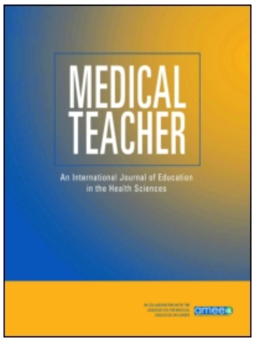

MEDICAL
TEACHER
An International Journal of Education in the Health Sciences
amee

Taylor & Francis
Taylor & Francis Group

Medical Teacher

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/imte20

# What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy

Seetha U. Monrad , Nikki L. Bibler Zaidi , Karri L. Grob , Joshua B. Kurtz , Andrew W. Tai , Michael Hortsch , Larry D. Gruppen & Sally A. Santen

**To cite this article:** Seetha U. Monrad , Nikki L. Bibler Zaidi , Karri L. Grob , Joshua B. Kurtz , Andrew W. Tai , Michael Hortsch , Larry D. Gruppen & Sally A. Santen (2021): What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy, Medical Teacher, DOI: 10.1080/0142159X.2021.1879376

**To link to this article:** https://doi.org/10.1080/0142159X.2021.1879376

View supplementary material

Published online: 16 Feb 2021.

Submit your article to this journal

Article views: 113

View related articles

View Crossmark data

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2021.1879376

# What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy

Seetha U. Monrad[a], Nikki L. Bibler Zaidi[b], Karri L. Grob[c], Joshua B. Kurtz[d*], Andrew W. Tai[e], Michael Hortsch[f], Larry D. Gruppen[g] and Sally A. Santen[h]

[a]Division of Rheumatology, Department of Internal Medicine, University of Michigan Medical School (UMMS), Ann Arbor, MA, USA; [b]RISE innovation unit, University of Michigan Medical School, Ann Arbor, MA, USA; [c]Office of Medical School Education, University of Michigan Medical School, Ann Arbor, MA, USA; [d]University of Michigan Medical School, Ann Arbor, MA, USA; [e]Division of Gastroenterology, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MA, USA; [f]Department of Cell and Developmental Biology, University of Michigan Medical School, Ann Arbor, MA, USA; [g]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MA, USA; [h]Department of Emergency Medicine, Virginia Commonwealth University School of Medicine, Richmond, VA, USA

**ABSTRACT**

**Background:** Using revised Bloom's taxonomy, some medical educators assume they can write multiple choice questions (MCQs) that specifically assess higher (analyze, apply) versus lower-order (recall) learning. The purpose of this study was to determine whether three key stakeholder groups (students, faculty, and education assessment experts) assign MCQs the same higher- or lower-order level.

**Methods:** In Phase 1, stakeholders' groups assigned 90 MCQs to Bloom's levels. In Phase 2, faculty wrote 25 MCQs specifically intended as higher- or lower-order. Then, 10 students assigned these questions to Bloom's levels.

**Results:** In Phase 1, there was low interrater reliability within the student group (Krippendorf's alpha = 0.37), the faculty group (alpha = 0.37), and among three groups (alpha = 0.34) when assigning questions as higher- or lower-order. The assessment team alone had high interrater reliability (alpha = 0.90). In Phase 2, 63% of students agreed with the faculty as to whether the MCQs were higher- or lower-order. There was low agreement between paired faculty and student ratings (Cohen's Kappa range .098–.448, mean .256).

**Discussion:** For many questions, faculty and students did not agree whether the questions were lower- or higher-order. While faculty may try to target specific levels of knowledge or clinical reasoning, students may approach the questions differently than intended.

**KEYWORDS**
Multiple choice questions; assessment; basic science; medical student

## Introduction

Medical education strives to promote critical thinking and clinical reasoning in trainees to foster the skills needed to provide care to increasingly complex patients (Eva 2005). This process requires the ability to synthesize large amounts of information, critically assess different diagnostic and therapeutic strategies, and evaluate possible outcomes. Proficient clinical reasoning is thought to be a combination of cognitive processes—pattern recognition and analytic reasoning (Eva et al. 2007; Pelaccia et al. 2011). In pattern recognition, the person considers a compilation of findings and determines the most plausible explanation for the pattern (Eva 2005; Norman et al. 2007; Brush et al. 2017). This occurs with some degree of automaticity, occurring outside conscious awareness of analytic processes involved. Clinicians are often faced with clinical situations where they need not "reason" at all, instead can identify patterns to make a diagnosis. The more analytic approach is a deliberate and iterative process by which diagnostic hypotheses are generated and tested to arrive at a diagnosis. Experts efficiently use a combination of pattern recognition and analytical reasoning. When the diagnostician does not

### Practice points

- Faculty write multiple choice questions based on the level of revised Bloom's taxonomy, some that specifically regard higher (analyze, apply) versus lower-order (recall).
- Faculty, educators, and students did not agree whether specific questions were lower- or higher-order.
- When faculty intentionally wrote higher-/lower-order questions, only 63% of the time did students agree with that designation.
- This study calls into question whether the modified Bloom's taxonomy is sufficiently compatible with writing MCQs to justify its routine application.

