


Assessing the Difficulty and Long-Term Retention of Factual and Conceptual Knowledge Through Multiple-Choice Questions: A Longitudinal Study

Neil G Haycocks^{1,2}, Jessica Hernandez-Moreno³, Johan C Bester⁴, Robert Hernandez Jnr³, Rosalie Kalili³, Daman Samrao³, Edward Simanton³, Thomas A Vida³ 

¹Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT, USA; ²Utah Office of the Medical Examiner, Taylorsville, UT, USA; ³Department of Medical Education, Kirk Kerkorian School of Medicine at UNLV, Las Vegas, NV, USA; ⁴Office of Curricular Affairs and Department of Family and Community Medicine, Saint Louis University School of Medicine, Saint Louis, MO, USA

Correspondence: Thomas A Vida, Email thomas.vida@unlv.edu

Purpose: Multiple choice questions (MCQs) are the mainstay in examinations for medical education, physician licensing, and board certification. Traditionally, MCQs tend to test rote recall of memorized facts. Their utility in assessing higher cognitive functions has been more problematic to determine. This work evaluates and compares the difficulty and long-term retention of factual versus conceptual knowledge using multiple-choice questions in a longitudinal study.

Patients and Methods: We classified a series of MCQs into two groups to test recall/verbatim and conceptual/inferential thinking, respectively. We used the MCQs to test a two-part hypothesis: 1) scores for recall/verbatim questions would be significantly higher than for concept/inference questions, and 2) memory loss over time would be more significant for factual knowledge than conceptual understanding compared with a loss in the ability to reason about concepts critically. We first used the MCQs with pre-clinical medical students on a summative exam in 2020, which served as a retrospective benchmark of their performance characteristics. After two years, the same questions were re-administered to volunteers from the same cohort of students in 2020.

Results: Retrospective analysis revealed that recall/verbatim questions were answered correctly more frequently (82.0% vs 60.9%, $P = 0.002$). Performance on concept/inference questions showed a significant decline, but a larger decline was observed for recall/verbatim questions after two years. Performance on concept/inference questions showed a slight decline across quartiles, while two years later, recall/verbatim questions experienced substantial performance loss. Subgroup analysis indicated convergence in performance on both question types, suggesting that the clinical relevance of the MCQ content may have influenced a regression toward a baseline mean.

Conclusion: These findings suggest conceptual/inferential thinking is more complex than rote memorization. However, the knowledge acquired is more durable in a longitudinal fashion, especially if it is reinforced in clinical settings.

Keywords: medical education assessment, multiple choice questions (MCQs), long-term knowledge retention, conceptual learning, cognitive function testing

Introduction

Multiple choice questions (MCQs) are ubiquitous in the medical profession. In the United States, the Medical College Admissions Test (MCAT), the United States Medical Licensing Examination (USMLE) examinations, specialty-specific board certification examinations, and other essential and clinical sciences examinations are built using MCQs. We may take them for granted today, but in previous decades, the usefulness of MCQs in clinical medical education was disputed.¹⁻³ Subsequently, data emerged that MCQs could be reliable, valid, and efficient,^{4,5} but questions persisted about the ability of MCQs to evaluate cognitive levels above rote recall.⁶ Several studies have attempted to classify MCQs and link them to the six cognitive domains described in Bloom's taxonomy with varying levels of success.⁷ This stratification of question content predates contemporary neuroscientific understanding of memory formation and retrieval

with even the most basic mapping of brain network functionality to learning (for review, see.⁸ Bloom's taxonomy does not adequately capture the complexities of learning in the brain, even in revision,⁹ making it difficult to demonstrate this framework's effectiveness in assessing learning.^{6,10,11}

Recent research in medical education highlights the impact of various educational strategies on knowledge retention, emphasizing the value of theoretical and practical approaches. Long-term knowledge retention is generally higher with active, blended-learning approaches than passive learning methods, with some variations in retention rates between student groups.¹² Spaced and repeated testing and retrieval-based formats have consistently outperformed mass practice or cramming, improving long-term retention.^{13,14} Feedback further solidifies learning, as knowledge retention can decline rapidly within days to weeks without reinforcement.¹⁵ Problem-based learning (PBL) and small group learning promise to enhance problem-solving skills and retention,^{16,17} while spaced education and distributed practice effectively improve clinical performance.¹⁸ Active learning strategies, though challenging to implement, have been recognized for their benefits,¹⁹ with collaborative reflection gaining traction as a method to engage students.²⁰ Podcasts, smartphones, and clinical applications have emerged as valuable tools for contemporary medical education, offering alternative and flexible learning modalities.^{14,21,22} Incorporating intentional difficulties, or "desirable difficulties",²³ or knowledge decrement²⁴ has also proven effective in enhancing long-term retention, while strategies focusing on rural workforce retention have demonstrated considerable success.²⁵ For long-term retention of medical knowledge, multiple-choice (MCQs) and short-answer (SAQs) improve delayed retention compared to no testing.^{26,27} While MCQs offer high reliability and easy marking, SAQs may require more effortful retrieval and promote better retention.^{28,29} Repeated testing enhances long-term retention compared to repeated study.³⁰ To construct high-quality MCQs, educators should focus on assessing higher-order thinking skills and use test blueprints.³¹ Ultimately, varied teaching approaches and well-prepared lectures are crucial for engaging students and fostering durable knowledge retention.^{32,33}

A challenge related to the widespread use of MCQs in medical education is the issue of assessing different levels of cognitive function: that is, testing critical thinking and understanding as opposed to rote memorization. To test critical thinking, a MCQ often needs considerable revision, which can be successful if the MCQ is crafted appropriately.³⁴ This can be coupled to the systematic design of instruction that typically begins with objectives and ends with an assessment to determine if the objectives (ie learning outcomes) have been achieved. Promoting higher-order thinking is desirable in contemporary pre-clinical curricula given the complexity of clinical medicine. In medical education, a shift from memorized facts to applied concepts occurs in the clinical years where diagnosis and treatment skills are honed. The cognitive scientific principles at play during this transition are complex and ultimately give rise to melding the art and science of medicine.³⁵ Therefore, an ideal MCQ-based exam should use questions that provide a meaningful assessment of higher cognitive functions, testing understanding and applying knowledge beyond rote memorization. Studies suggest that examinations should prioritize comprehension-level (higher-order thinking) questions over knowledge-level questions to better predict successful learning, according to Bloom's taxonomy.³⁶

Assessing higher cognitive function also has possible implications for education research. For example, the *posttest-only control group* design is a gold-standard intervention study design commonly found in such research.³⁷ Briefly, an intervention is administered to an experimental group but not a control group. A comparative assessment of the two groups, which may take the form of MCQs, then tests the effect of the intervention. Since MCQs can be written to different difficulty levels and potentially assess various levels of cognitive function, they introduce an additional variable to the experimental design and a potential source of measurement bias. The selection of MCQs for comparative assessments may influence whether an effect is observed, how large it is, and how it evolves.

