Differences in Item Statistics Between Positively and Negatively Worded Stems on Histology Examinations

Sara Klender, BS, Andrew Ferriby, BS, and Andrew Notebaert, PhD.
University of Mississippi Medical Center, Department of Neurobiology and Anatomical Sciences, Jackson, MS 39216
sklender@umc.edu, aferriby@umc.edu, anotebaert@umc.edu

Abstract
Multiple-choice questions (MCQ) are commonly used on histology examinations. There are many guidelines for how to properly write MCQ and many of them recommend avoiding negatively worded stems. The current study aims to investigate differences between positively and negatively worded stems in a medical histology course by comparing the item difficulty and discrimination index between matched MCQs. When questions were matched by modified Bloom's Taxonomy classification, presence or absence of an image, and timing of content presentation, negatively worded lower level Bloom's questions were less difficult and had a lower discrimination index.

Key Words: negative questions, multiple-choice questions, item writing flaws, stem orientation, histology

Introduction
Item Writing Flaws and Negative Questions

Multiple choice question (MCQ) based examinations are a popular form of student assessment that is common among undergraduate and professional school programs. Some of the perceived benefits of using MCQs have been detailed in Table 1. There are a number of guidelines for properly writing MCQs. Haladyna, Downing, and Rodriguez (2002) have complied a list of 31 item-writing recommendations from educational measurement textbooks. These recommendations include:

1. Avoid testing on trivial content.
2. Format items vertically instead of horizontally.
3. Use correct grammar, punctuation, and spelling.
4. Include the central idea in the stem and not in the answer choices.
5. Ensure there is only one correct answer.
6. Keep answer choices independent from one another.
7. Avoid questions with "all-of the-above" as an answer choice.

The National Board of Medical Examiners (NBME), which is responsible for writing items for the United States Medical Licensing Examination® (USMLE®) three-step examinations, has also published MCQ guidelines (Case and Swanson 2002). The NBME addresses issues related to "testwiseness" such as the inclusion of grammatical cues, logical cues, using absolute terms like always or never, and having the correct answer include the most elements in common with the other options. The NBME also cautions item writers to avoid issues related to irrelevant difficulty, which includes using long, complicated answer options, using vague frequency terms like sometimes and rarely, and using questions with "none-of-the-above" as an answer choice (Case and Swanson 2002).

continued on next page

Benefits of using MCQs	Citation
MCQ grading is objective.	Butler 2018
MCQs allow the exam to cover more content due to the small amount of time needed to respond to each question.	Butler 2018
Students' perceived anxiety is reduced.	Snow 1993
MCQs allow students to receive feedback sooner, possibly even immediately after an examination by utilizing electronic examinations.	Delgado and Prieto 2003; Epstein and Brosvic 2002
The grading process for MCQs is efficient and accurate, even for large number of exam takers.	Dufresne et al. 2002; Walstad and Becker 1994

Table 1. *Perceived benefits of using MCQ examinations.*

Questions that fail to adhere to item-writing guidelines can be described as having item-writing flaws (IWFs). When using a list of 15 item-writing guidelines, a medical school found that 8% of their examination questions suffered from IWFs (Ware and Vik 2009). Another medical school using the 31 guidelines compiled by Halayna et al. (2002), discovered that as many as 36-65% of MCQ on a series of four examinations from different disciplines were flawed (Downing 2005). Similarly, Tarrant and Ware (2008) found that 47% of items were flawed on ten high stakes nursing examinations. These high rates of IWF are alarming and have also been shown to affect student examination performance. For instance, Downing (2005) found that because of the IWFs, as many as 10-15% of the students may have been incorrectly classified as having "failed". On the other hand, Tarrant and Ware (2008) reported that students on the borderline of passing or failing performed better on an examination when flawed items were present compared to when the flawed items were removed. In both of these studies, flawed items may have led to unfair and inaccurate assessment of students' learning.