CONTACT Seetha U. Monrad seetha@med.umich.edu Assistant Dean of Evaluation, Assessment, and Quality Improvement and Scholarship and Associate Professor of Internal Medicine and Learning Health Sciences, University of Michigan Medical School, 6125 Taubman Health Sciences Library, 1135 Catherine Street, Ann Arbor 48109-5726, MI, USA
*Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA.

Supplemental data for this article can be accessed here.

recognize a story/pattern, one approach involves first generating a list of diagnostic hypotheses which are then verified or rejected through analytic reasoning—leading to a diagnostic conclusion (Norman et al. 2007; Brush et al. 2017). The strategy employed by trainees is similar to those employed by experienced doctors, although less efficiently and with less breadth based on limited foundational knowledge and experience.

Ideally, educational experiences and associated assessments are designed to support the development of clinical reasoning, based on our understanding of human cognition and learning. Medical educators often utilize Bloom's revised taxonomy for teaching, learning, and assessment, as it provides an easily understandable and practical framework with which to develop curricula (Anderson and Krathwohl 2001; Krathwohl 2002). The revised taxonomy organizes the cognitive processes with which learners engage with knowledge into six categories of increasing complexity; remembering, understanding, applying, analyzing, evaluating, and creating (Bloom et al. 1956; Krathwohl 2002). The taxonomy is also hierarchical, with engagement in higher order processes (such as applying and analyzing information) relying on lower order processes (such as recall and comprehension). When considering the application of the modified Bloom's taxonomy to clinical reasoning education, lower levels of Bloom's could theoretically target pattern recognition cognitive processes, while higher order levels could target analytical reasoning.

Assessment must reinforce the cognitive processes underpinning clinical reasoning. Assessments can help support trainees' understanding of core concepts of medical knowledge and patient care as well as foster their ability to integrate and synthesize information to gain deeper understanding and application (Buckwalter et al. 1981; Epstein 2007; Cilliers et al. 2012). Therefore, it is important that sound assessments support the learning necessary to develop clinical reasoning skills.

Multiple-choice questions (MCQs) are commonly used to assess student learning in the pre-clinical phase. While they can clearly assess factual knowledge, well-written MCQs can support learner engagement in higher levels of cognitive reasoning such as application or synthesis of knowledge (ten Cate et al. 2018). The use of MCQs to engage different levels of cognitive levels has been shown in several studies (Jensen et al. 2014; Ali and Ruit 2015; Billings et al. 2016; Kibble 2017; Choudhury and Freemont 2017). Testing of lower (factual recall) rather than higher (application of knowledge) cognitive function is noted to be a significant impediment to the quality of MCQs (Tarrant and Ware 2008; Tarrant et al. 2009). High-quality MCQ examinations include items that test the learning objectives and target the cognitive levels appropriate for a given subject and learner (Thompson and O'Loughlin 2015).

It is often suggested that applying the cognitive domains of Bloom's when creating MCQs will result in items that measure higher-order thinking, rather than simply an examinee's ability to recall factual information (Crowe et al. 2008; Kim et al. 2012; Jensen et al. 2014; Thompson and O'Loughlin 2015; Karpen and Welch 2016; Thompson et al. 2016; Cecilio-Fernandes et al. 2018). Thus, a practice adopted by some medical educators is to create assessments that target specific levels of Bloom's learning hierarchy in order to promote higher-order learning; however, it is unclear as to the effectiveness of this approach (Kim et al. 2012; Thompson and O'Loughlin 2015). The National Board of Medical Examiners notes that - "Rule 2: Each item should assess application of knowledge, not recall of an isolated fact (Billings et al. 2016, p. 29)." "In addition to considering topics that are important to include on a test, the item writer should think about how to structure those questions to test more than just recall of isolated facts (Billings et al., 2016 page 29)." Some studies of cognitive testing have used test questions that were designated by faculty members and/or researchers as Bloom's higher or lower orders (Buckwalter et al. 1981; Palmer and Devitt 2007; Burns 2010; Jensen et al. 2014; Freiwald et al. 2014; Thompson and O'Loughlin 2015; Cecilio-Fernandes et al. 2018). An underlying premise of these studies is that assessment questions can be categorized to be higher- or lower-order. Nevertheless, whether a trainee taking the examination will use the same cognitive level as the faculty who wrote the question is not clear. Aligning student and faculty question writer determination of cognitive levels can be difficult as the two parties present with disparate levels of knowledge (Thompson and O'Loughlin 2015; Zaidi et al. 2018).

The purpose of this study was to explore whether MCQs could reliably be categorized as higher-order (application, analysis, synthesis, and evaluation levels) and lower-order (knowledge and comprehension levels) by three stakeholder groups (students, faculty, assessment team).