Karpicke and Blunt report such a finding,³⁸ where in certain experimental groups, the posttest questions that they classified as "inference" showed superior recall than those classified as "verbatim". While not the focus of that particular study, the observation is nonetheless intriguing. Zaidi et al showed a related observation using a dichotomized Bloom's taxonomy to classify MCQs as "lower order" and "higher order" based on cognitive level.³⁹ They found a modest but statistically significant difference in performance on the two question types, with lower-order questions answered correctly more often than higher-order questions. Both studies suggest it is possible to differentiate MCQs based on function: those designed to test memorization and others designed to test deeper conceptual understanding.

Our study used Karpicke and Blunt's³⁸ work as a starting point to develop a classification scheme for medically relevant MCQs. We created three MCQ categories, including (1) recall/verbatim, (2) concept/inference, and (3) mixed/ambiguous. The hallmark of recall/verbatim questions is that they are based primarily on facts and assess whether a learner has retained relevant information. In contrast, concept/inference questions are based primarily on relationships and, therefore, consider whether a learner knows how facts are connected, which provides greater context and enables the ability to make predictions. In the present study, we added a third category, mixed/ambiguous, for MCQs that could be answered with the recall of facts or inference and interpretation. This study aimed to assess whether conceptual/inferential knowledge is retained longer than factual/verbatim knowledge among medical students over two years.

We formed and tested a two-part hypothesis using this classification system. First, after preparing for an exam, we posited that 1) scores for recall/verbatim questions would be significantly higher than for concept/inference questions, and 2) memory loss over time would be more significant for factual knowledge than conceptual understanding compared to a loss in the ability to reason about concepts critically. This hypothesis predicts that concept/inference questions will generally be more challenging based on the relatively more sophisticated cognitive processes needed to answer them. Further, the erosion over a defined time interval of conceptual knowledge, as assessed with concept/inference questions, will be slower than the erosion of straight factual knowledge, as assessed with recall/verbatim questions. This part of the hypothesis predicts that conceptual knowledge is relatively complex to acquire but, once encoded, is more durable, especially in a longitudinal manner.

Materials and Methods

Participants and Procedure

Question Categorization

The Kirk Kerkorian School of Medicine curriculum includes a comprehensive coverage of basic and clinical sciences, and questions were selected from summative examinations covering critical pulmonary and renal physiology topics. Forty-five questions were extracted from a faculty-authored summative examination. One of the authors (NH) developed a system to categorize each question as (1) recall/verbatim or (2) concept/inference. A third category, mixed/ambiguous, was for questions that could be answered using recall or inference. Seven faculty members, representing a mixture of MDs and PhDs with content expertise in assessment, pulmonary and renal physiology, and curriculum development, reviewed the questions independently and as a group, then reached a consensus on the best category for each question.

Recall/verbatim questions were categorized as having relatively simple cognitive processes to answer. These include straightforward recall of simple facts, solving arithmetic equations, recognizing basic patterns, or deductive reasoning that requires little or no content mastery. Verbatim/inquiry questions involving more esoteric and less familiar information are expected to have relatively high difficulty indices (the percentage of students who answer a question correctly). A question may also be classified as recall/verbatim if answering it requires recall of a specific, non-intuitive piece of information, regardless of what other cognitive processes are involved. Concept/inference questions require fairly complex cognitive processes to answer them, which commonly involve deductive reasoning to discern or predict relationships, recognizing subtle patterns, and solving problems that require substantial practice and content mastery. Mixed/ambiguous questions are answerable through more than one pathway. For example, diagnosing an acid-base disorder based on laboratory values can be done through logical inference or straight pattern recognition.

Assessment Implementation

The categorization system described above was first applied retrospectively. Performance data on each question were collected from medical student results in a summative pre-clinical basic science exam in April 2020. The same questions were then re-administered to volunteers from the same cohort of medical students in January-February 2022, after they had progressed through most of the clerkship year. To reduce the testing burden, the questions were broken into four quizzes (acid-base physiology, renal physiology, pulmonary physiology, and pulmonary/renal anatomy). To be included in the analysis, a medical student must have taken the 2020 exam and completed at least one 2022 quiz.

Ethical Considerations

Ethical approval was obtained from the University at Las Vegas Biomedical IRB on April 3, 2017 (Protocol 1030906–1). The approval permits the use of aggregated, de-identified data. The data was primarily collected to assess and inform curriculum, specifically to inform curricular use of faculty-authored examinations. The de-identified data was then used secondarily for research, as per the protocol, and is not considered human subjects' research. The UNLV Biomedical Institutional Review Board designated that no individual informed consent is required as per an excluded protocol since this is not human subjects' research. Further, all procedures performed in studies involving human participants were following the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards (Goodyear et al, 2007).

Statistical Analysis

All scores from student assessments were individually de-identified after the 2020 and 2022 data were linked to avoid bias and preserve anonymity. The included students ($n = 56$), from a total cohort of 120 pre-clinical students, were organized into subgroups (top/bottom half or quartiles) based on their overall basic science exam scores relative to the class median of 84.1%. Statistical analyses were performed using paired Student's *t*-tests with a two-tailed distribution and two-sample equal variance. The paired *t*-tests were used as the same group of students was assessed at two different time points (2020 and 2022). Such analysis is warranted, given we are comparing the means of two groups.^{40,41} Using paired *t*-tests has established statistical significance in various medical education assessment studies.^{42–45} A homoscedastic linear regression model⁴⁶ was applied to analyze the relationship between student performance on different MCQ types and the associated variables over time. This model assumes constant variance (homoscedasticity) across all levels of the independent variables, making it appropriate for examining the stable patterns in student performance without introducing bias due to unequal variances.^{47,48} The model helped evaluate how performance on recall/verbatim and concept/inference questions changed between 2020 and 2022, and whether these changes were influenced by factors such as student quartile and overall test difficulty. All calculations were performed using Microsoft Excel's regression analysis functions, which allowed for precise computation of *p*-values, confidence intervals, and other relevant statistical measures. A *P*-value threshold of 0.050 was used to determine statistical significance, ensuring a rigorous interpretation of the results.

