The Quality of In-house Medical School Examinations

Article in Academic Medicine • February 2002
DOI: 10.1097/00001888-200202000-00016 • Source: PubMed

CITATIONS: 73
READS: 119

6 authors, including:

Robert M Galbraith
143 PUBLICATIONS 3,750 CITATIONS

David B Swanson
American Board of Medical Specialties
152 PUBLICATIONS 3,898 CITATIONS

Available from: David B Swanson
Retrieved on: 23 September 2016
The Quality of In-house Medical School Examinations

Ralph F. Jozefowicz, MD, Bruce M. Koeppen, MD, PhD, Susan Case, PhD, Robert Galbraith, MD, David Swanson, PhD, and Robert H. Glew, PhD

ABSTRACT

Purpose. Most medical schools test their students throughout the curriculum using in-house examinations written by the faculty who teach the courses. The authors assessed the quality of in-house examinations used in three U.S. medical schools.

Method. In 1998, nine basic science examinations from the three schools were gathered and each question was subjected to quality assessment by three expert biomedical test developers, each of whom had had extensive experience in reviewing and evaluating questions for the United States Medical Licensing Examination (USMLE) Steps 1 and 2. Questions were rated on a five-point scale: 1 = tested recall only and was technically flawed to 5 = used a clinical or laboratory vignette, required reasoning to answer, and was free of technical flaws. Each rater made independent assessments, and the mean score for each question was calculated. Mean quality scores for National Board of Medical Examiners (NBME) who were trained question writers were compared with the mean scores for question writers without NBME training. The raters' quality assessments were made without knowledge of the test writers' training background or the study's hypothesis.

Results. A total of 555 questions were analyzed. The mean score for all questions was 2.39 ± 1.21. The 92 questions written by NBME-trained question writers had a mean score of 4.24 ± 0.85, and the 463 questions written by faculty without formal NBME training had a mean score of 2.03 ± 0.90 (p < .01).

Conclusions. The in-house examinations were of relatively low quality. The quality of examination questions can be significantly improved by providing question writers with formal training.


Most medical students will tell you that few aspects of their undergraduate medical education are as stressful to them as are the many periodic examinations that constitute the gauntlet they must run from entry into medical school through the time just before they receive their diplomas and recite the Hippocratic oath. These examinations not only determine whether a student passes a unit, course, or clerkship, at many medical schools they also determine class rank and influence the quality and strength of the dean's letter that goes out to directors of residency programs.

Usually two kinds of tests are used to assess medical students' knowledge and competency. The majority of these examinations are in-house test papers of the multiple-choice variety written by the faculty who teach the courses; the remainder are extramural, national examinations prepared by discipline-based societies or licensing organizations. While much has been written about the advantages and disadvantages of multiple-choice questions, at the moment and probably for some time to come, this is the kind of questions medical students will face on both in-house and national licensing examinations.

Commonly, every four to six weeks throughout the four years of medical school, a student will sit for a high-stakes one- or two-hour examination composed of 50 or more questions. Paradoxically, all too often these critical examinations are written and assembled at the last minute. Furthermore, by the time the questions have been assembled from the half dozen or more instructors who taught the course, there is inade-
quate opportunity to review the submitted questions or evaluate the examination for balance and overall quality. In addition, while most faculty who lecture medical students would argue that they write reasonably good exam questions, the truth is that few of them have benefited from formal training in the art of question writing. Often, the result is a heterogeneous collection of single-best-answer, fill-in-the-blank, and true-or-false questions of highly variable quality. Thus, while many faculty take seriously their formal lecturing duties and invest considerable thought into the preparation of lectures, it has been our experience that few commit commensurate effort to the task of preparing examination questions on the material they teach. The end result is that many excellent teachers assess their students using examinations of questionable quality. This problem is exacerbated by the fact that instructors often teamwork the particular section of the curriculum upon which the medical students are examined.

The literature contains few studies that evaluate the quality of in-house basic science examinations at medical schools. Also lacking are studies of the effectiveness of development programs aimed at improving the quality of multiple-choice examinations at medical schools. Our central hypothesis contends that in-house examinations currently being administered to U.S. medical students are of variable and generally low quality. To test this hypothesis, we gathered nine examinations from three U.S. medical schools and subjected the 555 questions to (blinded) quality assessment by three experts in educational measurement in medicine, each of whom had had extensive experience in item review for Steps 1 and 2 of the United States Medical Licensing Examination (USMLE) and the National Board of Medical Examiners’ (NBME’s) Subject Examinations (shelf examinations) program. This paper reports their ratings of the quality of the items on these nine in-house examinations.

**METHOD**

In 1998, nine basic science examinations from three different medical schools were gathered, for a total of 555 questions. Of these, 92 had been written by individuals who had been trained to write items for USMLE Step examinations.

The quality of each question was rated by three expert biomedical test developers with extensive experience reviewing items for USMLE Steps 1 and 2. Each rated the items independently, and they were not informed of the purpose of the study. In addition, they did not have knowledge of the medical schools where the questions had originated or the item writers’ identities or exam-writing training.