An item-writing guideline that continues to be debated is the use of negatively worded questions. These are question stems that include the words "not", "except", or "incorrect". The NBME recommends against using these questions because the answer choices cannot easily be ranked on a continuum. This increases the difficulty and inhibits the examinee's ability to rank answer choices as "most" or "least" correct (Case and Swanson 2002). Others agree that negatively worded question stems should be avoided since they make the question unnecessarily difficult and confusing by forcing the examinee to change their tactics from finding the correct answer to finding the incorrect answer (Boland et al. 2010; Smith 2018).

The recommendation to avoid negative questions has been supported by a number of studies. Harasym et al. (1992) found that when comparing single response negatively worded and multiple response positively worded questions, multiple response positively worded questions were a more reliable and valid method of assessing student achievement. Additionally, negative questions have been associated with a lower Bloom's Taxonomy level, meaning they do not require the examinee to use higher cognitive functions (Maher et al. 2016).

Other studies have suggested that there is no harm in utilizing negative questions if item-writers highlight, bold, or underline the negation (Haladyna et al. 2002). Results in an early study on this topic found no difference in item difficulty or discrimination index when comparing negatively and positively worded question stems (Violato and Marini 1989). These results were confirmed by Caldwell and Pate (2013), which found that while negatively worded questions had higher item difficulty the difference did not reach statistical significance. There was also no significant difference in the discrimination index of the positively and negatively worded questions, however, the study compared only five pairs of negatively and positively worded questions (Caldwell and Pate 2013).

continued on next page

Assessing MCQs

There are a number of ways to assess MCQs. The current study will consider item difficulty, discrimination index (DI) values, and Blooms Taxonomy categorizations. Item difficulty can be defined as the percentage of the examinees that answered the question correctly. Thus, the higher the percentage of students who answered the question correctly, the "easier" the question is judged to be (Ebel and Frisbie 1991). Discrimination index is the difference between the percentage of correct responses from the upper and lower performers, which indicates an item's capacity to differentiate between high scorers and low scorers on an examination (Rush, Rankin, and White 2016). Therefore, a high discrimination index value indicates that the upper performers did better on the item compared to the lower performers. Different values can be used to define upper and lower performers, but often the highest and lowest quartiles are used (Rush, Rankin, and White 2016). A high quality item should be of appropriate difficulty for the students being assessed and should have the capacity to discriminate between students.

Bloom's Taxonomy is a tool created to assess the cognitive functions and level of reasoning needed to answer a question. The original version consisted of six different levels: the lowest being Knowledge, the levels increase to Comprehension, Application, Analysis, Synthesis, and Evaluation (Blooms 1956). The taxonomy has undergone several revisions since its inception; the first rewording the dimensions to fall under two more generalized categories of "Knowledge and Cognitive Process" (Krathwohl 2002). While the original taxonomy had six dimensions, some claim that the higher levels (Synthesis and Evaluation) cannot be assessed using MCQ (Crowe et al. 2008; Huxham and Naeraa 1980).

Some researchers suggest the original Bloom's Taxonomy tool may be too general for specific academic disciplines and certain levels of instruction (Hussey and Smith 2002). This limitation has led to the creation of discipline specific Bloom's Taxonomy tools. One such adaptation is the Bloom's Taxonomy Histology Tool (BTHT), which allows for evaluators to properly categorize histology related MCQs. It also places a larger focus on questions involving images due to the visual nature of histology (Zaidi et al. 2017).

Purpose

Despite conflicting recommendations, negative questions continue to be utilized. Downing (2005) found that negatively worded questions were the second most common IWF on medical school examinations. Other medical schools have reported that 7-23% of examination questions were negatively written (Maher et al. 2016; Ware and Vik 2009). To investigate the appropriateness of using negatively worded question stems, this study aims to compare the item difficulty and discrimination index of positively and negatively worded questions. Our first hypothesis was that compared to positively worded questions stems, negatively worded question stems will be more difficult, meaning that fewer students will correctly answer these questions. Our second hypothesis was that negatively worded stems would have a lower discrimination index value compared to positively worded stems.

Methods

Context

The University of Mississippi Medical Center (UMMC) is a large academic medical center which educates future healthcare providers within the schools of medicine, dentistry, pharmacy, nursing, allied health science, and graduate studies. The medical school is the state's only allopathic program and typically accepts only in state residents.