Results

Question Categorization and Initial Performance

Using a basic framework (Table 1 and Table 2 for examples), the consensus of seven faculty members categorized 10 questions (22%) as concept/inference, 33 questions (73%) as recall/verbatim, and 2 (4%) as mixed/ambiguous. The mixed/ambiguous questions were excluded from further analysis.

Performance on recall/verbatim questions was significantly higher than on concept/inference questions, but concept/inference questions were more effective in differentiating student performance overall, and this differentiation persisted over time (Table 3). A retrospective analysis of question performance on the 2020 exam showed the recall/verbatim questions were answered correctly more often than the concept/inference questions (82.0% vs 60.9%, $P = 0.002$). When the top and bottom halves of the class were compared, they answered recall/verbatim questions correctly at approximately the same

Table 1 Basic Framework for Categorizing Multiple Choice Questions

Processing	Information	Category
Simple	Familiar	Recall/Verbatim (easier)
	Esoteric	Recall/Verbatim (more difficult)
Complex	Familiar	Concept/Inference
	Esoteric	Recall/Verbatim (more difficult)

Table 2 Example Recall/Verbatim and Concept/Inference Questions

Recall/Verbatim	Recall/Verbatim	Concept/Inference	Concept/Inference
<p>Rationale: The question is primarily organized around factual recall: major toxic component of car exhaust (carbon monoxide) and properties of carboxyhemoglobin.</p> <p>A 45-year-old man with a history of depression and substance abuse attempts suicide by sitting in his idling car while enclosed in a garage. His teenage daughter finds him and summons an ambulance. Which of the following most accurately describes the effect of the car exhaust on his hemoglobin?</p> <p>A. Decreases affinity for oxygen, no change to oxygen saturation B. Increases affinity for oxygen, no change to oxygen saturation C. No change to affinity for oxygen, decreases oxygen saturation D. Decreases affinity for oxygen, decreases oxygen saturation E. Increases affinity for oxygen, decreases oxygen saturation</p>	<p>Rationale: The question requires recall of the equation for calculating renal clearance. The remaining arithmetic operations require simple processing.</p> <p>A healthy 23-year-old man takes part of a research study. He is injected with an experimental drug X that has no known hepatic metabolism. A 24-hour urine collection amounts to 1.72 L with a urine [X] of 500 mg/L. His plasma [X] is 30 mg/L. Calculate the renal clearance of drug X.</p> <p>A. 10 mL/min B. 20 mL/min C. 30 mL/min D. 40 mL/min E. 50 mL/min</p>	<p>Rationale: The question requires conceptualization of different physiologic and pathophysiologic states, judging how they will be affected by changing one variable (alveolar oxygen concentration).</p> <p>In which of the following conditions will supplemental oxygen be least helpful in raising the PO₂ of systemic arterial blood?</p> <p>A. Ventral septal defect (VSD) B. Respiratory depression from drug overdose C. Pulmonary fibrosis D. Pulmonary embolism E. High altitude</p>	<p>Rationale: The question requires a conceptual understanding of how and why the kidney modifies urine osmolality, ie, the purpose of the countercurrent multiplication mechanism, which informs its anatomic organization.</p> <p>A 32-year-old man goes for a hike in Death Valley National Park in July. By the time he finishes the hike he is fatigued, dizzy, and extremely thirsty. The osmolality of his urine will be most similar to the osmolality of his:</p> <p>A. Distal convoluted tubules B. Outer medullary collecting ducts C. Plasma D. Proximal convoluted tubules E. Renal papillae</p>

Notes: the familiar recall/verbatim questions tend to ask “what”, while esoteric concept/inference questions tend to ask “why” or for the application of knowledge.

Table 3 Comparative Performance by Question Type and Year

	Percent Correct		P-value
	Recall/Verbatim	Concept/Inference	
2020 Exam			
All Participants (n = 56)	82.0	60.9	0.002
Top Half (n = 28)	84.0	68.6	0.020
Bottom half (n = 28)	80.2	53.2	0.001
First Quartile (n = 14)	84.4	70.7	0.060 (NS)
Second Quartile (n = 14)	83.5	66.4	0.009
Third Quartile (n = 14)	80.7	57.9	0.003
Fourth Quartile (n = 14)	79.4	48.6	0.001
2022 Quizzes			
All Participants (n = 56)	59.8	42.9	0.079 (NS)
Top Half (n = 28)	62.5	44.2	0.078 (NS)
Bottom half (n = 28)	57.1	41.7	0.110 (NS)
First Quartile (n = 14)	64.1	49.7	0.161 (NS)
Second Quartile (n = 14)	61.2	38.8	0.044
Third Quartile (n = 14)	54.3	39.5	0.135 (NS)
Fourth Quartile (n = 14)	59.8	43.5	0.118 (NS)

Notes: percentages are calculated from the total number of quizzes attempted by each subgroup.

Abbreviation: NS, not significant.

frequency (84.0% vs 80.2%, $P = 0.433$). The top half answered concept/inference questions correctly more often than the bottom half (68.6% vs 53.2%, $P = 0.027$). Further delineation showed a progressive decline in performance on concept/inference questions by quartile, while performance on recall/verbatim questions was relatively stable (Figure 1).

Performance on Recall/Verbatim and Concept/Inference Questions

Analysis of question performance on the 2022 quizzes showed no statistically significant difference between performance on recall/verbatim questions versus concept/inference questions (59.8% vs 42.9%, $P = 0.079$). Likewise, no observed significant difference occurred between the top and bottom halves of the class on recall/verbatim questions (62.5% vs 57.1%, $P = 0.432$) or concept/inference questions (44.2% vs 41.7%, $P = 0.827$).

