COLLEGE of AMERICAN PATHOLOGISTS

ARCHIVES
of Pathology & Laboratory Medicine

EARLY ONLINE RELEASE

Note: This article was posted on the *Archives* Web site as an Early Online Release. Early Online Release articles have been peer reviewed, copyedited, and reviewed by the authors. Additional changes or corrections may appear in these articles when they appear in a future print issue of the *Archives*. Early Online Release articles are citable by using the Digital Object Identifier (DOI), a unique number given to every article. The DOI will typically appear at the end of the abstract.

The DOI for this manuscript is doi: 10.5858/arpa.2020-0624-OA

The final published version of this manuscript will replace the Early Online Release version at the above DOI once it is available.

Assessment Question Characteristics Predict Medical Student Performance in General Pathology

Tahyna Hernandez, MD; Margret S. Magid, MD; Alexandros D. Polydorides, MD, PhD

• *Context.*—**Evaluation of medical curricula includes appraisal of student assessments in order to encourage deeper learning approaches. General pathology is our institution's 4-week, first-year course covering universal disease concepts (inflammation, neoplasia, etc).**

Objective.—**To compare types of assessment questions and determine which characteristics may predict student scores, degree of difficulty, and item discrimination.**

Design.—**Item-level analysis was employed to categorize questions along the following variables: type (multiple choice question or matching answer), presence of clinical vignette (if so, whether simple or complex), presence of specimen image, information depth (simple recall or interpretation), knowledge density (first or second order), Bloom taxonomy level (1–3), and, for the final, subject familiarity (repeated concept and, if so, whether verbatim).**

Results.—**Assessments comprised 3 quizzes and 1 final exam (total 125 questions), scored during a 3-year period (total 417 students) for a total 52 125 graded attempts. Overall, 44 890 attempts (86.1%) were correct. In multivariate analysis, question type emerged as the most significant predictor of student performance, degree of difficulty, and item discrimination, with multiple choice questions being significantly associated with lower mean scores ($P = .004$) and higher degree of difficulty ($P = .02$), but also, paradoxically, poorer discrimination ($P = .002$). The presence of a specimen image was significantly associated with better discrimination ($P = .04$), and questions requiring data interpretation (versus simple recall) were significantly associated with lower mean scores ($P = .003$) and a higher degree of difficulty ($P = .046$).**

Conclusions.—**Assessments in medical education should comprise combinations of questions with various characteristics in order to encourage better student performance, but also obtain optimal degrees of difficulty and levels of item discrimination.**

(*Arch Pathol Lab Med.* doi: 10.5858/arpa.2020-0624-OA)

There are multiple challenges facing the design and implementation of continuously changing and evolving undergraduate medical education curricula. In terms of content, educators need to include ever accumulating medical knowledge, modify emphasis and perspective in order to adapt to shifting societal contexts, and attempt to better integrate the dual pillars of basic science and clinical application (as exemplified by learning and practicing evidence-based medicine).[1–6] In terms of delivery, there is increasing pressure to emulate the competence-based training models adopted in graduate medical education, a desire to embrace more updated and advanced pedagogic techniques (such as self-directed and team-based learning), and a necessity to incorporate current technologic advances in health care (eg, digital and digitized medicine).[7–9] Students use 3 main approaches to learning and studying: deep/thoughtful (seeking to understand, relating new concepts to prior knowledge, and critically examining evidence); superficial/surface (completing the task, memorizing information, and focusing on individual points without recognizing wider context or reflecting on the process); and strategic/efficient (organizing work, managing time, and aiming to efficiently pass any assessment).[10,11]

In this context, student assessments that are perceived to reward the understanding of knowledge, rather than simple data regurgitation (ie, recall), would encourage deeper learning approaches. Furthermore, assessments that focus on and evaluate clinical reasoning, judgment, management of ambiguity, lifelong learning, and teamwork strategies would appraise learners in a more reliable and comprehensive manner.[12,13] Assessments should provide both insight into the actual students' performance as well as the tools to recognize areas for change and improvement, thereby helping students identify and respond to their own learning needs.[14] Therefore, assessments in medical education (quizzes, practical tests, final exams, observations, evaluations, and simulations) have to be accurate, reliable, impactful, and constructive. In particular, formative (as opposed to summative) assessments provide timely feedback and suggestions for improvement and are thus

Accepted for publication October 22, 2020.

From the Department of Pathology, Molecular and Cell Based Medicine, Icahn School of Medicine at Mount Sinai, New York, New York (Hernandez, Polydorides); and the Department of Pathology, New York University Grossman School of Medicine, New York, New York (Magid).

The authors have no relevant financial interest in the products or companies described in this article.

This work was partially presented during the 109th annual meeting of the United States and Canadian Academy of Pathology (USCAP); March 2, 2020; Los Angeles, California.

Corresponding author: Alexandros D. Polydorides, MD, PhD, Department of Pathology, Molecular and Cell Based Medicine, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1194, New York, NY 10029 (email: alexandros.polydorides@mountsinai.org).

uniquely able to contribute to enhanced learning outcomes.[15]

A formal way to evaluate and improve assessments in order to ensure that they measure and encourage higher-order thinking rather than simple recall of facts, is Bloom taxonomy.[16] This premise involves the hierarchical ordering of 6 cognitive domains employed with the intention to acquire, retain, and manage knowledge: recall, comprehension, application, analysis, synthesis, and evaluation (also referred to as: remember, understand, apply, analyze, create, and evaluate). Thus, in order to achieve the more desirable higher-order levels of learning, such as synthesis and evaluation, students must first apply the more basic lower-level domains of recall and comprehension. This is particularly applicable to multiple choice questions (MCQs), the most commonly used assessment tool during written examination in medical education, whether in-house at specific medical schools or nationalized, for example, in the US Medical Licensing Examination (USMLE). Accordingly, the National Board of Medical Examiners item-writing guide recommends that both basic science and clinical MCQs begin with a clinical vignette that provides context and requires examinees to determine relevant information by accessing higher-order cognitive skills.[17] Nevertheless, studies have reported that, measured this way, the quality of MCQs throughout medical education, and especially during in-house written examinations, is relatively poor, highlighting an area of need in this process.[18,19]