At the time of this study, the medical curriculum included two years of basic science courses followed by two years of clinical rotations. First year medical students took Gross Anatomy, Histology and Cell Biology, Developmental Anatomy, Physiology, Neuroscience, Biochemistry, and Introduction to the Medical Profession. Medical Histology and Cell Biology was taught as a stand-alone course throughout the entire fall semester and half of the spring semester. The course consisted of six credit hours and was divided into seven blocks.

Medical Histology and Cell Biology included both lecture and laboratory components. Fifty-minute lectures given by anatomy faculty covered basic histology content and were not mandatory for students to attend. At the end of each of these lectures, a five-question multiple-choice bonus quiz was given using TurningPoint electronic polling software (www.turningtechnologies.com). At the end of each block, each student's average bonus quiz score was added to their written examination grade as percentage points. Each block also had one clinical correlation lecture, presented by a physician, which connected the basic science content to clinical practice. These lectures were mandatory and did not include a bonus quiz. For laboratory sessions, students were divided into two groups. Each group had an hour and half in the laboratory to work with a partner to identify cells, tissues, and organs using light microscopy and electron microscope images. Students were given a guide for each laboratory session that indicated the structures they should identify. Anatomy faculty members and teaching assistants were also available to answer questions. At the end of each lab students were given a bonus quiz that consisted of five light microscopy questions and one electron microscopy identification question. Each question on the laboratory bonus quiz was fill-in-the-blank style. At the end of each block, each student's average bonus lab quiz score was added to their practical examination grade as percentage points.

continued on next page

Seven lecture examinations, three practical examinations, and the NBME Histology and Cell Biology Subject Examination were used to determined grades. Lecture exams consisted of 30-36 multiple-choice questions administered using ExamSoft® software (www.examsoft.com). Five answer options were provided for positively worded stems and four were provided for negatively worded stems. In negatively worded questions, the negation was capitalized and bolded. Two clinical vignette questions were also included as extra credit. The practical examination consisted of approximately 60 fill-in-the-blank style questions that required students to identify cells, tissues, and organs using light microscopy and electron microscope images. Students had one minute at each station to view the slide or image and then they rotated to the next station. The average of the seven lecture examinations was worth 45% of the final grade, the average of the three practical examinations was 45%, and the NBME Subject Examination was 10%. Table 2 demonstrates how the average lecture and practical exam grades were calculated.

Block	Topics	% of Lecture Grade	% of Practical Grade
1	Cellular and Molecular Biology	15.0	
2	Basic Tissues 1	15.0	33.87
3	Basic Tissues 2	15.0	
4	Digestive System and Immune	15.0	32.26
5	Blood, Ear, and Eye	12.5	
6	Cardiovascular, Urinary, Endocrine, and Cell Division	12.5	33.87
7	Reproductive and Respiratory	15.0	

Table 2. *Contribution of each exam to the lecture and practical exam grades.*

Participants

At the time of the study there were a total of 163 students in the Medical Histology and Cell Biology course. However, the number of students taking the exam using the ExamSoft® software varied between exams due to personal computer issues or absences (Table 3). Students who had computer issues or absences took the exam on paper and were excluded from our calculations.

Exam	# of Exam Takers (n)
1	163
2	162
3	163
4	162
5	162
6	157
7	150

Table 3. *Number of ExamSoft® exam takers at each block.*

continued on next page

Materials

Each Histology and Cell Biology question was rated using a modified Bloom's Taxonomy tool (Table 4). This tool was based on the frequently used Revised Bloom's Taxonomy (Krathwohl 2002). Material from the Bloom's Taxonomy Histology Tool (Zaidi et al. 2017) was also incorporated since it addresses some discipline specific information that is pertinent to histology MCQs. This tool was used by two authors who were anatomy graduate students with prior histology experience (SK and AF). The raters were instructed to make judgements based on the assumption that images and questions were novel to students. Raters were required to assign each item to the Remember, Understand, Apply, Analyze, or Evaluate category. To ensure the tool was being used in a similar manner by both raters, inter-rater reliability was determined using Kohen's Cappa and a moderate level of agreement was found (Table 5). Questions that were classified differently by the two raters were discussed and agreed upon using the modified Bloom's Taxonomy criteria.