The rating scheme was developed to reflect accepted item-writing principles, as well as the results of empirical studies of factors influencing item quality. Items were rated on a five-point scale. A score of 5 was awarded to any item that included a vignette (either patient or laboratory), a one-best-answer format (i.e., not true-or-false), and no item flaw. A score of 4 was given to an item that failed to completely satisfy one of these three conditions. The rating scheme required vignette items to be assessed lower if they had significant item flaws, or if the vignette was not necessary to answer the question. If an item did not involve a vignette, it was awarded a maximum of 3; to get this value, the item had to be focused, contain no item flaw, and involve more than recall of an isolated fact. An item was rated 2 if it failed to satisfy one of the criteria for a 3. An item was rated 1 if it failed to satisfy two or more of the criteria for a 3. Because true-or-false items generally had item flaws, they were awarded low ratings (i.e., 1 or 2).

The average quality assessment score (QAS)—the mean of the three assessors’ ratings—was then calculated for all test items, and for subsets of items based on school of origin and the authors’ training in question writing. Examples of the items (quoted and unedited) and the QAS scores they were assigned are shown in the Appendix.

To evaluate inter-rater agreement, a random-effects items-by-raters analysis of variance was performed to obtain variance components and a generalizability coefficient was calculated, with differences in rater stringency included in measurement error.

**RESULTS**

Variance components for items, raters, and error were 1.3174, 0.0641, and 0.4287, respectively. Pearson correlation coefficients between pairs of raters were .70, .78, and .79. The generalizability of the QAS for each item was .89, indicating a high level of inter-rater agreement for subsequent analyses.

As shown in Table 1, the mean QAS for all 555 questions was 2.39 ± 1.21. The qualities of the examinations at the three schools varied: the 222 questions from School A had a mean QAS of 1.94 ± 0.90, the 180 questions from School B had a mean QAS of 3.26 ± 1.28, and the 153 questions from School C had a mean QAS of 2.03 ± 0.94. The significantly higher mean QAS for School B likely reflects the fact that 44% of the questions submitted by School B had been written by an NBME-trained item writer.

When QAS scores for the 555 items were analyzed in relation to whether items’ writers had been trained by the NBME, the 92 questions written by writers who had been trained had a mean QAS of 4.24 ± 0.85, compared with a mean of 2.03 ± 0.90 for the 463 questions drafted by faculty without NBME training. The difference between these two means was highly significant (p < .001).

The overall distribution of QAS for
formats for questions, or may not have developed question-writing guidelines to assist in this process. Finally, examinations are usually assembled at the last minute, which constrains the faculty charged with assembling the examination and precludes developing and using strategies that could improve the overall quality of the examinations.

A limitation of our study is that the 555 questions were assessed by raters who were NBME staff, and the item writers whose questions received high QAS ratings were ones trained by these staff. Although the NBME-staff raters were blinded to the sources of the items, a reasonable argument could be made that we have simply shown that the raters recognized items prepared in accord with NBME training. One way to address this criticism would be to have the same pool of examination questions evaluated by raters who were not NBME staff or individuals who have not been trained by the NBME. For example, they could be assessed by faculty at the three medical schools (or other U.S. medical schools) who are highly regarded for the quality of the examinations they write, but who were not trained by NBME staff. While the results of our study do suggest that providing training to examination writers improves the quality of the multiple-choice questions they write, our study did not specifically assess the effects such a training program might have on the test development skills of a particular group of exam writers. We acknowledge that faculty who have not been NBME-trained may produce multiple-choice questions of high quality and that other kinds of training may be beneficial in this regard.

Another pertinent question raised by this study is whether the quality of the examination questions matters. That is, do the best students do well regardless of the quality of the examination's questions? We will address this question in a future study.

Since modes of instruction vary across medical schools, with some institutions emphasizing traditional basic science disciplines in the first two years and others integrating the teaching of basic science and clinical material from the outset of the curriculum, some medical educators might argue that vignette-type questions are inappropriate for assessing students' performances in courses presented in the first one or two semesters and that there is a place for recall- or identification-type questions in the first few semesters. While we concede that it is possible to write multiple-choice questions that perform well from the standpoint of discriminating students' performances in a course or bloc of lectures, we nevertheless maintain that testing knowledge through contextual, vignette, or problem-solving questions that require reasoning skills is preferable to testing the recall of isolated facts. We therefore acknowledge that if in the present study we had not used a rating system that downgraded questions that lacked a vignette, the results of the assessments of the questions might well have turned out differently. Furthermore, we concede that in some settings, particularly where the medical curriculum is one in which core basic science subject matter is taught in the first two years, it may be appropriate to write multiple-choice questions that consist of a mix of vignette and recall-type items.

What can be done to improve the quality of "in-house" medical school examinations? First, faculty responsible for writing examinations should be trained; writing good examination questions is a skill that can be learned. Second, at the outset of the examination writing process, all participants should agree on the format and guidelines for writing items. For example, those types of items that involve "not" and "except" formats should be disallowed. Third, the exam-
ination should be prepared well in advance (e.g., two to three weeks) of the date the students will sit for the examination. Fourth, a committee should review, critique, and approve the content and format of the final draft of the examination.