Undergraduate medical education in pathology has similarly faced significant challenges as it attempts to adjust in this fluctuating environment.[20] Although there is great variability among medical schools in the number of instruction hours allotted, educational modalities used, types and content of assessments, and methods of outcome measurements, there is emerging consensus that "evidence-based education" (meaning the use of data to inform decisions about teaching) is an appropriate and worthy goal.[21] As such, and mirroring recent advances in graduate medical education, there have been recommendations for the development, implementation, and assessment of core competencies in undergraduate pathology that would function as national standards.[22] These competencies would be particularly applicable to schools that implement instruction in the setting of an integrated (as opposed to discipline-specific) curriculum and would focus on the 3 mainstays underlying the central dogma of the pathologic basis of disease, namely, disease mechanisms, organ system pathology, and diagnostic medicine.[12]

Most schools use a combination of instructional methods, including lectures, image-centered portfolios, case-based teaching, and team-based or self-directed learning, to accomplish these goals in pathology education.[23–28] However, there is a lack of robust guidelines and experimental approaches in the evaluation and critical appraisal of these educational interventions.[29,30] To that end, we sought to systematically characterize, evaluate, and compare the various types of assessment questions employed within an introductory course in general pathology during the first year of medical school instruction at our institution. Our goal was to use item-level analysis in order to identify specific assessment characteristics that might influence and predict student performance in this setting and to provide data that could be used to guide future evidence-based decision-making in promoting higher-level learning approaches.

MATERIALS AND METHODS

The study was approved by our school's Institutional Review Board. Deidentified student scores were retrospectively collected during 3 years of instruction (2017, 2018, and 2019). Scores were retrieved from a Web-based course management platform (Blackboard Learn, Blackboard Inc) that administers the assessments and automatically grades and records student attempts. Student attempts were recorded for all assessments (3 short quizzes and 1 longer final examination) in the general pathology course, and average responses (percent correct) were calculated for each question during the 3 events (years). Quizzes occurred once a week, tested new knowledge each time, and were deemed to be formative assessments because they were administered during the course and answers with detailed explanations were provided to the students on an ongoing basis throughout the duration of remaining instruction time. In contrast, the final examination was categorized as a summative assessment because it was administered at the very end of the course, it had a high point value, and it largely evaluated information already tested at least once (during the quizzes) for the purpose of dispensing final course grades.

Each question was characterized across a number of categoric variables, including type of answer option (multiple choice or matching); the presence of a clinical vignette in the question stem (and, if so, whether this vignette was simple or complex); the presence of an accompanying specimen image (gross or microscopic), the depth of information being tested (simple recall or interpretation); the density of knowledge required (first or second order); and the level in a modified Bloom taxonomy pyramid (levels 1 to 3, in increasing order of question complexity). For the final exam, which largely assessed recurring knowledge, questions were also categorized in terms of subject familiarity, that is, whether they tested brand-new information (ie, never before tested) or knowledge repeated from prior assessments (ie, quizzes) and, if so, whether the information tested was repeated verbatim or with slight modifications. Clinical vignettes were designated as simple when specific signs and symptoms were listed, or complex when a synthesis of laboratory and imaging data was required to arrive to a particular conclusion about the patient's condition and/or diagnosis. Depth of information was defined as the level of data collection required by students in order to answer the question, that us, whether simple recollection of facts was sufficient or whether a certain degree of knowledge interpretation was necessary. Density of knowledge required by students measured the initial level of understanding necessary in terms of whether the question tested direct knowledge (ie, first order) or whether it required deductive reasoning to reach a first level of comprehension before answering (ie, second order).

Item difficulty refers to the proportion of students answering a particular question correctly out of all participants and was graded as follows (according to previously published guidelines[31]): very difficult (hard) if $<30\%$, moderate (ie, the desired range) if 30% to 80%, and easy if $>80\%$. Item discrimination, defined as the correlation (2-point correlation coefficient) between a student's score on a specific item (ie, question) and their score on the entire exam (ie, particular quiz or final examination), was calculated and measured based on the point biserial index (PBI).[32] The PBI values were further graded as follows: below 0.2 was considered poor, 0.2 to 0.29 was considered fair, 0.3 to 0.39 was considered good, and 0.4 to 0.7 was considered very good.

Continuous variables (average student scores) were compared across question variables using 1-way analysis of variance. Categoric variables (tiers of item difficulty and item discrimination) were compared using Pearson χ^2 tests. Multivariate logistic regression was performed using the indicated question characteristics as independent variables, with calculated odds ratio (OR) and 95% CI. For the purposes of statistical analysis, item difficulty levels of very difficult (hard) and moderate (ie, $\leq 80\%$ overall) were combined and together compared to items of easy difficulty (ie, $>80\%$). For item discrimination, good and very good PBI values were combined (ie, ≥ 0.3 overall) and compared to poor and fair PBI together (ie, <0.3 overall). Bloom taxonomy level (1–3) was