Remember (1)	Recall facts and basic concepts (ex. recall, define, memorize) (Krathwohl 2002).
Understand (2)	Explain ideas or concepts, without relating to anything else (ex. classify, identify, locate) (Krathwohl 2002). "Requires recall and com- prehension of facts. Image questions asking to identify a structure/cell type without requiring a full understanding of the relationship of all parts" (Zaidi 2017).
Apply (3)	Use information in new situations (ex. apply, implement, use) (Krathwohl,2002). "Two-step questions that require image-based identification as well as the application of knowledge (e.g., identify structure and know function/ purpose)" (Zaidi 2017).
Analyze (4)	Draw connections among ideas (ex. organize, analyze, calculate, compare, contrast, attribute) (Krathwohl 2002). "Students must call upon multiple independent facts and properly join them together." (Zaidi 2017).
Evaluate (5)	Justify a decision (ex. critique, judge, predict, appraise) (Krathwohl 2002).

Table 4. *Modified Bloom's Taxonomy criteria used to classify histology questions.*

	Value	Asymptotic Standard Error[a]	Approximate T[b]	Significance
Cohen's Kappa	.520	.041	12.914	.000

Table 5. *Inter-rater reliability for classification of questions using the modified Bloom's Taxonomy criteria. a: Not assuming the null hypothesis. b: Using the asymptotic standard error assuming the null hypothesis.*

continued on next page

Procedure

Lecture examination questions from the 2017 - 2018 school year were retrospectively analyzed. A total of 240 items were classified based on several criteria:

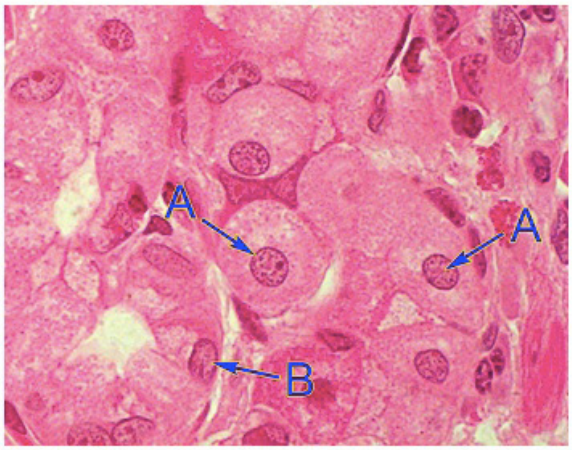
1. The designation of a negative or positive stem.
2. The modified Bloom's Taxonomy rating.
3. The presence or absence of an image relating to the question.
4. The timing of the content presentation.

Question stems using negative phrases such as "Which of the following is **NOT**…", "All of the following **EXCEPT**…", or "Identify the **FALSE** statement" were classified as negatively worded, while all others were classified as positively worded. Next, every question was rated using the modified Bloom's Taxonomy tool.

In order to match each negative question, the authors first found positively worded question stems with the same modified Bloom's Taxonomy rating. From this group of positively worded questions, the negatively worded question was then matched with a positively worded question based on the presence or absence of an image. Finally, the negative question was matched based on timing of content presentation. Ideally, content tested in the positive and negative matched questions was presented during the same lecture. If not, the authors moved to a positive question that tested content presented in the same exam block. An example of a matched positive and negative question can be seen in Figure 1.

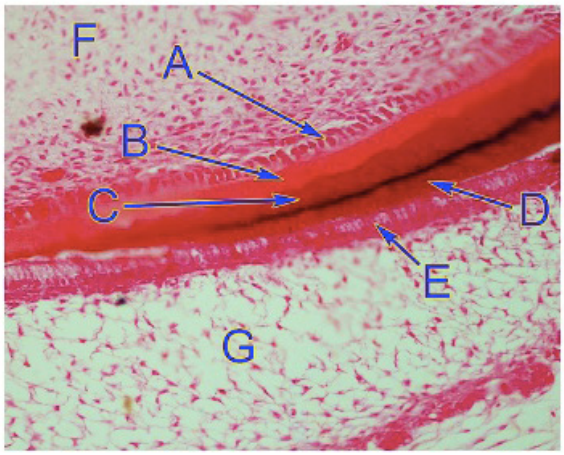
Item difficulty and discrimination index values were obtained from the ExamSoft® post-test summary report. Item difficulty was calculated as the number of students who correctly answered the item divided by the total number of exam takers. Discrimination index was calculated by subtracting the item difficulty of the lower 27% of the class from the item difficulty of the upper 27% of the class.