## Methods

The University of Michigan Medical School developed a process to help guide faculty in creating MCQ examinations for preclerkship courses (Bibler Zaidi et al. 2016, 2017, 2018; Zaidi et al. 2017; Santen et al. 2019). The Evaluation and Assessment (E&A) team, staff with advanced degrees in education and applied assessment, created a framework for categorizing MCQs into "lower-order" and "higher-order," according to a dichotomized Bloom's taxonomy (Bloom et al. 1956; Krathwohl 2002; Bibler Zaidi et al. 2016; Zaidi et al. 2017). The purpose was to guide faculty to write questions aimed at higher-order thinking rather than lower-order recall/identification-type questions. As part of the development and implementation of this process, the E&A team participated in multiple norming sessions to refine categorizations and increase interrater agreement among these non-content expert staff reviewers. The guidelines were (1) focus only on *how* the MCQ was written—not on content; (2) assume that the MCQ was not explicitly used for teaching purposes, (3) assume all item content (e.g. information provided in the question stem) was germane to the MCQ, (4) focus on the stem only and do not factor the response options into consideration (Zaidi et al. 2018). The team neither attempted to identify correct answer options nor make connections between the questions posed and actual clinical or basic science content. Using classifications made by this team, we found that the modified Bloom's successfully identified MCQs that were more difficult for students, adding some preliminary validity to our process.

While this process was grounded in theory and consistently achieved high interrater agreement in categorizations by the E&A team, the Bloom's categorizations were made by non-content experts, independent of input from question writers or consideration of the content of the question, and thus were not calibrated to medical students and faculty categorizations. To determine the generalizability of the dichotomized Bloom's, we analyzed the application of the framework by varying levels of content expertise using three stakeholder groups: E&A team, faculty, and students. Our goal was to examine whether all stakeholder groups assign MCQs to the same higher or lower division in Bloom's. There were two phases to this process.

### *Phase 1*

A convenience sample of six teaching faculty members from the foundational science (pre-clerkship) curriculum and five medical students, who served as curriculum representatives for their class, were invited to participate in Bloom's coding sessions led by the E&A team. The intent of each session was to review the MCQs from administered examinations and discuss how members of each group applied Bloom's to the categorization of MCQs.

We conducted three independent coding sessions lasting approximately three hours—two sessions with students and one session with faculty members. For all three sessions, participants were provided an overview of Bloom's and the dichotomized framework. A total of 90 exam items from three different pre-clerkship exams were reviewed by all student and faculty participants and categorized as higher- or lower-order. The questions being reviewed were not ones the reviewing faculty had written. Participants independently reviewed sets of 10 MCQs at a time and then reported their categorizations. When disagreement occurred, the item would be discussed to understand differing perspectives. Participants and investigators recorded field notes of observations during each session. Notes focused on the review process, specifically addressing what elements of an MCQ made an item higher- or lower-order for each group.

### *Phase 2*

In Phase 1, it was not known whether the questions were intended to be higher- or lower-order by the question authors. Therefore, in Phase 2, faculty were asked to intentionally write higher- or lower-order questions. Three faculty members from the pre-clerkship gastroenterology course (a physiologist, histologist, and clinician) intentionally wrote a total of 25 higher-order or lower-order questions. We used this intention as the gold standard rating, and then recruited 10 students across all quartiles of student performance to review the questions and determine if they were higher- or lower-order. The aim was to examine whether students classified questions as faculty intended.

### *Analysis*

Inter-rater reliability for each group was calculated for each phase using SPSS (IBM SPSS, V 22.0). Phase 1 used multiple raters and in some cases, varying numbers of raters rated each item. Therefore, a Krippendorf's alpha (where 0 indicates absence or reliability and 1 is perfect agreement) was calculated. This test was used because it ignores missing data and can handle various categories and numbers of raters. Phase 2 involved paired student-faculty ratings, so a Cohen's kappa was calculated for each student-faculty pairing separately. This study was determined to be exempt from ongoing review (IRB HUM00130655).

## Results

### *Phase 1*

Figure 1 provides an example MCQ that demonstrates these three groups' different approaches to Bloom's categorization. (Correct answers were not provided to categorizers.)