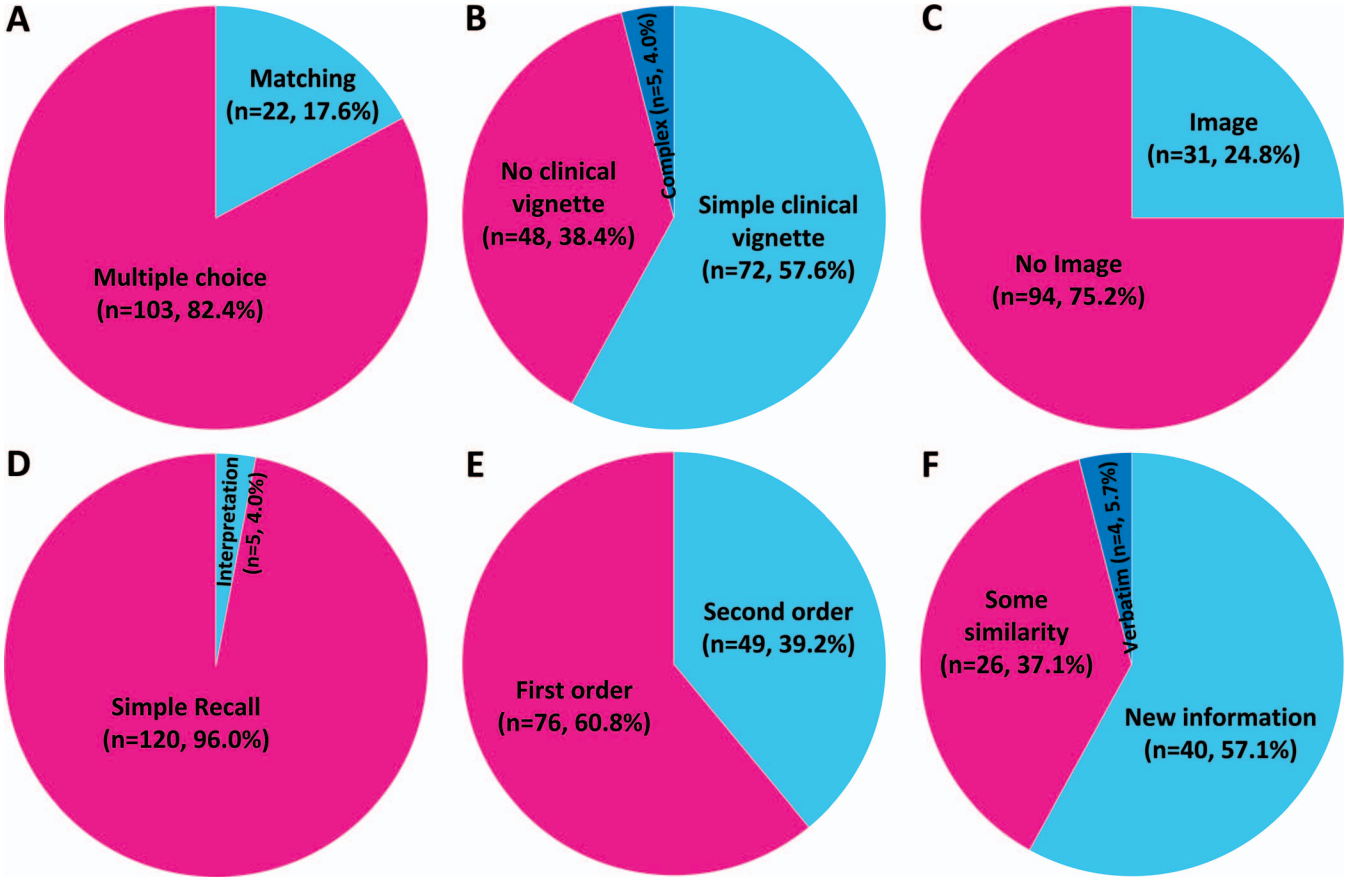

Figure 1. *Pie charts illustrating the number (n) and percentage (%) of the total examination questions falling under each category among the characteristics evaluated in the study, including: (A) answer choice type (multiple choice or matching answer), (B) presence and type of clinical vignette in the question stem, (C) presence of a specimen image, (D) information depth (simple recall or interpretation), (E) knowledge density (first or second order), and (F) subject familiarity in the final examination (repeated information tested and, if so, extent of repetitive information). The total number of questions was 125 for all parameters, except for subject familiarity in the final exam, where the total number of questions was 70.*

categorized as an ordinal variable. Questions in the final examination that tested repeated information from prior quizzes (whether verbatim or with slight modifications) were combined and together compared to questions testing new information. Questions with clinical vignettes (whether simple or complex) were compared to questions without. All statistical analysis was carried out using Statistical Package for the Social Sciences software (SPSS, build 1.0.0.1327, copyright 2019, IBM), with $P < .05$ considered significant throughout.

RESULTS

General pathology is a 4-week, first-year medical school course at our institution that introduces universal disease concepts and pathogenesis mechanisms, before the second-year curriculum delves into more specific organ system pathology. The course is structured across 6 thematic modules (cell injury, inflammation and repair, hemodynamic disorders, immunologically mediated disease, neoplasia, and nutritional pathology), each with 1 or 2 lectures introducing basic concepts followed by small-group laboratory sessions reviewing gross specimens and computer-based microscopic images. In the labs, students in teams of 3 are asked to consider clinical scenarios examining previously introduced concepts as applied to specific patient-based cases. Student assessments during the course consist of 3 formative quizzes (1 at the end of each of the first 3 weeks) covering 1 or 2 modules each, followed by a summative final examination covering material from the entire course. As a result, some of the questions in the final exam may retest knowledge previously assessed in one of the quizzes. The first quiz comprises 15 questions, the second and third quizzes have 20 questions each, and the final exam contains 70 questions, for a total of 125 questions.

These questions were differentiated across a number of characteristics (Figure 1). Of all 125 questions, 103 (82.4%) were MCQs, whereas 22 (17.6%) required matching different answer choices; 77 questions (61.6%) included a clinical vignette in the question stem, 5 (6.5%) of which were complex and 72 (93.5%) were simple vignettes, whereas 48 questions (38.4%) did not contain a clinical vignette; 31 questions (24.8%) contained a specimen image (6 gross, 22 microscopic, and 4 with both), whereas 94 (75.2%) did not. In terms of the depth of information tested, 120 questions (96.0%) were simple recall, whereas 5 (4.0%) required a level of interpretation on the part of the students. As far as knowledge density, 76 questions (60.8%) were first order, whereas 49 (39.2%) were second order. Of the 70 questions in the final examination, 40 questions (57.1%) tested new information, whereas 30 (42.9%) assessed repeated concepts and, of these, 26 (86.7%) contained some similarity to prior questions, whereas 4 (13.3%) repeated concepts verbatim compared with questions in past quizzes.