The cells whose nuclei are indicated by the letter **A**:



A. are located in the crypts of Lieberkühn
B. are located within the villi of the small intestine
C. secrete gastrin to increase gastric motility
D. secrete motilin necessary for the absorption of vitamin B_{12}
E. utilize CO_2 and H_2O to produce hydrochloric acid

Which of the following is **NOT** true of the layer labeled **A**?



A. This layer is maintained in adult teeth.
B. This layer originates from neural crest cells.
C. This layer produces pre-enamel that is mineralized to make enamel.
D. This layer sends cell processes into dentinal tubules.

Figure 1. *Example of matched positive and negative questions. These items were matched based on their modified Bloom's Taxonomy rating (Apply), the presence of image, and the timing of content presentation (concepts were presented during the digestive system lecture).*

continued on next page

Analysis

Descriptive statistics were used to summarize question characteristics. Paired t-tests were conducted in order to compare the item difficulty of matched positive and negative questions. To compare discrimination index values, questions were categorized based on the criteria recommended by Roa et al. (2016). This criteria categorizes items with DI values between 0.00-0.19 as "poor", 0.20-0.29 as "acceptable", 0.30-0.39 as "good", and >0.40 as "excellent". Fisher's exact tests were then used to compared the likelihood of positive or negative questions having a better DI in the lower (Remember and Understand) and higher (Apply and Analyze) Bloom's Taxonomy categories. All statistical analyses were completed using SPSS version 24 (IBM Inc., Armonk, NY) with a significance level of $p < 0.05$.

Results

The majority of questions were classified as Remember (54%) and none were classified as Evaluate. Frequency counts and mean item difficulty of each category is shown in Table 6.

	Frequency, n(%)	Item Difficulty (M±SD)
Bloom's Taxonomy Rating		
Remember	129 (53.8)	78.74 ± 17.1
Understand	30 (12.5)	76.20 ± 18.6
Apply	50 (20.8)	74.04 ± 18.0
Analyze	31 (12.9)	75.39 ± 15.3
Evaluate	0 (0.0)	-
Image		
Yes	92 (38.3)	77.99 ± 17.5
No	148 (61.7)	75.43 ± 16.9

Table 6. *Frequency and mean item difficulty of all questions by modified Bloom's Taxonomy classification and presence or absence of an image.*

Of the 240 questions on the seven histology lecture exams, 27 (11.3%) were classified as negatively worded. Each negative question was matched with a positive question based on Bloom's Taxonomy rating, presence of an image, and timing of content delivery. Only one negative question required matching from a different exam, while all other negative questions were successfully matched with positive questions that tested content presented in the same exam block. Table 7 shows the mean item difficulty of positively and negatively worded questions in each Bloom's Taxonomy rating.

In order to address our first hypothesis, which stated that negatively worded questions would be more difficult, paired t-tests were conducted to compare item difficulty of the matched positive and negative questions. There was no significant difference in difficulty found between 27 pairs of matched positive and negative questions; t (26) = .884, p=.385. However, when comparing difficulty of positive and negative questions in each Bloom's Taxonomy category, there was a significant difference in the Remember category; t (15) = 2.258, p=.039. No significant differences were found between positive and negative questions in the Understand category; t (3) = -.545, p=.624 or the Apply category; t (5) = -.037, p=.972. Paired t-test could not be conducted to compare positive and negative questions in the last two categories (Analyze and Evaluate) because there were too few questions or no pairs of questions to analyze.