Comparison of Knowledge Retention Over Time

A comparison of question performance between 2020 and 2022 showed a statistically significant decrease in performance on recall/verbatim questions (82.0% vs 59.8%, $P < 0.001$) and a significant decrease in performance on concept/inference questions ($P = 0.050$) (Figure 2a). The decrease on recall/verbatim questions was similar for both the top half (84.0% vs 62.5%, $P < 0.001$) and the bottom half (80.1% vs 57.1%, $P < 0.001$) of participants. A statistically significant decrease was observed on concept/inference questions for only the top half (68.6% vs 44.2%, $P = 0.024$). No significant difference was found on concept/inference questions for the bottom half (53.2% vs 41.7%, $P = 0.204$) (Figure 2b).

Subgroup Analysis of Performance Decline

The decline in performance was different for all analyzed subgroups. The decline in recall/verbatim questions was similar between the top and bottom halves of the class. However, this was not observed for concept/inference questions. A statistically significant decrease of 24.4 points ($P = 0.024$) was observed for the top half of students on concept/inference questions and the bottom half averaged a statistically insignificant 11.5-point decrease ($P = 0.204$). Table 4 shows that the decline in performance on concept/inference questions was significant for the top half of the class. In contrast, performance on recall/verbatim questions declined equally for all groups. The gap between these two groups averaged 15.4 points in 2020, and this difference narrowed to a statistically insignificant 2.5 points in 2022 ($P = 0.827$). In fact, by 2022, the performance of both groups on both question types had become statistically indistinguishable in all but one comparison (top half, recall/verbatim questions vs bottom half, concept/inference questions ($P = 0.036$)).

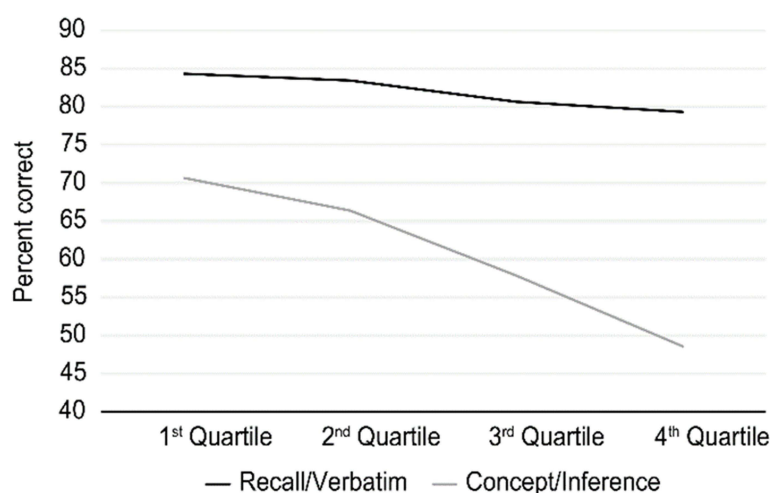


Figure 1 Baseline Performance on Multiple Choice Questions by Question type and Quartile. A summative exam with multiple choice questions was administered to 56 medical students in 2020. A panel of medical school faculty retrospectively categorized the questions as recall/verbatim or concept/inference using a metric described in Materials and Methods. The percent correct was then computed and expressed by quartile. Question performance on the 2020 exam showed the recall/verbatim questions was answered correctly more often than the concept/inference questions (this is a graphical depiction of data in Table 3).

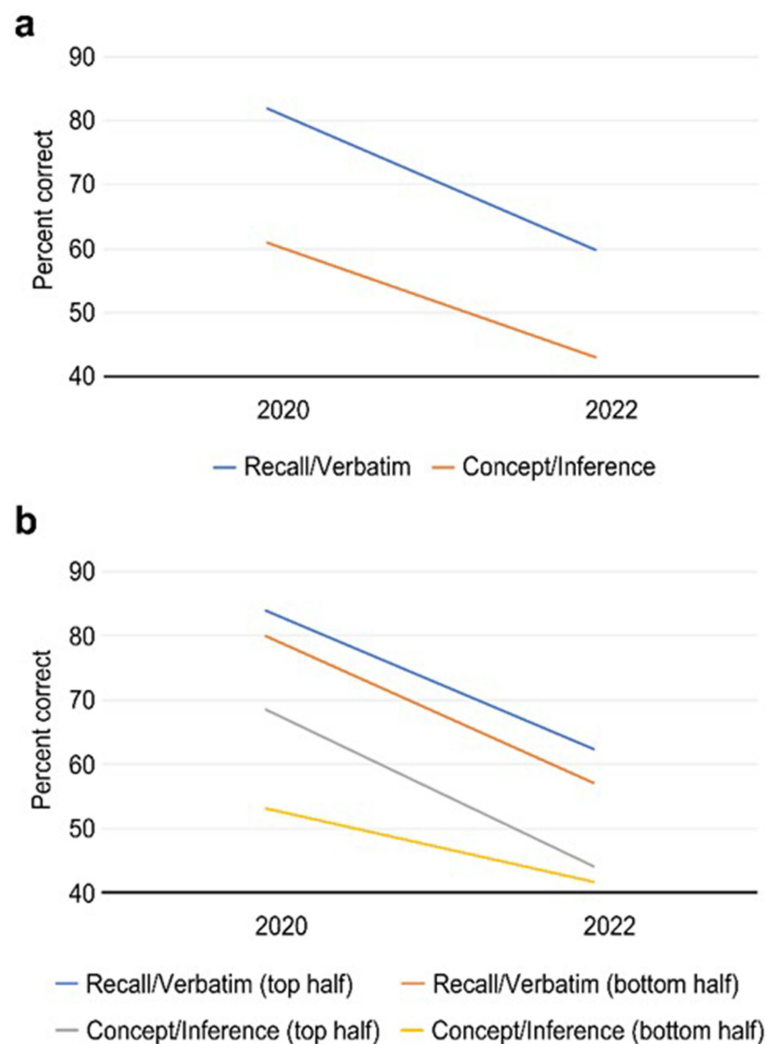


Figure 2 Comparison of Exam Question Performance over Two Years. The same questions used to measure baseline performance were administered a second time two years later to the same cohort of students (see Figure 1). (a) Depicts the percent correct for all students over two years using recall/verbatim or concept/inference questions. (b) Depicts the percent correct for the top and bottom half of the students over two years using recall/verbatim or concept/inference questions. The decrease in performance on recall/verbatim questions was similar for both the top and bottom halves. A statistically significant decrease in performance occurred on concept/inference questions for only the top half and no significant difference was found on concept/inference questions for the bottom half (this is a graphical depiction of data in Table 3).