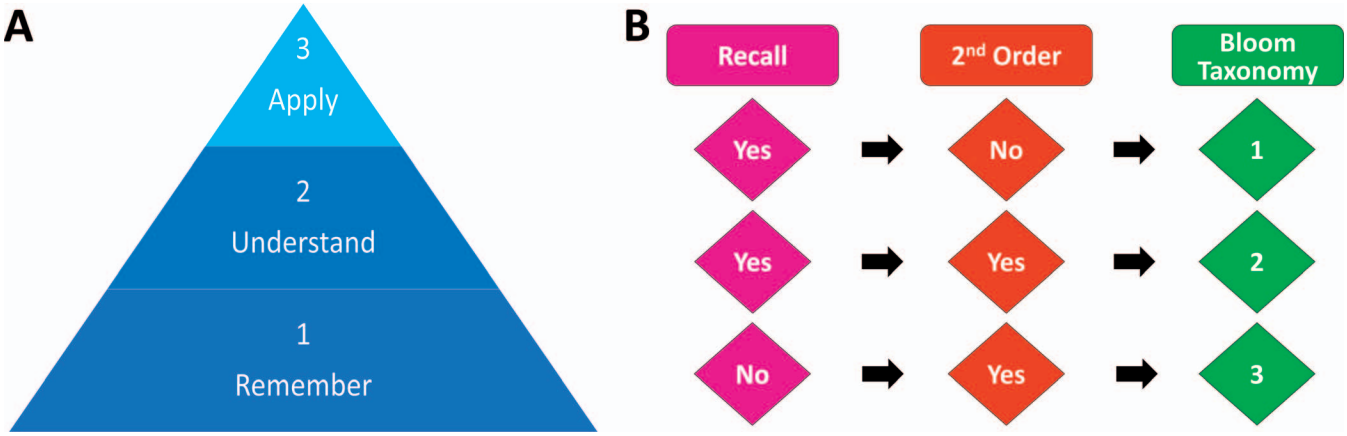

Figure 2. *Schematic depicting the modified Bloom taxonomy used to evaluate assessment questions. (A) Three-tiered Bloom pyramid (from lower/broader meaning simpler to higher/narrower meaning more complex) representing a hierarchic grading of question complexity in terms of the learning processes required by students in order to answer correctly: tier 1, remember (simple recall); tier 2, understand (explanation of concepts); and tier 3, apply (new setting). (B) Algorithm used to assign Bloom levels to examination questions: first-order questions testing simple recall were assigned Bloom level 1; second-order questions that still required simple recall were assigned Bloom level 2; and second-order questions that required interpretation (not just simple recall) were assigned Bloom level 3.*

In order to assess the cognitive level required in each question, we employed a 3-tiered modified Bloom taxonomy pyramid, based on previously described recommendations and examples.[16,33–35] Briefly, questions were assigned levels across a hierarchic ordering of question complexity, based on specified criteria (Figure 2, A). Level 1, also defined as knowledge ("remember"), required only remembering facts and basic concepts in order to answer the question. Questions of this type typically ask students to define, duplicate, list, memorize, repeat, recall, and state knowledge. Level 2, also referred to as comprehension ("understand"), required a level of understanding on the part of the student before arriving to the correct answer. Such questions used verbs like classify, describe, discuss, explain, and identify. Level 3, also defined as application ("apply"), required students to use acquired information and concepts in new settings and situations and included questions with wording such as connect, execute, implement, solve, use, and demonstrate. To minimize subjectivity in the application of these criteria, we used an algorithm in order to assign Bloom levels to various assessment questions, according to our defined parameters (Figure 2, B). Bloom level 1 questions tested simple recall and required first-order density of knowledge. Bloom level 2 questions still assessed simple recall, but they required a certain level of understanding of the concepts involved (second order). Finally, Bloom level 3 questions necessitated a deeper level of information understanding (interpretation, not simple recall) and were second order in terms of knowledge density. Based on this taxonomy, 76 of 125 questions (60.8%) were classified as level 1, 44 (35.2%) were level 2, and 5 (4.0%) were level 3. There were no first-order questions requiring interpretation.

We collected student scores for all 125 assessment questions during 3 class years (2017, 2018, and 2019), comprising 417 students (138, 139, and 140 students, respectively) for a total of 52 125 scored attempts. Each of these attempts had been marked as correct or incorrect, and there were 44 890 correct answers overall, leading to an all-inclusive correct score of 86.1%. Percent correct scores were subsequently broken down according to question characteristics (Table 1). In univariate analysis, students scored significantly lower in MCQs as opposed to questions with matching answers ($P = .02$), in questions requiring interpretation as opposed to those testing simple recall of information ($P = .003$), and in questions of Bloom level 3

Table 1. Average Student Scores (Percent Correct) for Questions With Various Characteristics

	Average Score ± SD, %	*P* Value	
		Univariate	Multivariate
Question type			
Multiple choice	84.85 ± 0.14	.02[a]	.004[a]
Matching	92.06 ± 0.06		
Clinical vignette			
Absent	87.05 ± 0.12	.54	N/A
Present	85.54 ± 0.14		
Simple	86.22 ± 0.13	.48	N/A
Complex	75.73 ± 0.30		
Specimen image			
Absent	86.76 ± 0.12	.36	N/A
Present	84.20 ± 0.16		
Information depth			
Simple recall	86.83 ± 0.12	.003[a]	.003[a]
Interpretation	69.25 ± 0.28		
Knowledge density			
First order	86.63 ± 0.13	.59	N/A
Second order	85.33 ± 0.15		
Bloom taxonomy			
Level 1	86.63 ± 0.13	.01[a]	.79
Level 2	87.15 ± 0.11		
Level 3	69.25 ± 0.28		
Subject familiarity			
New information	84.20 ± 0.12	.001[a]	N/A
Repeated information	92.46 ± 0.08		
Some similarity	92.93 ± 0.07	.40	N/A
Verbatim	89.42 ± 0.12		

Abbreviation: N/A, not applicable.

[a] Statistically significant *P* values (ie, <.05).