continued on next page

Bloom's Taxonomy Rating	Pairs, n(%)	Positive Item Difficulty (M±SD)	Negative Item Difficulty (M±SD)
Remember	16 (59.3)	72.38 ± 16.5*	83.94 ± 17.2*
Understand	4 (14.8)	81.75 ± 5.1	73.00 ± 31.7
Apply	6 (22.2)	71.33 ± 19.6	71.00 ± 12.5
Analyze	1 (3.7)	98.00	61.00
Evaluate	0 (0.0)	-	-
Total	27 (100.0)	74.48 ± 27.3	78.59 ± 23.6

Table 7. *Mean item difficulty of positive and negative questions by Bloom's Taxonomy rating. *Significant difference at the 0.05 level.*

In order to address the second hypothesis, which stated that negatively worded stems would have a lower discrimination index value compared to positively worded stems, each of the 27 paired questions were placed into a discrimination index category. Table 8 details the discrimination index categories for each pair and which question of the pair received the better DI category. The Fisher's exact test was then used to compare the likelihood of positive or negative questions having a better DI in the lower (Remember and Understand) and higher (Apply and Analyze) Bloom's Taxonomy categories. Fisher's exact test shows no significant association between Bloom's Taxonomy level and frequency of pairs having a better positive question DI or having equal DI ($p = 1.000$). Likewise, there was no significant association between Bloom's Taxonomy level and frequency of pairs having a better negative question DI or equal DI ($p = 0.138$). However, there was a significant association between Bloom's Taxonomy level and frequency of pairs having a better positive question DI or better negative question DI ($p = 0.014$), with positively worded questions tending to have a better DI category when written at a lower Bloom's Taxonomy level.

continued on next page

Bloom's Taxonomy	Pair	Positive DI	Positive Category	Negative DI	Negative Category	Better DI
Remember	1	0.44	Excellent	0.23	Acceptable	Positive
	2	0.31	Good	0.21	Acceptable	Positive
	3	0.36	Good	0.20	Acceptable	Positive
	4	0.21	Acceptable	0.23	Acceptable	Same
	5	0.61	Excellent	0.33	Good	Positive
	6	0.25	Acceptable	0.05	Poor	Positive
	7	0.05	Poor	0.11	Poor	Same
	8	0.53	Excellent	0.00	Poor	Positive
	9	0.26	Acceptable	0.41	Excellent	Negative
	10	0.04	Poor	0.02	Poor	Same
	11	0.27	Acceptable	0.09	Poor	Positive
	12	0.16	Poor	0.29	Acceptable	Negative
	13	0.01	Poor	0.15	Poor	Same
	14	0.55	Excellent	0.29	Acceptable	Positive
	15	0.19	Poor	0.21	Acceptable	Negative
	16	0.30	Good	0.17	Poor	Positive
Understand	17	0.27	Acceptable	0.00	Poor	Positive
	18	0.36	Good	0.29	Acceptable	Positive
	19	0.25	Acceptable	0.24	Acceptable	Same
	20	0.28	Acceptable	0.09	Poor	Positive
Apply	21	0.10	Poor	0.34	Good	Negative
	22	0.49	Excellent	0.36	Good	Positive
	23	0.40	Excellent	0.48	Excellent	Same
	24	0.16	Poor	0.42	Excellent	Negative
	25	0.14	Poor	0.28	Acceptable	Negative
	26	0.35	Good	0.42	Excellent	Negative
Analyze	27	0.03	Poor	0.45	Excellent	Negative

Table 8. *Discrimination index of paired positively and negatively worded questions by modified Bloom's Taxonomy rating. The final column indicates whether the positively or negatively worded question of each pair had a better DI. DI = discrimination index.*

	Better Positive DI (n)	Same DI (n)	Better Negative DI (n)
Lower Blooms	12	5	3
Higher Blooms	1	1	5