Discussion

This study investigated the performance and retention of medical students on multiple-choice questions (MCQs) that assessed either recall/verbatim knowledge or conceptual/inferential reasoning. The results demonstrate that while students performed better on recall/verbatim questions initially, their performance on these questions declined significantly over two years. In contrast, conceptual/inferential questions, although more challenging initially, showed a smaller decline in performance over time. These findings suggest that conceptual understanding, though harder to acquire, is more resilient to memory decay and is potentially reinforced by clinical experience. This has important implications for the design of medical education assessments, emphasizing the value of testing higher-order cognitive skills to promote long-term knowledge retention.

Question Performance

Our first hypothesis was that recall/verbatim and concept/inference questions would have significantly different performance characteristics. The data from the 2020 exam support this hypothesis, with concept/inference questions having a substantially higher difficulty index than recall/verbatim questions. Moreover, all students had similar performance on

Table 4 Statistical Comparisons (P-values) of Top/Bottom Half Subgroups by Question Type and Year

			2020				2022			
			Recall/Verbatim		Concept/Inference		Recall/Verbatim		Concept/Inference	
			Top Half	Bottom Half	Top Half	Bottom Half	Top Half	Bottom Half	Top Half	Bottom Half
2020	Recall/ Verbatim	Top Half	–	0.433	0.020	<0.001	<0.001	<0.001	<0.001	<0.001
		Bottom Half	0.433	–	0.001	0.107	0.005	<0.001	<0.001	<0.001
	Concept/ Inference	Top Half	0.020	0.001	–	0.027	0.502	0.205	0.024	0.003
		Bottom Half	<0.001	0.107	0.027	–	0.323	0.670	0.408	0.204
2022	Recall/ Verbatim	Top Half	<0.001	0.005	0.502	0.323	–	0.432	0.078	0.036
		Bottom Half	<0.001	<0.001	0.205	0.670	0.432	–	0.204	0.110
	Concept/ Inference	Top Half	<0.001	<0.001	0.024	0.408	0.078	0.204	–	0.827
		Bottom Half	<0.001	<0.001	0.003	0.204	0.036	0.110	0.827	–

Notes: Values >0.050 are in bold.

the recall/verbatim questions. In contrast, it varied considerably for the concept/inference questions, which were the primary determinant of the final exam score. All participants, therefore, demonstrated a similar ability to memorize and perform simple cognitive tasks and were differentiated based on a subset of more challenging questions.

Our second hypothesis was that retention of more fact-oriented knowledge, as evidenced by performance on recall/verbatim questions, would degrade more rapidly than the knowledge tested by concept/inference questions. Both question types showed similar proportional declines in performance from 2020 to 2022, but the decline in recall/verbatim performance was statistically significant, while the decline in concept/inference was nearly statistically significant. These results concerning our second hypothesis are, therefore, inconclusive.

We observed that student performance on both question types tended to converge over time, suggesting a regression toward core clinical knowledge but the reasons for this are unclear. One possibility is that the participants in this study were near the end of their clerkship year and, therefore, had spent significant time participating in patient care and studying for clinical subject examinations. The “floor” performance, the lowest observed performance in each group for each question type in 2022 (~60% for recall verbatim, ~43% for concept/inference), may, in part, represent knowledge that was relevant to their clinical learning and reinforced during the clerkships. Conversely, the questions that showed a substantial decline in performance may reflect knowledge that is obscure or of low clinical importance and was consequently not revisited. In other words, while the 2020 exam could stratify students mainly based on their performance on concept/inference questions, performance tends to converge toward a baseline of core clinical knowledge essential for practice.

If borne out by additional studies, this work has implications for exam design. The high performance seen on recall/verbatim questions and their lack of ability to discriminate between different examinees calls into question the utility of even including such questions on summative examinations. It is well recognized that assessment drives learning,⁴⁹ and testing relatively basic information with recall/verbatim questions could appropriately incentivize students to learn foundational knowledge. However, overreliance on them could disincentivize more significant mastery of the content by allowing those with superficial knowledge to pass. Rather than designing a summative exam with a high proportion of relatively easy recall/verbatim questions, it may be better to compose a shorter exam with a higher proportion of more difficult concept/inference questions and a commensurately lower passing threshold. In theory, such an approach could incentivize students to focus on higher-level cognitive processes that would presumably be of greater value than transient memorization.

Defining Question Types

The notion that MCQs can be categorized based on the level of cognitive function they assess has been introduced previously. In the 1970s and 1980s, several studies attempted to classify MCQs according to Bloom’s taxonomy. This approach is both tempting and problematic. Learning and memory are complicated phenomena, and their complexity is

not necessarily reflected in the six clean layers that comprise Bloom's hierarchy. Zaidi et al proposed that Bloom's taxonomy may have no meaningful utility in categorizing MCQs. Instead, they developed a dichotomized scheme of "higher order" and "lower order" questions.¹¹

The classification scheme developed for this study started with the very basic idea that answering a given MCQ involves (1) information and (2) cognitive processing. Information was divided into *familiar* or *esoteric* forms. Familiar information is integral to the learning of a given subject. An example from pulmonary physiology is compliance, a core concept underpinning an accurate understanding of how the lungs function. By contrast, examples of esoteric information (also from the lung) are the locations and nomenclature for the twenty individual lung segments. While this information is important to a subset of practitioners, it would be a time-consuming and low-yield endeavor for the average medical student to commit the segments to memory.