Table 2. Distribution of Degrees of Difficulty Among Various Question Characteristics				
	Degree of Difficulty, No. (%)		*P* Value	
	Easy (n = 93)	Moderate–Hard (n = 32)	Univariate	Multivariate
Question type				
Multiple choice	72 (69.9)	31 (30.1)	.01[a]	.02[a]
Matching	21 (95.5)	1 (4.5)		
Clinical vignette				
Absent	39 (81.3)	9 (18.8)	.17	N/A
Present	54 (70.1)	23 (29.9)		
Simple	51 (70.8)	21 (29.2)	.61	N/A
Complex	3 (60.0)	2 (40.0)		
Specimen image				
Absent	71 (75.5)	23 (24.5)	.61	N/A
Present	22 (71.0)	9 (29.0)		
Information depth				
Simple recall	91 (75.8)	29 (24.2)	.07	.046[a]
Interpretation	2 (40.0)	3 (60.0)		
Knowledge density				
First order	58 (76.3)	18 (23.7)	.54	N/A
Second order	35 (71.4)	14 (28.5)		
Bloom taxonomy				
Level 1	58 (76.3)	18 (23.7)	.20	N/A
Level 2	33 (75.0)	11 (25.0)		
Level 3	2 (40.0)	3 (60.0)		
Subject familiarity				
New information	28 (70.0)	12 (30.0)	.10	N/A
Repeated information	26 (86.7)	4 (13.3)		
Some similarity	23 (88.5)	3 (11.5)	.46	N/A
Verbatim	3 (75.0)	1 (25.0)		

Abbreviation: N/A, not applicable.

[a] Statistically significant *P* values (ie, <.05).

($P = .01$). Students also had significantly higher scores in questions of the final examination containing repeated information from prior quizzes ($P = .001$), even though within this group, students did better on questions with some similarity as opposed to verbatim repeated questions (but the difference between the last 2 variables was not significant). Questions with lower mean scores also included those containing clinical vignettes, specimen images (whether gross or microscopic), and second-order questions, but these differences did not reach statistical significance. In multivariate analysis using only variables that reached significance (ie, question type, information depth, and Bloom taxonomy level, all with $P < .05$) and excluding subject familiarity (which concerned questions on the final exam only), both question type and information depth characteristics retained statistical significance. Students were significantly more likely to score higher in questions with matching answers compared with MCQs ($P = .004$) and in questions testing simple recall as opposed to those requiring some interpretation ($P = .003$).

Degree of difficulty (easy, moderate, or very difficult/hard) was determined for each assessment question as described in Materials and Methods, and questions were compared across their various characteristics in terms of being easy or moderate-hard (Table 2). In univariate analysis, MCQs were significantly more difficult than matching questions ($P = .01$), and questions assessing data interpretation were more likely to be of moderate difficulty or hard compared with questions testing simple recall, although this did not quite reach statistical significance ($P = .07$). Questions with clinical vignettes (especially complex ones) and specimen images, second-order questions and questions of higher Bloom taxonomy level were slightly more difficult, but these differences were not significant. Unsurprisingly, questions with a familiar subject (in terms of being repeated from earlier assessments) were easier (even though within this group, questions repeated verbatim from prior quizzes were harder), but this did not reach statistical significance. Question type and information depth (variables that reached or approached significance in univariate analysis, ie, $P < .10$) both retained statistical significance when compared in multivariate analysis. Multiple choice questions were significantly more difficult than matching questions ($P = .02$) and, when present, MCQs increased the likelihood of the degree of difficulty being designated as moderate or hard by an OR of 15.4 (CI, 1.4–166.7). Similarly, questions requiring data interpretation were significantly more likely to be designated as moderate or hard in terms of difficulty ($P = .046$) and, compared with simple recall questions, increased that likelihood with an OR of 12.7 (CI, 1.1–153.9).

Finally, the extent of item discrimination (very good, good, fair, or poor) was determined for each question as described in Materials and Methods and compared across assessment characteristics (Table 3). In univariate analysis,

Table 3. Extent of Item Discrimination for Questions With Various Characteristics				
	Item Discrimination, No. (%)		*P* Value	
	Good–Very Good (n = 40)	Fair–Poor (n = 85)	Univariate	Multivariate
Question type				
Multiple choice	27 (26.2)	76 (73.8)	.003[a]	.002[a]
Matching	13 (59.1)	9 (40.9)		
Clinical vignette				
Absent	16 (33.3)	32 (66.7)	.80	N/A
Present	24 (31.2)	53 (68.9)		
Simple	22 (30.5)	50 (69.5)	.66	N/A
Complex	2 (40.0)	3 (60.0)		
Specimen image				
Absent	26 (27.6)	68 (72.3)	.07	.04[a]
Present	14 (45.2)	17 (54.8)		
Information depth				
Simple recall	37 (30.8)	83 (69.2)	.17	N/A
Interpretation	3 (60.0)	2 (40.0)		
Knowledge density				
First order	21 (27.7)	55 (72.4)	.19	N/A
Second order	19 (38.7)	30 (61.2)		
Bloom taxonomy				
Level 1	21 (27.7)	55 (72.4)	.24	N/A
Level 2	16 (36.4)	28 (63.6)		
Level 3	3 (60.0)	2 (40.0)		
Subject familiarity				
New information	10 (25.0)	30 (75.0)	.62	N/A
Repeated information	6 (20.0)	24 (80.0)		
Some similarity	5 (19.2)	21 (80.8)	.79	N/A
Verbatim	1 (25.0)	3 (75.0)		

Abbreviation: N/A, not applicable.

[a] Statistically significant *P* values (ie, <.05).

matching type questions had better discrimination compared with MCQs ($P = .003$), as did questions with a specimen image, the latter approaching but not quite reaching significance ($P = .07$). The presence of a clinical vignette, information depth, degree of knowledge density, level of Bloom taxonomy, and subject familiarity did not significantly correlate with the discriminating ability of assessment questions. In multivariate analysis including question type and presence of a specimen image as variables ($P < .10$), both remained statistically significant. Questions requiring matching answers showed better ability at discrimination among students ($P = .002$) and when present, increased the likelihood of a question being able to have good or very good discrimination by an OR of 4.6 (CI, 1.7–12.4). Similarly, a specimen image increased the student discrimination of a given question ($P = .04$), and when present, increased the likelihood of good or very good discrimination by an OR of 2.6 (CI, 1.1–6.2).