#### *Consistency among the E&A team*

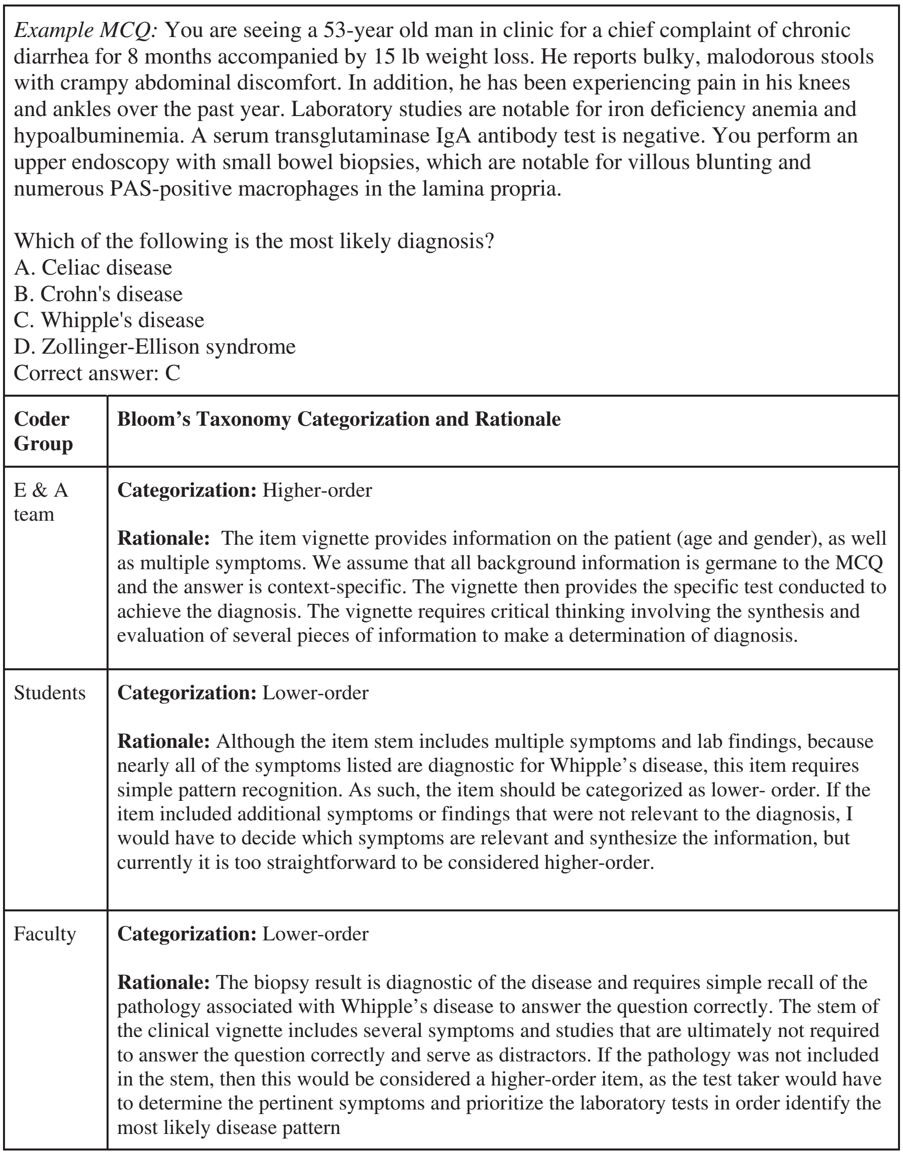
The E&A team did not participate in a formal coding session; they had calibrated over a three-year period based on content-independent guidelines for characteristics of higher- versus lower-order questions. Due to this shared mental model, the two core members who reviewed MCQs for each exam demonstrated very high interrater agreement (Krippendorf's alpha = .90; CI = .80–.97).

#### *Consistency among student group*

Consistency among the students' categorizations was low (Krippendorf's alpha = 0.37; CI = 0.18–0.54). Overall, we found that for any given MCQ, some students approached the question as lower-order (either a knowledge and comprehension task) while other students applied higher-order approaches (application, analysis, synthesis, and/or evaluation). The perspective of the students is published elsewhere and summarized below (Zaidi et al. 2018). In general, student categorizations largely depended on the framing faculty used when presenting information to students; whether students focused on details or broader concepts while studying; and their confidence with the information presented in the question. In addition, the format of the questions affected students' categorization. For example, MCQs containing clinical vignettes were more likely to be seen as higher order, unless the vignettes did not provide meaningful information or had a pathognomonic identifier (e.g. Kayser-Fleischer rings indicating that the vignette was about Wilson's disease, regardless of additional details provided).

#### *Consistency among faculty group*

The faculty were of different clinical specialties (Gastroenterology, Emergency Medicine, Endocrine, Rheumatology, Cardiology). Similar to the student group, consistency among faculty categorizations was low (Krippendorf's alpha = .37; CI = 0.24–0.50). Faculty content expertise influenced categorizations as faculty drew from a larger body of heuristic techniques and knowledge than students in the process of answering an MCQ. This allowed one or more cognitive steps to be skipped, rendering a question to be perceived as lower-order by the faculty. For example, many MCQs were written in the style of a clinical vignette that related to a classic presentation of a disease. If the item stem then asked for the "most likely diagnosis," faculty might categorize this as a lower-order item, even when the MCQ did not explicitly ask for recall of a fact. This was attributed to the faculty member's ability to quickly extract the key diagnostic criteria for a disease from the vignette by nonanalytic pattern recognition without



*Example MCQ:* You are seeing a 53-year old man in clinic for a chief complaint of chronic diarrhea for 8 months accompanied by 15 lb weight loss. He reports bulky, malodorous stools with crampy abdominal discomfort. In addition, he has been experiencing pain in his knees and ankles over the past year. Laboratory studies are notable for iron deficiency anemia and hypoalbuminemia. A serum transglutaminase IgA antibody test is negative. You perform an upper endoscopy with small bowel biopsies, which are notable for villous blunting and numerous PAS-positive macrophages in the lamina propria.