Cognitive processing was likewise subdivided into *simple* or *complex* forms. Simple processing encompasses functions that all medical students should be able to carry out with minimal cognitive load. Complex processing involves all cognitive functions that go beyond the simple. The classification of a given MCQ depends on the interplay between information and processing. Most will be classified based on processing type, with simple processing indicating a verbatim/recall question, and complex processing indicating a concept/inference question. An exception to this rule exists where recall of specific *esoteric information* is necessary to answer a question, regardless of the processing involved correctly. For example, consider a question where one must determine if a patient with an acid-base disorder has the expected compensatory response. Interpreting this question involves (among other things) understanding the relationship between PaCO_2 and HCO_3^- , but to answer it confidently one must know the precise quantitative aspect of this relationship. A "cognitive bottleneck" such as this one, which requires recall of information that is not intuitive and cannot be deduced, classifies such a question as recall/verbatim. It should be recognized that the lines dividing simple from complex and familiar from esoteric are sometimes not easily drawn or static. Complex processing may become simple with practice and acquisition of expertise, and esoteric information may become familiar to the point of being implicit.

These findings also underscore the importance of MCQ selection for educational research experiments that utilize them. For example, if the MCQs from this study were used to form a question bank for a prospective experiment, their known performance characteristics would show a variation in performance between 2020 and 2022, ranging from a 64-point decrease to a 17-point increase, representing significant differences in how students retained factual versus conceptual knowledge. An alternative approach to assessing critical reasoning in medical education could be relying on short-answer, free-response questions. This could also avoid question writing flaws that often appear in MCQs. However, MCQs have an established ability to address higher-order reasoning used in critical thought when their stems and answer choices have been carefully constructed.⁵⁰ The results of our present study should also aid in the design and writing of MCQ-based examinations that are aimed at assessing critical thinking, which will reflect higher-order cognitive function.

In designing assessments for educational research, it would be ideal to mitigate such potential distortions. MCQs should be chosen based on pre-defined learning objectives, categorized by cognitive level, and their respective performance characteristics determined in various testing environments. Such deliberation would ensure that the assessments used in educational research experiments are optimal regarding validity and difficulty. The dichotomous categorization of questions in this and other studies⁵¹ may be a more direct method to assess critical thinking vs factual recall. Even with carefully written or selected MCQs, perception of question difficulty varies widely for both students and faculty when based on the six levels of Bloom's taxonomy, blurring lower-order vs higher-order cognition performance.⁵²

Limitations

This study used questions from a limited-scope exam administered to a single cohort of just 56 medical student participants. This is a relatively small sample size. Further, the number of included concept/inference questions was modest at just 43 total, limiting the analysis's power. Participants had completed a substantial portion of the clerkship year by the time of the 2022 assessments, which is a confounding variable in our interpretation of the findings. Whereas the 2020 assessment was a summative examination, participation in the 2022 assessments was voluntary and had no stakes, which may have impacted the amount of effort expended on the questions. Finally, the mindset of participants is impossible to control and will always be

an uncontrolled variable in assessment studies for medical education or research. For example, just being sleepy can affect cognition and clinical reasoning.⁵³

Conclusion

The findings from our study provide critical insights into how different types of cognitive assessments impact learning outcomes in medical education. Our results show that conceptual knowledge, though initially more challenging to master, demonstrates greater durability over time than factual recall, particularly when reinforced through clinical practice. This observation raises important questions about the efficacy of traditional recall-based assessments, which may promote short-term performance but fail to support long-term knowledge retention. These results align with emerging research suggesting that recall-based methods are less effective than those that foster deeper conceptual understanding.^{54,55} Prior work demonstrates that implementing retrieval practice questions and spaced learning supports this approach and helps prevent knowledge decay.⁵⁶

Our findings advocate for reevaluating current medical assessment strategies, emphasizing the need to shift the focus from recall-driven evaluations to ones that promote critical thinking and conceptual integration. This aligns with the broader competency-based medical education (CBME) trend, prioritizing practical skills and attitudes alongside clinical knowledge.⁵⁷ However, defining and assessing competencies remains challenging.⁵⁸ Our findings suggest innovative approaches, such as cognitive diagnostic assessments, may offer valuable ways to measure factual recall and retention of more significant, clinically relevant knowledge.⁵⁹

Our results further highlight the need for frequent, longitudinal assessments, which provide formative feedback and support lifelong learning among physicians.⁶⁰ Ultimately, our research points toward integrating conceptual/inferential questions in medical education to better prepare students for long-term success in clinical practice and to encourage the retention of knowledge critical to patient care. Moving forward, medical educators should explore the mechanisms behind knowledge retention and the role of applied clinical learning to refine these assessment practices.

Abbreviations

MCQs, multiple choice questions; MCAT, Medical College Admissions Test; USMLE, the United States Medical Licensing Examination.

Data Sharing Statement

The data was obtained through internal quality improvement. This published article includes all data generated or analyzed during this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

The author(s) report no conflicts of interest in this work. A previous unpublished version of this work resides on the preprint server Research Square. Ethical approval was obtained from the University at Las Vegas Biomedical IRB on April 3, 2017 (Protocol 1030906-1).

References

1. Dudley HAF. Multiple-Choice Tests: Time For A Second Look ? *Lancet*. 1973;302(7822):195–196. doi:10.1016/S0140-6736(73)93021-3
2. Joorabchi B. How to...construct problem-solving Mcqs. *Med Teach*. 1981;3(1):9–13. doi:10.3109/01421598109081736
3. Joorabchi B, Chawhan AR. Multiple choice questions The debate goes on. *Medical Education*. 2009;9(4):275–280. doi:10.1111/j.1365-2923.1975.tb01938.x