DISCUSSION

In this study, we identified assessment question characteristics that are associated with student scores, degree of difficulty, and item discrimination during a first-year introductory course in pathology. Specifically, multivariate item-level analysis of 52 125 attempts showed that MCQs (versus matching answer) and those requiring some interpretation of data (as opposed to simple recall) resulted in lower scores and were more likely to be hard or moderately difficult. In addition, questions with matching answers and those including a gross or microscopic specimen image were more likely to be associated with good or very good discrimination among students.

As medical faculty aim to improve the quality of administered assessments, they may lack appropriate familiarity, expertise, and training in the composition and evaluation of exam questions and in the interpretation of performance data, such as item-level analysis.[36] In particular, it is proposed that MCQs may be able to provide adequate discrimination among students, but they are often thought to assess lower-order (recall) skills rather than higher-order thinking and comprehension. Thus, although item-level analysis reports provide data on the difficulty level and discrimination ability of assessment questions, guidelines have been developed to help faculty interpret and correct outliers based on this methodology. For example, at the University of Michigan Medical School, a document titled "Guidelines for Interpreting Item Analysis Reports" provides basic definitions of these metrics and recommendations for using the data contained therein during decision-making in the evaluation and revision of assessment questions.[36] These guidelines include reviewing items with a difficulty index ≤50% (particularly if also <0.50 in the discrimination scale), reviewing items with a nonsignificant (ie, poor) PBI (<0.20) if difficulty index is ≤75%, and

dropping items with a negative discrimination index. Based on these guidelines, 3 of 125 questions (2.4%) in our assessments would require a review/revision and 6 (4.8%) would need to be dropped. Although we did not retroactively apply any grade adjustments based on items that should have been dropped, it is our hope that future assessments based on guidelines such as these will lead to increased satisfaction among students and faculty in terms of grades accurately reflecting and evaluating knowledge.

Overall, calculated difficulty indices in our study tended to correlate with student scores across question types, and, for the most part, question types with increased difficulty also had better discrimination. Although MCQs, those requiring interpretation, Bloom taxonomy level 3 questions, and those testing new information were all associated with lower student scores, only MCQs and those necessitating information interpretation remained significantly associated with lower performance on multivariate analysis and were also independently associated with higher degrees of difficulty. Multiple choice questions have been associated with a phenomenon called cueing, which refers to the fact that answer choices, when provided, allow examinees to recognize the correct option when they may not have been able to do so otherwise, in the absence of answer options.[37] This would be especially problematic when measuring skills related to real-life diagnostic and clinical reasoning where answer options are not available and where premature decision-making may be detrimental to patients' health outcomes.[14] To ameliorate this, it has been proposed that matching answer-type questions are used, which minimize cueing (because all possible options are usually provided).[14] Our results indicate that matching answer questions also achieve better discrimination than MCQs, despite being associated with higher student scores and a lower degree of difficulty. This may be ideal because it attains performance discrimination among students without unnecessarily difficult questions that may otherwise discourage learner participation and engagement.

Prior studies have found that recall questions may be less difficult and less discriminating than questions with clinical vignettes during first-year medical school exams, even though these differences tend to disappear during the second year.[38] We similarly observed that simple recall questions were significantly less difficult, even though they were not less discriminating. An important difference, however, was that we compared simple recall questions to those requiring interpretation rather than vignette questions, and the latter categories did not align in our data, even when considering complex vignettes. Initial test question-writing guidelines seemed to favor MCQs with stems that were short, easy to read, and free of extraneous information, but which, however, may end up testing simple recall of facts rather than the desired higher-level thinking.[39] Thus, in recognition of the ambiguity and uncertainty often encountered in clinical care and of the need for students to learn to familiarize and acclimate themselves to this reality, the USMLE has transitioned to completely vignette-based stems.[40] Reflecting this change, most questions (62%) in our assessments featured a clinical vignette, which, however, did not result in greater difficulty or better discrimination among students. Interestingly, other studies have found that clinical vignettes in the question stem are associated with better discrimination but only among second-year, as opposed to first-year, medical students.[41] It would therefore be important to examine the frequency of questions with clinical vignettes as well as their relative contribution to student performance and item discrimination in other courses in our school, particularly as students progress through the medical education curriculum toward clerkships.

An interesting finding of this study concerns the fact that image-based questions, whether gross or microscopic, were independently correlated with better student discrimination, even though they were not significantly associated with lower student scores or with higher degrees of difficulty. It is tempting to speculate that the elucidation of medical data depicted in a specimen image may require a different set of skills that, although not directly associated with poorer scores, may identify students who are able to better analyze visual information. Interestingly, this may be specific to medical specialties, such as pathology and radiology, that rely heavily upon such types of interpretation. The absence of images from medical school assessments (even in our pathology course only 25% of all test questions contained an image) and students' discomfort in their interpretation may exacerbate some of the deficiencies of medical graduates as they pursue these particular specialties.[42] Importantly, a question with an image may still test memorization skills (ie, simple recall) rather than deductive reasoning if it includes an image that was previously shown, for example in a lab or lecture.[16] Importantly, of 31 questions with an image in our assessments, 4 (12.9%) required interpretation (versus simple recall), significantly more than the 1 of 94 questions (1.1%) without an image requiring interpretation ($P = .02$), perhaps suggesting a reason for the better discrimination observed with these types of questions in our study. Underscoring the hypothesis that images may be particularly important for item discrimination but only in medical disciplines concerned with disease diagnosis, such as pathology, a study of normal histology MCQs found no differences in difficulty or discrimination between questions with an illustration and those without.[43]

The Bloom taxonomy has been often used to assess students' critical thinking skills and to determine whether higher-order MCQs encourage deeper conceptual understanding of medical and scientific concepts.[16] Some authors have found that higher-level Bloom taxonomy items correlate with higher discrimination and difficulty indices, particularly in earlier, formative assessments.[33] Although they are hierarchically ranked, the various levels in a Bloom taxonomy are not meant to be directly analogous to degrees of difficulty.[44] Thus, just because a question is testing basic recall does not necessarily mean it has to be easier than a higher-order question, particularly if the content being assessed is obscure, detailed, or esoteric.[16] Similarly, item discrimination should not automatically improve with questions of a higher Bloom taxonomy level. The intent of Bloom taxonomies is to measure the cognitive level associated or targeted with each question and only secondarily to determine whether students are having difficulty with such higher-level questions. Underscoring this fact, our data suggest that, although students performed worse on average with questions of higher levels in our Bloom taxonomy, these did not have higher degrees of difficulty and were not better able to discriminate between students.