Which of the following is the most likely diagnosis?
A. Celiac disease
B. Crohn's disease
C. Whipple's disease
D. Zollinger-Ellison syndrome
Correct answer: C

| **Coder Group** | **Bloom's Taxonomy Categorization and Rationale** |
|---|---|
| E & A team | **Categorization:** Higher-order<br>**Rationale:** The item vignette provides information on the patient (age and gender), as well as multiple symptoms. We assume that all background information is germane to the MCQ and the answer is context-specific. The vignette then provides the specific test conducted to achieve the diagnosis. The vignette requires critical thinking involving the synthesis and evaluation of several pieces of information to make a determination of diagnosis. |
| Students | **Categorization:** Lower-order<br>**Rationale:** Although the item stem includes multiple symptoms and lab findings, because nearly all of the symptoms listed are diagnostic for Whipple's disease, this item requires simple pattern recognition. As such, the item should be categorized as lower- order. If the item included additional symptoms or findings that were not relevant to the diagnosis, I would have to decide which symptoms are relevant and synthesize the information, but currently it is too straightforward to be considered higher-order. |
| Faculty | **Categorization:** Lower-order<br>**Rationale:** The biopsy result is diagnostic of the disease and requires simple recall of the pathology associated with Whipple's disease to answer the question correctly. The stem of the clinical vignette includes several symptoms and studies that are ultimately not required to answer the question correctly and serve as distractors. If the pathology was not included in the stem, then this would be considered a higher-order item, as the test taker would have to determine the pertinent symptoms and prioritize the laboratory tests in order identify the most likely disease pattern |

Figure 1. Example of Bloom's taxonomy categorizations, by Coder group.

Table 1. Percentage of students' agreement with faculty for each multiple-choice question.

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Faculty rating | Low | Low | High | Low | High | High | High | Low | Low | High | High | Low |
| % agreement | 80 | 80 | 90 | 100 | 60 | 100 | 40 | 0 | 40 | 80 | 70 | 70 |
| | **Q13** | **Q14** | **Q15** | **Q16** | **Q17** | **Q18** | **Q19** | **Q20** | **Q21** | **Q22** | **Q23** | **Q24** |
| Faculty rating | High | High | High | High | Low | Low | High | Low | High | Low | Low | High |
| % agreement | 70 | 10 | 100 | 20 | 50 | 90 | 90 | 100 | 70 | 40 | 60 | 70 |

having to use analytic or higher order reasoning. Additionally, faculty considered the content of MCQ responses, rather than only the item stem, to determine if items were higher or lower-order—even if the mechanics of the item stem seemed like simple recall. The lack of agreement may have occurred because some faculty were not content experts in the material and thus either remembered a key feature, making the question lower order, or were unfamiliar with the content and therefore thought the question was higher order.

The use of images in items was also a factor in faculty categorizations. For example, an item using a histopathology image was generally considered a lower-order item if it named the organ in the item stem and asked an examinee to identify a specific structure or cell type. On the other hand, the same item could be considered a higher-order if it omitted the name of the organ in the item stem, as this would require the examinee to first identify the tissue based on the image to answer the item correctly.

### Cross-group agreement

We found overall agreement among the three groups to be very low (Krippendorf's alpha = .34; CI = .16–.52).

## Phase 2

For the 25 questions that were intentionally written by faculty as higher- or lower-order, on average 63% of students agreed with the faculty categorizations (Table 1). There was

66% agreement for lower-order items (12) and 62% agreement for higher-order items (with a range of 0–100% agreement for each item). Cohen's kappa measuring the agreement between paired student-faculty ratings was low, ranging from .098 to .448 with a mean .256. The Supplementary Appendix document provides examples of MCQs and rationales for higher or lower order classifications.

## Discussion

In our study, we found there was a lack of agreement whether questions were higher-order or lower-order amongst and between students, faculty, and educators. Although differences among coding groups were anticipated, the amount of variability within and between groups was surprising. The two phases of our study allowed for nuanced analysis of discrepancies between faculty and student categorization.

