Intercept	.012	1	.012	.993	.323	.017	.993	.165
Flaw	.001	1	.001	.057	.812	.001	.057	.056
Disposition	5.446E-7	1	5.446E-7	.000	.995	.000	.000	.050
Flaw * Disposition	.011	1	.011	.903	.346	.016	.903	.154
Error	.702	57	.012					
Total	.765	61						
Corrected Total	.714	60						

a. R Squared = .018 (Adjusted R Squared = -.034)

b. Computed using alpha = .05

Tests of Between-Subjects Effects

Dependent Variable: Disc Delta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	.015 ^a	3	.005	.239	.869	.012	.718	.093
Intercept	2.778E-5	1	2.778E-5	.001	.971	.000	.001	.050
Flaw	8.850E-5	1	8.850E-5	.004	.949	.000	.004	.050
Disposition	.001	1	.001	.064	.801	.001	.064	.057
Flaw * Disposition	.013	1	.013	.603	.441	.010	.603	.119
Error	1.264	59	.021					
Total	1.288	63						
Corrected Total	1.279	62						

a. R Squared = .012 (Adjusted R Squared = -.038)

b. Computed using alpha = .05

Tests of Between-Subjects Effects

Dependent Variable: AAT Delta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	17.176 ^a	3	5.725	.026	.994	.001	.077	.054
Intercept	23.563	1	23.563	.106	.746	.002	.106	.062
Flaw	.043	1	.043	.000	.989	.000	.000	.050
Disposition	11.616	1	11.616	.052	.820	.001	.052	.056
Flaw * Disposition	.043	1	.043	.000	.989	.000	.000	.050
Error	12859.033	58	221.707					
Total	12925.000	62						
Corrected Total	12876.210	61						

a. R Squared = .001 (Adjusted R Squared = -.050)

b. Computed using alpha = .05