Table 9. *Frequency of pairs with the positively worded question having a better DI, pairs with negatively worded question having a better DI, and pairs with equal DI categories; separated by lower (Remember and Understand) and higher (Apply and Analyze) Bloom's Taxonomy categories. For example, of the lower Bloom's Taxonomy pairs, in 12 of the 20 pairs the positively worded question had a better DI. DI = discrimination index.*

continued on next page

Discussion

There is still some debate about whether negative questions are appropriate to use on MCQ examinations (Haladyna et al. 2002). In order to investigate this issue we compared the item difficulty and discrimination index of 27 pairs of positively and negatively worded question stems. Our first hypothesis was that compared to positively worded questions, negatively worded questions would be more difficult. This hypothesis was not supported by our data. When comparing the mean item difficulty of all positive and negative questions there were no significant differences found. However, when comparing matched questions from the same Bloom's Taxonomy category, there was a significant difference in item difficulty between positively and negatively worded questions in the Remember category, with negative questions being less difficult. The low level Remember questions consist mostly of true or false fact answer options. The lack of complexity in these options may limit the variety of distractors in these questions, making it easier for students to identify the false answer option.

These results contradict the findings of Caldwell and Pate (2013) and Violato and Marini (1989) that claimed there was no significant difference in item difficulty of positively and negatively worded stems on pharmacy or undergraduate MCQ examinations. However, these studies did not compare positively and negatively worded questions within Bloom's Taxonomy levels and therefore potential differences at certain levels may have been overlooked.

Our data also revealed a trend that shows as the Bloom's Taxonomy classification of negative questions becomes higher; fewer numbers of students answered the question correctly. A future investigation may consider analyzing a larger sample of negative questions in order to see if this trend persists.

Our second hypothesis was that negatively worded stems would have a lower discrimination index value compared to positively worded stems. This hypothesis was only somewhat supported by our data. At the lower level Bloom's Taxonomy questions there were significantly more pairs of questions where the positive question had a better DI categorization compared to the negative question. Once again, these results differ from the findings of Caldwell and Pate (2013) and Violato and Marini (1989).

These results indicate that negatively worded questions written at a lower level were easier for students, since more students correctly answered these questions, and tended to have a lower DI as compared to matched positively worded questions. Therefore, negative questions at a lower Bloom's level may not be as effective at differentiating between high and low performing students. For this reason, our results seem to support the widely held belief that negative question stems should be avoided when possible, particularly when evaluating lower levels of knowledge based on Bloom's Taxonomy (Boland et al. 2010; Harasym et al. 1992; Smith 2018; Xu et al. 2016).

There are several limitations to the current study. Because the questions were analyzed retrospectively, we were only able to control for a limited number of variables; specifically Bloom's Taxonomy classification, the presence or absence of an image, and the timing of content presentation. Future studies may consider matching questions prior to administration of the exam in order to make questions identical in all ways except the negation of the stem. Secondly, due to small numbers of questions categorized at higher Bloom's taxonomy levels, trends in this data may have been overlooked. Future studies should aim to include a sufficient number of pairs of positive and negative questions at each Bloom's taxonomy level. Finally, due to faculty members' perceived difficulty of negatively worded questions, only four answer options were provided for these questions while positively worded questions had five answer options. Once again, future studies should consider prospectively designing questions in order to create questions that are similar in all ways except stem negation.

The current study found that compared to positively worded questions at a low Bloom's level, negatively worded questions of the same level are significantly easier and have a lower DI category. This suggests that negative questions may not be ideal for differentiating between high and low performing students, particularly on lower level Bloom's questions. Based on these results, instructors may want to limit their use of negatively worded questions on MCQ examinations. However, further research is needed to confirm these preliminary findings.

About the Authors

Sara Klender, BS, is a fourth-year graduate student in the Clinical Anatomy program at the University of Mississippi Medical Center (UMMC). Her dissertation research focuses on the relationship between fear of death and performance in gross anatomy.

Andrew Ferriby, BS, is a second-year graduate student in the Clinical Anatomy program at UMMC. His current research focuses on the relationship between students' anatomical self-efficacy and burnout.

Andrew Notebaert, PhD, is an Assistant Professor and Program Director for the Clinical Anatomy program at UMMC. He teaches education courses to graduate students and conducts research on student perceptions

continued on next page

References

Bloom, BS. 1956. Taxonomy of Educational Objectives. Vol. 1: Cognitive Domain. New York (NY): McKay.