The criteria that define assessments of good quality are continuously updated, particularly as they pertain to the ultimate purpose of the assessment (ie, whether summative or formative) and the vested interest of respective stake-

holders (eg, students, faculty, administration).[45] Thus, although particular characteristics, such as exam length or test-taker cohort size, may influence the evaluation of individual assessments, item difficulty and discrimination are universally accepted as quality criteria in medical literature.[46–48] Nevertheless, alternative approaches, such as using open-book tests for formative assessment or employing different question formats (eg, essay, open-ended, uncued) have been recommended for use, especially in order to test higher-level problem-solving skills and provide effective discrimination among students.[49,50] However, it is still not entirely clear that these types of questions offer a better appraisal of higher-order cognitive functioning or contribute to assessments of higher validity.[51] Ultimately, the goal should be assessments with a combination of questions of varying difficulty and item discrimination that are able to identify both students struggling with the material as well as concepts that are not effectively taught. In that respect, assessments should be specifically connected to articulated medical education program objectives and associated with efforts to increase individual student learning ("assessment *for* learning").[52]

This study has a few limitations. Given its design nature (retrospective data collection from a single institution), the study's findings may not be easily generalizable. Differences in frequency, content, and length of assessments may prohibit the application of our conclusions to other medical education settings. This is especially true because most schools do not offer an introductory, stand-alone course in general pathology.[22,53] Furthermore, because grading scales are not uniform between medical schools, inferences about degree of difficulty and item discrimination may not be transferable to other courses. Nevertheless, studies have found that preclinical course grades in pathology are the strongest predictor of performance in USMLE Step exams, and thus evaluating pathology assessments may offer precious insight on students' future achievement.[54] On the other hand, our study evaluated question characteristics as they apply specifically to undergraduate medical education in pathology (eg, in terms of including a specimen image), and our conclusions may be valuable to other pathology instructors. We tried to minimize sampling, selection, and reporting bias by including all assessment questions in our analysis, by using standardized definitions for difficulty and discrimination indices, and by using an objective algorithm for the assignment of Bloom taxonomy levels. Finally, the inclusion of a large number of scored attempts and the use of robust statistical analysis allowed us to interpret our results with some confidence.

In conclusion, our findings suggest that certain types of assessment questions, particularly MCQs and those testing deeper levels of information (ie, interpretation) are associated with lower student scores and higher degrees of difficulty whereas others, such as matching answer-type and those containing a specimen image, are better at discriminating among students' performances. Therefore, assessments in medical education should include various combinations of question types as needed in order to achieve optimal levels of student performance in terms of difficulty and discrimination.

References

1. Albarqouni L, Hoffmann T, Glasziou P. Evidence-based practice educational intervention studies: a systematic review of what is taught and how it is measured. *BMC Med Educ*. 2018;18(1):177.
2. Claramita M, Setiawati EP, Kristina TN, Emilia O, van der Vleuten C. Community-based educational design for undergraduate medical education: a grounded theory study. *BMC Med Educ*. 2019;19(1):258.
3. Gonzalo JD, Chang A, Dekhtyar M, Starr SR, Holmboe E, Wolpaw DR. Health systems science in medical education: unifying the components to catalyze transformation. *Acad Med*. 2020;95(9):1362–1372.
4. Lewis JH, Lage OG, Grant BK, et al. Addressing the social determinants of health in undergraduate medical education curricula: a survey report. *Adv Med Educ Pract*. 2020;11:369–377.
5. Ma J, Stahl L, Knotts E. Emerging roles of health information professionals for library and information science curriculum development: a scoping review. *J Med Libr Assoc*. 2018;106(4):432–444.
6. Zanting A, Meershoek A, Frambach JM, Krumeich A. The 'exotic other' in medical curricula: rethinking cultural diversity in course manuals. *Med Teach*. 2020;42(7):791–798.
7. Ellis PM, Wilkinson TJ, Hu WC. Differences between medical school and PGY1 learning outcomes: an explanation for new graduates not being "work ready"? *Med Teach*. 2020;42(9):1043–1050.
8. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. *BMC Med Educ*. 2019;19(1):460.
9. van Schaik S, Plant J, O'Sullivan P. Promoting self-directed learning through portfolios in undergraduate medical education: the mentors' perspective. *Med Teach*. 2013;35(2):139–144.
10. Patell R, Raska P, Lee N, et al. Development and validation of a test for competence in evidence-based medicine. *J Gen Intern Med*. 2020;35(5):1530–1536.
11. Reid WA, Duvall E, Evans P. Can we influence medical students' approaches to learning? *Med Teach*. 2005;27(5):401–407.
12. Cooke M, Irby DM, Sullivan W, Ludmerer KM. American medical education 100 years after the Flexner report. *N Engl J Med*. 2006;355(13):1339–1344.
13. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA*. 2002;287(2):226–235.
14. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–396.
15. Velan GM, Jones P, McNeil HP, Kumar RK. Integrated online formative assessments in the biomedical sciences for medical students: benefits for learning. *BMC Med Educ*. 2008;8:52.
16. Zaidi NLB, Grob KL, Monrad SM, et al. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? *Acad Med*. 2018;93(6):856–859.
17. Paniagua MA, Swygert KA, eds. *Constructing Written Test Questions For the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners; 2016. https://www.nbme.org/sites/default/files/2020-01/IWW_Gold_Book.pdf. Accessed August 2020.
18. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ Theory Pract*. 2006;11(1):61–68.
19. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. *Acad Med*. 2002;77(2):156–161.
20. Marshall R, Cartwright N, Mattick K. Teaching and learning pathology: a critical review of the English literature. *Med Educ*. 2004;38(3):302–313.
21. Taylor CR, DeYoung BR, Cohen MB. Pathology education: quo vadis? *Hum Pathol*. 2008;39(11):1555–1561.
22. Sadofsky M, Knollmann-Ritschel B, Conran RM, Prystowsky MB. National standards in pathology education: developing competencies for integrated medical school curricula. *Arch Pathol Lab Med*. 2014;138(3):328–332.
23. King TS, Sharma R, Jackson J, Fiebelkorn KR. Clinical case-based image portfolios in medical histopathology. *Anat Sci Educ*. 2019;12(2):200–209.
24. Koles P, Nelson S, Stolfi A, Parmelee D, Destephen D. Active learning in a year 2 pathology curriculum. *Med Educ*. 2005;39(10):1045–1055.
25. Koles PG, Stolfi A, Borges NJ, Nelson S, Parmelee DX. The impact of team-based learning on medical students' academic performance. *Acad Med*. 2010;85(11):1739–1745.
26. Hwang JE, Kim NJ, Song M, et al. Individual class evaluation and effective teaching characteristics in integrated curricula. *BMC Med Educ*. 2017;17(1):252.
27. Talaulikar D, Akerlind G, Potter JM. Bench work and clinical relevance: a new strategy in pathology education. *Pathology*. 2008;40(7):707–710.
28. Tan KB. Some variations of case-based techniques for the teaching of undergraduate pathology. *Malays J Pathol*. 2005;27(2):127–128.
29. McBrien S, Bailey Z, Ryder J, Scholer P, Talmon G. Improving outcomes. *Am J Clin Pathol*. 2019;152(6):775–781.
30. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. *Med Educ*. 2010;44(1):109–117.
31. McDonald ME. *The Nurse Educator's Guide to Assessing Learning Outcomes*. 4th ed. Sudbury, MA: Jones & Bartlett Learning; 2018.
32. McGahee TW, Ball J. How to read and really use an item analysis. *Nurse Educ*. 2009;34(4):166–171.
33. Zaidi NB, Hwang C, Scott S, Stallard S, Purkiss J, Hortsch M. Climbing Bloom's taxonomy pyramid: lessons from a graduate histology course. *Anat Sci Educ*. 2017;10(5):456–464.
34. Thompson AR, O'Loughlin VD. 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*. 2015;8(6):493–501.