Phase 1 questions were drawn from existing examinations and analyzed retrospectively. E&A team members focused only on the mechanics of the question stem, not the content of the question or how the material was taught. Students' individual experiences and perspectives influenced their interpretations (Zaidi et al. 2018). A student who could not simply remember the answer to a MCQ would generally categorize the question as higher-order because the item required cognitive steps beyond recall. Conversely, when students recognized a pattern or buzzword, the question became lower-order, even though the case vignette or the pattern may have been complex.

Faculty often relied on their broad knowledge base and experience with multiple clinical presentations; they used pattern recognition techniques to skip levels of cognitive processing, thereby categorizing items that by their mechanics appeared to be higher-order to a lower-order designation. In approaches to clinical reasoning, it has been observed that non-analytic approaches to clinical reasoning are not conducive to retrospection/introspection (Norman et al. 2007). We postulate that items might have been categorized differently by faculty in Phase 1 had they been explicitly asked to do so from the perspective of the medical student. Yet, even in Phase 2 where the faculty intentionally wrote questions to be higher- or lower-order, students did not consistently identify the questions as such. This demonstrates that it can be challenging to write questions using a different frame of reference than one's own.

We hypothesize that the discrepancy between student and faculty categorizations arises for two reasons: (1) the use of different clinical reasoning strategies based on different underlying knowledge by the two groups, as described above, and (2) the impact of pedagogy (instruction). Kern and Thomas's framework for curriculum development emphasizes the interrelatedness of all elements in the instruction - learning - assessment cycle (Kern 2009). Assessments must be matched not only to learning objectives but also to instructional strategies. For example, how to assess a learning objective such as "Differentiate between a physiologic split S2 and a pathophysiologic S3" will depend on whether the students learn this distinction through review of text descriptions or via auscultation of different heart sounds. How faculty design curriculum and associated assessments may differ from how students experience them, depending on students' approaches to learning and faculty/learner choice of educational strategies. For example, some items might be drawn directly from a point made explicitly in lecture, with the lecturer occasionally stating that students "should know this for the exam"; these items would therefore be categorized as lower-order, even if the question was designed to be higher-order.

Phase 2 found that even when questions were intentionally developed as higher- or lower-order, independent of pedagogy, there was still disagreement between the students' and faculty's perspectives. There was not a pattern of question type or content between the questions with high agreement and those without. It has been reported previously that both students and faculty utilize qualitatively identical processes in clinical reasoning involving both nonanalytic and analytic reasoning (Neufeld et al. 1981), but more experienced clinicians are more likely to arrive at the correct diagnostic conclusion given more previous experience to draw upon. It is possible that part of the discrepancy in categorization between students and faculty is secondary to faculty overestimating the breadth of the student knowledge base due to the "curse of knowledge" phenomenon (Camerer et al. 1989). Faculty may therefore assume in writing certain lower-order Bloom's questions that students will be able to utilize nonanalytic diagnostic reasoning, whereas students' developing knowledge bases force them to utilize a more analytic diagnostic approach, thereby identifying questions requiring such an approach as being appropriately categorized as higher-order Bloom's tasks (i.e. analysis and evaluation). Conversely, faculty may underestimate the breadth of student knowledge (for example, due to incomplete knowledge of what students have already learned prior to encountering their material) and assign questions as higher order based on such assumptions.

This study brings into question whether the modified Bloom's taxonomy is sufficiently compatible with writing MCQs to justify its routine application in medical trainee assessment. In exploring this issue, we returned to the literature. A number of important studies in the educational psychology literature have addressed this issue in non-medical education fields. For example, Chi et al. found that physics experts engage in qualitative analysis to recognize patterns prior to problem solving (pattern recognition), but then abstract principles to solve problem representations (pattern recognition → analytic reasoning); whereas physics novices primarily use literal features to base their representations (Chi et al. 1981). A similar approach may occur in medical students. Students often study by taking numerous practice questions. In this process, they may be learning patterns that they then apply to the examination questions.

Our results lead us to conclude that Bloom's taxonomy may not align sufficiently with the cognitive processes underlying medical expertise, or the impact of instruction/pedagogy on learning, to allow for *a priori* application. Individual approaches to questions factored heavily into decision-making around final Bloom's categorizations for both faculty and students, and they were not aligned, either with each other or with the non-content expert E&A

team. Based on the results of this study, the E&A team has stopped assigning Bloom's categorizations to institutionally generated MCQs. However, we believe that a better understanding of the cognitive processes underlying approaches to question answering may enable us to push faculty and students to higher-order thinking through MCQs. Further, faculty development is needed to write questions that reinforce higher level clinical reasoning skills.

There are some limitations to this study. For the purpose of this study, we dichotomized Bloom's taxonomy which may overly simplify the approach to reasoning and answering questions. While we carefully instructed the students and faculty on Bloom's categorization, we deliberately did not calibrate the different groups to each other in order to elicit differences in perspective. It is therefore possible that some of the findings are due to construct-irrelevant variance such as lack of shared understanding of the process or difficulty applying the coding schema. Despite framing of the goals of the study, it is possible that faculty and students may be confusing "higher-order" with "more difficult" and "lower-order" with "easier." The numbers of participants are small and may not reflect the student or faculty population. Finally, the relationship between learning and assessing clinical reasoning and the role of MCQs in that process is not well understood.