Boland, RJ, Lester, NA, & Williams, E. 2010. Writing multiple-choice questions. *Acad Psychiatry*. 34(4):310-316.

Butler, AC. 2018. Multiple-choice testing in education: Are the best practices for assessment also good for learning?. *J Appl Res Mem Cogn*. 7(3):323-331.

Caldwell, DJ and Pate, AN. 2013. Effects of question formats on student and item performance. *Am J Pharm Educ*. 77(4):71.

Case, SM and Swanson, DB. 2002. Constructing Written Test Questions for the Basic and Clinical Sciences. 3rd Edition. Philadelphia (PA): National Board of Medical Examiners.

Crowe, A, Dirks, C and Wenderoth, MP. 2008. Biology in bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE Life Sci Educ.* 7(4):368-381.

Delgado, AR and Prieto, G. 2003. The effect of item feedback on multiple-choice test responses. *Br J Psychol.* 94(1):73-85.

Downing, SM. 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. *Adv Health Sci Educ.* 10(2):133-143.

Dufresne, RJ, Leonard, WJ and Gerace, WJ. 2002. Marking sense of students' answers to multiple-choice questions. *The Phys Teach*. 40(3):174-180.

Ebel, RL and Frisbie, DA. 1991. Essentials of Educational Measurement. 5th Edition. Prentice-Hall: Englewood Cliffs.

Epstein, ML and Brosvic, GM. 2002. Immediate feedback assessment technique: Multiple-choice test that "behaves" like an essay examination. *Psychol Rep*. 90(1):226.

Haladyna, TM, Downing, SM and Rodriguez, MC. 2002. A review of multiple choice item-writing guidelines for classroom assessment. *Applied Measurement in Education.* 15(3):309-334.

Harasym, PH, Price, PG, Brant, R, Violato, C and Lorscheider, FL. 1992. Evaluation of negation in stems of multiple-choice items. *Eval Health Prof*. 15(2):198-220.

Hussey, Tand Smith, P. 2002. The trouble with learning outcomes. *Active Learning in Higher Education*. 3(3):220-233.

Huxham, GJ and Naeraa, N. 1980. Is Bloom's taxonomy reflected in the response pattern to MCQ items? *Med Educ*. 14(1):23-26.

Krathwohl, DR. 2002. A revision of Bloom's taxonomy: An overview. *Theory Pract*. 41(4):212-218.

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

Rao, C, Kishan Prasad, HK, Sajitha, K, Permi, H and Shetty, J. 2016. Item analysis of multiple choice questions: Assessing an assessment tool in medical students. *International Journal of Educational and Psychological Researches*. 2(4):201-204.

Rush, BR, Rankin, DC and White, BJ. 2016. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med Educ.* 16(1):250.

Smith, L.S. 2018. How to write better multiple-choice questions. *Nursing2018.* 48(11):14-17.

Snow, RE. 1993. Construct validity and constructed-response tests. In: Bennet, RE, Ward, WC, editors. Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment. New York (NY): Routledge. p. 45-60.

Tarrant, M and Ware, J. 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med Educ*. 42(2):198-206.

Violato, C and Marini, AE. 1989. Effects of stem orientation and completeness of multiple-choice items on item difficulty and discrimination. *Educ Psychol Meas*. 49(1):287-295.

Walstad, WB and Becker, WE. 1994. Achievement differences on multiple-choice and essay tests in economics. *Am Econ Rev.* 84(2):193-196.

Ware, J and Vik, T. 2009. Quality assurance of item writing: during the introduction of multiple choice questions in medicine for high stakes examinations. *Med Teach*. 31(3): 238-243.

Xu, X, Kauer, S and Tupy, S. 2016. Multiple-choice questions: Tips for optimizing assessment in-seat and online. *Scholarsh Teach Learn Psychol*. 2(2):147-158.

Zaidi, NB, Hwang, C, Scott, S, Stallard, S, Purkiss, J and Hortsch, M. 2017. Climbing Bloom's taxonomy pyramid: Lessons from a graduate histology course. *Anat Sci Educ.* 10(5):456-464.

Back to TOC