4. Norcini JJ, Swanson DB, Grosso LJ, Shea JA, Webster GD. A comparison of knowledge, synthesis, and clinical judgment: multiple-choice questions in the assessment of physician competence. *Eval Health Prof.* 1984;7(4):485–499. doi:10.1177/016327878400700409
5. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education.* 1985;19(3):238–247. doi:10.1111/j.1365-2923.1985.tb01314.x
6. Ferland JJ, Dorval J, Levasseur L. Measuring higher cognitive levels by multiple choice questions: a myth? *Medical Education.* 1987;21(2):109–113. doi:10.1111/j.1365-2923.1987.tb00675.x
7. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DRA. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain.* DAvid McKay; 1956.
8. Kucewicz MT, Worrell GA, Axmacher N. Direct electrical brain stimulation of human memory: lessons learnt and future perspectives. *Brain.* 2023;146(6):2214–2226. doi:10.1093/brain/awac435
9. Airasian PW, Cruikshank KA, Mayer RE, Pintrich PR, Rath J, Wittrock MC. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Abridged Version.* 1st. Anderson L, Wittrock MC, eds.. Langman; 2001.
10. Huxham GJ, Naeraa N. Is bloom's taxonomy reflected in the response pattern to MCQ items? *Medical Education.* 1980;14(1):23–26. doi:10.1111/j.1365-2923.1980.tb02608.x
11. 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. doi:10.1097/ACM.0000000000002087
12. Alshamrani KM, Khan MA, Alyousif S. Assessment of radiological sciences students' and interns' long-term retention of theoretical and practical knowledge: a longitudinal panel study. *AMEP.* 2021;12:1549–1559. doi:10.2147/AMEP.S346802
13. Kerfoot BP, DeWolf WC, Masser BA, Church PA, Federman DD. Spaced education improves the retention of clinical knowledge by medical students: a randomised controlled trial. *Med Educ.* 2007;41(1):23–31. doi:10.1111/j.1365-2929.2006.02644.x
14. Yeh DD, Park YS. Improving learning efficiency of factual knowledge in medical education. *J Surg Educ.* 2015;72(5):882–889. doi:10.1016/j.jsurg.2015.03.012
15. Bell DS, Harless CE, Higa JK, et al. Knowledge retention after an online tutorial: a randomized educational experiment among resident physicians. *J GEN INTERN MED.* 2008;23(8):1164–1171. doi:10.1007/s11606-008-0604-2
16. Kolars JC, Gruppen LD, Traber PG, Paine ME, Davis WK, Woolliscroft JO. The effect of student- and teacher-centred small-group learning in medical school on knowledge acquisition, retention and application. *Med Teach.* 1997;19(1):53–57. doi:10.3109/01421599709019349
17. Trullàs JC, Blay C, Sarri E, Pujol R. Effectiveness of problem-based learning methodology in undergraduate medical education: a scoping review. *BMC Med Educ.* 2022;22(1):104. doi:10.1186/s12909-022-03154-8
18. Butler AC, Raley ND. The future of medical education: assessing the impact of interventions on long-term retention and clinical care. *J Graduate Med Educ.* 2015;7(3):483–485. doi:10.4300/JGME-D-15-00236.1
19. Bucklin BA, Asdigian NL, Hawkins JL, Klein U. Making it stick: use of active learning strategies in continuing medical education. *BMC Med Educ.* 2021;21(1):44. doi:10.1186/s12909-020-02447-0
20. Van Braak M, Veen M, Muris J, Van Den Berg P, Giroldi E. A professional knowledge base for collaborative reflection education: a qualitative description of teacher goals and strategies. *Perspect Med Educ.* 2021;11(1):53–59. doi:10.1007/S40037-021-00677-6
21. Kelly JM, Perseghin A, Dow AW, Trivedi SP, Rodman A, Berk J. Learning through listening: a scoping review of podcast use in medical education. *Acad Med.* 2022;97(7):1079–1085. doi:10.1097/ACM.0000000000004565
22. Simanton E, Hansen L. Long-term retention of information across the undergraduate medical school curriculum. *S D Med.* 2012;65(7):261–263.
23. Nelson A, Eliaz KL. Desirable difficulty: theory and application of intentionally challenging learning. *Medical Education.* 2023;57(2):123–130. doi:10.1111/medu.14916
24. Harrison A. Using knowledge decrement to compare medical students' long-term retention of self-study reading and lecture materials. *Assess Eval Higher Educ.* 1995;20(2):149–159. doi:10.1080/02602939508565717
25. Noya F, Carr S, Freeman K, Thompson S, Clifford R, Playford D. Strategies to facilitate improved recruitment, development, and retention of the rural and remote medical workforce: a scoping review. *Int J Health Policy Manag.* 2021;10(1):1. doi:10.34172/ijhpm.2021.160
26. Janse RJ, Van Wijk EV, Ruijter BN, et al. Comparison of very short answer questions and multiple choice questions in medical students: reliability, discrimination, acceptability and effect on knowledge retention. *medRxiv.* 2022. doi:10.1101/2022.07.13.22277583
27. Ramraje S. Comparison of the effect of post-instruction multiple-choice and short-answer tests on delayed retention learning. *AMJ.* 2011;4(6):332–339. doi:10.4066/AMJ.2011.727
28. Butler AC, Roediger HL. Testing improves long-term retention in a simulated classroom setting. *Eur J Cognit Psychol.* 2007;19(4–5):514–527. doi:10.1080/09541440701326097
29. Kang SHK, McDermott KB, Roediger HL. Test format and corrective feedback modify the effect of testing on long-term retention. *Eur J Cognit Psychol.* 2007;19(4–5):528–558. doi:10.1080/09541440601056620
30. Larsen DP, Butler AC, Roediger HL. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Medical Education.* 2009;43(12):1174–1181. doi:10.1111/j.1365-2923.2009.03518.x
31. Salam A, Yousuf R, Bakar SMA. Multiple choice questions in medical education: how to construct high quality questions. *Int J Hum Health Sci.* 2020;4(2):79. doi:10.31344/ijhhs.v4i2.180
32. DrK H, Es Dr F. Changing teaching strategies and lecture preparation to improve medical students' knowledge acquisition and retention. *Int J Adv Community Med.* 2022;5(1):50–54. doi:10.33545/comed.2022.v5.i1a.226
33. Silverberg J, Taylor-Vaisey A, Szalai JP, Tipping J. Lectures, interactive learning, and knowledge retention in continuing medical education. *J Contin Educ Health Prof.* 1995;15(4):231–234. doi:10.1002/chp.4750150407
34. Tractenberg RE, Gushta MM, Mulroney SE, Weissinger PA. Multiple choice questions can be designed or revised to challenge learners' critical thinking. *Adv in Health Sci Educ.* 2013;18(5):945–961. doi:10.1007/s10459-012-9434-4
35. Croskerry P, Campbell SG, Petrie DA. The challenge of cognitive science for medical diagnosis. *Cogn Res.* 2023;8(1):13. doi:10.1186/s41235-022-00460-z
36. Verenna AA, Noble KA, Pearson HE, Miller SM. Role of comprehension on performance at higher levels of Bloom's taxonomy: findings from assessments of healthcare professional students. *Anatomical Sciences Ed.* 2018;11(5):433–444. doi:10.1002/ase.1768
37. Tuckman BW, Harper BE. *Conducting Educational Research.* 6th ed. Rowman and Littlefield; 2012.