Next steps in this program of research include further empirical study of how different stakeholder groups (learners, faculty) conflate or distinguish between "easy/hard" and "lower order/higher order" questions. For example, question writers' explanations of why, for a particular question, the correct answer is correct and the distractors are not could be analyzed to understand faculty designations of higher- versus lower-order questions. Additionally, it is necessary to study whether the underlying skills and knowledge of the students makes a difference in their reflection on learning, particularly their interpretation of the underlying cognitive levels being assessed. Are students who have less confidence and knowledge more likely to approach questions as higher-order because they need to reason through the material to come to an answer? Further, are there student approaches to studying that encourage a higher-order approach (they prefer to think through things) or lower-order (they like to memorize as much as they can)? Finally, correlation with psychometric and other analytics from actual question administrations (difficulty index, discrimination index, time spent on a question, etc.) could provide important triangulation between perceptions of difficulty/higher versus lower order with actual performance and behavioral outcomes.

Through this work, we hope to better understand the relationship between frameworks for learning and their practical applicability to assessment of said learning. We also hope to add to the literature on medical students' approaches to learning and testing, in order to continuously refine and improve current assumptions and practices.

## Acknowledgements

The authors appreciate and acknowledge the contributions of the faculty and students who reviewed the questions, especially Dr. Asra Ahmed; Dr John Williams for providing questions; Jun Yang for statistical support; and Jonie Josephs and Heather Brickley for administrative support.

## Disclosure statement

The University of Michigan Medical School and Virginia Commonwealth University School of Medicine receive funding from the American Medical Association for the Accelerating Change in Medical Education Grant.

## Notes on contributors

*Seetha U. Monrad*, MD, is assistant dean of evaluation, assessment, and quality improvement and clinical associate professor of Internal Medicine and Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan.

*Nikki L. Bibler Zaidi*, PhD, is director of evaluation and assessment, Research. Innovation. Scholarship, Education (R.I.S.E.), University of Michigan Medical School, Ann Arbor, Michigan.

*Karri L. Grob*, EdS, is director of student services Office of Medical Student Education, University of Michigan Medical School, Ann Arbor, Michigan.

*Joshua B. Kurtz*, MD, is a first year pediatric resident at the Children's Hospital of Philadelphia, Pennsylvania.

*Andrew W. Tai*, MD, PhD, is associate professor of internal Medicine and Microbiology & Immunology, University of Michigan Medical School, Ann Arbor, Michigan.

*Michael Hortsch*, PhD, is professor of Cell and Developmental Biology and of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan.

*Larry D. Gruppen*, PhD, is professor in the Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan.

*Sally A. Santen*, MD, PhD, is the senior associate dean, assessment, evaluation and scholarship and professor, emergency medicine, Virginia Commonwealth University School of Medicine, Richmond, VA, and was the assistant dean, and professor, Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, Michigan.

## ORCID

Seetha U. Monrad http://orcid.org/0000-0002-3374-2989
Nikki L. Bibler Zaidi http://orcid.org/0000-0001-6364-9358
Karri L. Grob http://orcid.org/0000-0002-9883-7743
Joshua B. Kurtz http://orcid.org/0000-0001-7528-1722
Andrew W. Tai http://orcid.org/0000-0002-6877-450X
Michael Hortsch http://orcid.org/0000-0002-3750-737X
Larry D. Gruppen http://orcid.org/0000-0002-2107-0126
Sally A. Santen http://orcid.org/0000-0002-8327-8002

## References

Ali SH, Ruit KG. 2015. The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. Perspect Med Educ. 4(5):244–251.

Anderson LW, Krathwohl DR, editors. 2001. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. Complete ed. New York: Longman.

Bibler Zaidi NL, Grob KL, Monrad SU, Holman ES, Gruppen LD, Santen SA. 2018. Item quality improvement: what determines a good question? Guidelines for interpreting item analysis reports. Med Sci Educ. 28(1):13–17.

Bibler Zaidi NL, Grob KL, Yang J, Santen SA, Monrad SU, Miller JM, Purkiss JA. 2016. Theory, process, and validation evidence for a staff-driven medical education exam quality improvement process. Med Sci Educ. 26(3):331–336.

Bibler Zaidi NL, Monrad SU, Grob KL, Gruppen LD, Cherry-Bukowiec JR, Santen SA. 2017. Building an exam through rigorous exam quality improvement. Med Sci Educ. 27(4):793–798.

Billings MS, DeRuchie K, Haist SA, Hussie K, Merrell J, Paniagua MA, Swygert KA, Tyson J. 2016. Constructing written test questions for the basic and clinical sciences. 4th ed. Philadelphia (PA): National Board of Medical Examiners.