35. Armstrong P. Bloom's taxonomy. Vanderbilt University Center for Teaching. https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/. Accessed August 2020.

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

37. Veloski JJ, Rabinowitz HK, Robeson MR. A solution to the cueing effects of multiple choice questions: the Un-Q format. *Med Educ*. 1993;27(4):371–375.

38. Nazian S, Stevenson F. Difficulty and discriminative ability of various categories of multiple choice questions in medical school preclerkship examinations. *Med Sci Educ*. 2014;24:387–393.

39. Haas MJ, Dragan YP, Hikita H, et al. Transgene expression and repression in transgenic rats bearing the phosphoenolpyruvate carboxykinase-simian virus 40 T antigen or the phosphoenolpyruvate carboxykinase-transforming growth factor-alpha constructs. *Am J Pathol*. 1999;155(1):183–192.

40. Case SM, Swanson DB, Becker DF. Verbosity, window dressing, and red herrings: do they make a better test item? *Acad Med*. 1996;71(10 suppl):S28–S30.

41. Ikah DS, Finn GM, Swamy M, White PM, McLachlan JC. Clinical vignettes improve performance in anatomy practical assessment. *Anat Sci Educ*. 2015;8(3):221–229.

42. Naritoku WY, Vasovic L, Steinberg JJ, Prystowsky MB, Powell SZ. Anatomic and clinical pathology boot camps: filling pathology-specific gaps in undergraduate medical education. *Arch Pathol Lab Med*. 2014;138(3):316–321.

43. Holland J, O'Sullivan R, Arnett R. Is a picture worth a thousand words: an analysis of the difficulty and discrimination parameters of illustrated vs. text-alone vignettes in histology multiple choice questions. *BMC Med Educ*. 2015;15:184.

44. Thompson AR, Braun MW, O'Loughlin VD. A comparison of student performance on discipline-specific versus integrated exams in a medical school course. *Adv Physiol Educ*. 2013;37(4):370–376.

45. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206–214.

46. Aubin AS, Young M, Eva K, St-Onge C. Examinee cohort size and item analysis guidelines for health professions education programs: a Monte Carlo simulation study. *Acad Med*. 2020;95(1):151–156.

47. Kirschstein T, Wolters A, Lenz JH, et al. An algorithm for calculating exam quality as a basis for performance-based allocation of funds at medical schools. *GMS J Med Educ*. 2016;33(3):Doc44.

48. Young M, Cummings BA, St-Onge C. Ensuring the quality of multiple-choice exams administered to small cohorts: a cautionary tale. *Perspect Med Educ*. 2017;6(1):21–28.

49. Damjanov I, Fenderson BA, Veloski JJ, Rubin E. Testing of medical students with open-ended, uncued questions. *Hum Pathol*. 1995;26(4):362–365.

50. Magid MS, Schindler MK. Weekly open-book open-access computer-based quizzes for formative assessment in a medical school general pathology course. *Med Sci Educ*. 2007;17(1):45–51.

51. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Med Educ*. 2014;14:249.

52. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478–485.

53. Haspel RL, Bhargava P, Gilmore H, et al. Successful implementation of a longitudinal, integrated pathology curriculum during the third year of medical school. *Arch Pathol Lab Med*. 2012;136(11):1430–1436.

54. Hu Y, Martindale JR, LeGallo RD, White CB, McGahren ED, Schroen AT. Relationships between preclinical course grades and standardized exam performance. *Adv Health Sci Educ Theory Pract*. 2016;21(2):389–399.