38. Karpicke JD, Blunt JR. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*. 2011;331(6018):772–775. doi:10.1126/science.1199327
39. Bibler Zaidi NL, Grob KL, Yang J, et al. Theory, process, and validation evidence for a staff-driven medical education exam quality improvement process. *Med Sci Educ*. 2016;26(3):331–336. doi:10.1007/s40670-016-0275-2
40. Mishra P, Singh U, Pandey C, Mishra P, Pandey G. Application of student's t-test, analysis of variance, and covariance. *Ann Card Anaesth*. 2019;22(4):407. doi:10.4103/aca.ACA_94_19
41. Yan F, Robert M, Li Y. Statistical methods and common problems in medical or biomedical science research. *Int J Physiol Pathophysiol Pharmacol*. 2017;9(5):157–163.
42. Bußenius L, Harendza S, Van Den Bussche H, Selch S. Final-year medical students' self-assessment of facets of competence for beginning residents. *BMC Med Educ*. 2022;22(1):82. doi:10.1186/s12909-021-03039-2
43. Hong S, Go B, Rho J, et al. Effects of a blended design of closed-book and open-book examinations on dental students' anxiety and performance. *BMC Med Educ*. 2023;23(1):25. doi:10.1186/s12909-023-04014-9
44. Meng J, Love R, Rude S, Martzen MR. Enhancing student learning by integrating anatomy in pathology teaching. *Med Sci Educ*. 2021;31(4):1283–1286. doi:10.1007/s40670-021-01330-x
45. Norris ME, Cachia MA, Johnson MI, Martin CM, Rogers KA. Are clerks proficient in the basic sciences? Assessment of third-year medical students' basic science knowledge prior to and at the completion of core clerkship rotations. *Med Sci Educ*. 2021;31(2):709–722. doi:10.1007/s40670-021-01249-3
46. Clar M. Homoscedasticity. In: Michalos AC editor. *Encyclopedia of Quality of Life and Well-Being Research*. Springer Netherlands; 2014:2910–2911. doi:10.1007/978-94-007-0753-5_1305
47. Wallisch C, Bach P, Hafermann L, et al. Review of guidance papers on regression modeling in statistical series of medical journals. *PLoS One*. 2022;17(1):e0262918. doi:10.1371/journal.pone.0262918
48. Yang K, Tu J, Chen T. Homoscedasticity: an overlooked critical assumption for linear regression. *Gen Psych*. 2019;32(5):e100148. doi:10.1136/gpsych-2019-100148
49. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? *Anatomical Sciences Ed*. 2009;2(5):199–204. doi:10.1002/ase.102
50. Pham H, Trigg M, Wu S, et al. Choosing medical assessments: does the multiple-choice question make the grade? *Educ Health*. 2018;31(2):65. doi:10.4103/efh.Efh_229_17
51. Davies DJ, McLean PF, Kemp PR, et al. Assessment of factual recall and higher-order cognitive domains in an open-book medical school examination. *Adv in Health Sci Educ*. 2022;27(1):147–165. doi:10.1007/s10459-021-10076-5
52. Monrad SU, Bibler Zaidi NL, Grob KL, et al. What faculty write versus what students see? Perspectives on multiple-choice questions using Bloom's taxonomy. *Med Teach*. 2021;43(5):575–582. doi:10.1080/0142159X.2021.1879376
53. Cleland J, Gates LJ, Waiter GD, Ho VB, Schuwirth L, Durning S. Even a little sleepiness influences neural activation and clinical reasoning in novices. *Health Sci Rep*. 2021;4(4):e406. doi:10.1002/hsr.2.406
54. Fraundorf SH, Caddick ZA, Nokes-Malach TJ, Rottman BM. Cognitive perspectives on maintaining physicians' medical expertise: IV. Best practices and open questions in using testing to enhance learning and retention. *Cogn Res*. 2023;8(1):53. doi:10.1186/s41235-023-00508-8
55. Holmboe ES, Osman NY, Murphy CM, Kogan JR. The urgency of now: rethinking and improving assessment practices in medical education programs. *Acad Med*. 2023;98(8S):S37–S49. doi:10.1097/ACM.0000000000005251
56. Donker SCM, Vorstenbosch MATM, Gerhardus MJT, Thijssen DHJ. Retrieval practice and spaced learning: preventing loss of knowledge in Dutch medical sciences students in an ecologically valid setting. *BMC Med Educ*. 2022;22(1):65. doi:10.1186/s12909-021-03075-y
57. Ng IKS, Mok SF, Teo D. Competency in medical training: current concepts, assessment modalities, and practical challenges. *Postgraduate Med J*. 2024;qgae023. doi:10.1093/postmj/qgae023
58. Bhanji F, Naik V, Skoll A, et al. Competence by design: the role of high-stakes examinations in a competence based medical education system. *Perspectives Med Educ*. 2024;13(1):68–74. doi:10.5334/pme.965
59. Matus AR, Matus LN, Hiltz A, et al. Development of an assessment technique for basic science retention using the NBME subject exam data. *BMC Med Educ*. 2022;22(1):771. doi:10.1186/s12909-022-03842-5
60. Rottman BM, Caddick ZA, Nokes-Malach TJ, Fraundorf SH. Cognitive perspectives on maintaining physicians' medical expertise: i. Reimagining maintenance of certification to promote lifelong learning. *Cogn Res*. 2023;8(1):46. doi:10.1186/s41235-023-00496-9

Advances in Medical Education and Practice

Dovepress

Publish your work in this journal

Advances in Medical Education and Practice is an international, peer-reviewed, open access journal that aims to present and publish research on Medical Education covering medical, dental, nursing and allied health care professional education. The journal covers undergraduate education, postgraduate training and continuing medical education including emerging trends and innovative models linking education, research, and health care services. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <http://www.dovepress.com/advances-in-medical-education-and-practice-journal>