Bloom BS, Englehart MD, Furst EJ, Hill WH, Krathwohl DR. 1956. Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. London: Longmans, Green and Co LTD.

Brush JE, Sherbino J, Norman GR. 2017. How expert clinicians intuitively recognize a medical diagnosis. Am J Med. 130(6):629–634.

Buckwalter J, Schumacher R, Albright J, Cooper R. 1981. Use of an educational taxonomy for evaluation of cognitive performance. J Med Educ. 56(2):115–121.

Burns ER. 2010. "Anatomizing" reversed: use of examination questions that foster use of higher order learning skills by students. Anat Sci Educ. 3(6):330–334.

Camerer C, Loewenstein G, Weber M. 1989. The curse of knowledge in economic settings: an experimental analysis. J Polit Econ. 97(5): 1232–1254.

Cecilio-Fernandes D, Kerdijk W, Bremers AJ, Aalders W, Tio RA. 2018. Comparison of the level of cognitive processing between case-based items and non-case-based items on the Interuniversity Progress Test of Medicine in the Netherlands. J Educ Eval Health Prof. 15:28.

Chi MTH, Feltovich PJ, Glaser R. 1981. Categorization and representation of physics problems by experts and novices. Cogn Sci. 5(2): 121–152.

Choudhury B, Freemont A. 2017. Assessment of anatomical knowledge: approaches taken by higher education institutions. Clin Anat. 30(3): 290–299.

Cilliers FJ, Schuwirth LW, van der Vleuten CP. 2012. A model of the pre-assessment learning effects of assessment is operational in an undergraduate clinical context. BMC Med Educ. 12(1):9.

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

Epstein RM. 2007. Assessment in medical education. N Engl J Med. 356(4):387–396.

Eva KW. 2005. What every teacher needs to know about clinical reasoning. Med Educ. 39(1):98–106.

Eva KW, Hatala RM, LeBlanc VR, Brooks LR. 2007. Teaching from the clinical reasoning literature: combined reasoning strategies help novice diagnosticians overcome misleading information. Med Educ. 41(12):1152–1158.

Freiwald T, Salimi M, Khaljani E, Harendza S. 2014. Pattern recognition as a concept for multiple-choice questions in a national licensing exam. BMC Med Educ. 14(1):232.

Jensen JL, McDaniel MA, Woodard SM, Kummer TA. 2014. Teaching to the test … or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. Educ Psychol Rev. 26(2):307–329.

Karpen SC, Welch AC. 2016. Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's taxonomy classifications. Curr Pharm Teach Learn. 8(6):885–888.

Kern DE, editor. 2009. Curriculum development for medical education: a six-step approach. 2nd ed. Baltimore (MD): The Johns Hopkins University School of Medicine.

Kibble JD. 2017. Best practices in summative assessment. Adv Physiol Educ. 41(1):110–119.

Kim M-K, Patel RA, Uchizono JA, Beck L. 2012. Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. Am J Pharm Educ. 76(6):1–8.

Krathwohl DR. 2002. A revision of Bloom's taxonomy: an overview. Theory Pract. 41(4):212–218.

Neufeld V, Norman G, Feightner J, Barrows H. 1981. Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. Med Educ. 15(5):315–322.

Norman G, Young M, Brooks L. 2007. Non-analytical models of clinical reasoning: the role of experience. Med Educ. 41(12):1140–1145.

Palmer EJ, Devitt PG. 2007. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC Med Educ. 7(1):49.

Pelaccia T, Tardif J, Triby E, Charlin B. 2011. An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. Med Educ Online. 16. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3060310/.

Santen SA, Grob KL, Monrad SU, Stalburg CM, Smith G, Hemphill RR, Bibler Zaidi NL. 2019. Employing a root cause analysis process to improve examination quality. Acad Med. 94(1):71–75.

Tarrant M, 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.

Tarrant M, Ware J, Mohammed AM. 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ. 9(1):40.

ten Cate O, Custers E, Durning S, editors. 2018. Principles and practice of case-based clinical reasoning education: a method for preclinical students. 15th ed. Cham (Switerland): Springer International Publishing.

Thompson AR, Kelso RS, Ward PJ, Wines K, Hanna JB. 2016. Assessment driven learning: the use of higher-order and discipline-integrated questions on gross anatomy practical examinations. Med Sci Educ. 26(4):587–596.

Thompson AR, O'Loughlin VD. 2015. The Blooming Anatomy Tool (BAT): a discipline-specific rubric for utilizing Bloom's taxonomy in the design and evaluation of assessments in the anatomical sciences. Anat Sci Educ. 8(6):493–501.

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

Zaidi NLB, Grob KL, Monrad SM, Kurtz JB, Tai A, Ahmed AZ, Gruppen LD, Santen SA. 2018. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? Acad Med. 93(6): 856–859.