REFERENCES


APPENDIX

Examples of In-house Examination Questions (Quoted and Unedited) from Three U.S. Medical Schools

Questions 1–3 are examples of questions that received QAS scores of 4.0–5.0; questions 4–6 are examples of questions that scored in the intermediate range (QAS 3.0–4.0), and questions 7–9 are examples of questions that scored poorly (QAS 1.0–2.0). The mean QAS score for each question is indicated in the parentheses and the correct answer is marked by an asterisk.

1. (QAS = 4.3)

A cell has channels for Na⁺, K⁺, and Cl⁻ in its plasma membrane. The resting membrane potential is −60 mV (cell interior negative). The intracellular and extracellular concentrations for these ions are (note—the values of E, reflect the convention of placing the extracellular ion concentration in the numerator of the Nernst equation):

\[
\begin{align*}
\text{Intracellular} & \quad \text{Extracellular} \\
Na^+ & \quad 14 \text{ mmol/L} \quad 140 \text{ mmol/L} \\
K^+ & \quad 150 \text{ mmol/L} \quad 5 \text{ mmol/L} \\
Cl^- & \quad 10 \text{ mmol/L} \quad 100 \text{ mmol/L} \\
\end{align*}
\]

A. Net Cl⁻ movement out of the cell will be increased
B. Net Cl⁻ movement into the cell will be increased
C. There will be no change in the net movement of Cl⁻

2. (QAS = 4.7)

A patient is being mechanically ventilated (on a respirator). His arterial blood gas and serum electrolyte analysis shows:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P_O₂</td>
<td>150 mm Hg</td>
<td>P_CO₂</td>
</tr>
<tr>
<td>pH</td>
<td>7.52</td>
<td>Na⁺</td>
</tr>
<tr>
<td>K⁺</td>
<td>4.5 mEq/L</td>
<td>Cl⁻</td>
</tr>
<tr>
<td>HCO₃⁻</td>
<td>20 mEq/L</td>
<td></td>
</tr>
</tbody>
</table>

The acid–base abnormality is:

A. Acute (uncompensated) respiratory acidosis
B. Acute (uncompensated) respiratory alkalosis
C. Acute (uncompensated) metabolic acidosis
D. Acute (uncompensated) metabolic alkalosis

3. (QAS = 5.0)

A middle-aged male complains of difficulty climbing stairs. He describes weakness without pain in his right lower limb. He is able to place his right leg on each step without experiencing any problem, but has difficulty climbing the step, and must grasp the handrail to pull himself up. Climbing the next step with his left leg occurs normally. You also notice that his gait on a flat surface appears nearly normal, there is no weakness in extending the right knee against a considerable load. You suspect damage and/or malfunction in the:

A. Obturator nerve
B. Tibial nerve
C. Superior gluteal nerve
D. Femoral nerve
E. Inferior gluteal nerve

4. (QAS = 3.0)

The concentration of glucose in the plasma is 100 mg/100 ml. GFR is 120 ml/min. How much glucose is filtered each minute?

1. 100 mg
2. 1200 mg

160
*3. 120 mg
4. 1.2 mg
5. 1 mg

5. (QAS = 3.7)

A patient presents with megaloblastic anemia. Serum vitamin B₁₂ levels are below normal although a dietary assessment indicates the dietary intake of vitamin B₁₂ is adequate. The Schilling test indicates normal production of intrinsic factor. Which one of the following conditions could contribute to vitamin B₁₂ deficiency in this patient?

A. Excess acid production in the stomach
B. Consumption of a high-fiber diet
C. Excess intake of folic acid
*D. Pancreatic insufficiency
E. Excess excretion of vitamin B₁₂ in the urine

6. (QAS = 3.0)

An 18-year-old girl awakens at 7:50 A.M. for her 8:00 A.M. Monday class, dresses quickly, and starts running across the quadrangle to the Biology Building. She feels shaky and the next thing she knows she is lying on the path with two students standing over her telling her that an ambulance is on the way. She is brought to the emergency department but no students accompany her to describe what they saw. Past history is negative and there is no family history of seizures. Which of the following questions will LEAST likely help you determine what happened to her?

1. Does she experience jerkiness when she awakens in the morning?
2. Did she use marijuana the night before?
3. How much sleep did she get over the past weekend?
4. Was she drinking alcohol the night before?
5. Has she been having recent headaches?

7. (QAS = 1.0)

The glycolytic conversion of glucose to lactate: (single best answer).

A. Generates a net gain of two NADH's for each glucose consumed
B. Requires the direct participation of molecular oxygen
C. Cannot take place in cells lacking mitochondria

*D. Is stimulated by a high intracellular concentration of fructose-2,6-bis-phosphate
E. Involves a single dehydrogenase

8. (QAS = 1.3)

Louder sound causes:

A. Increased amplitude of action potential
B. Increased frequency of action potential
C. Increased receptor potential amplitude
D. Recruitment of different hair cells
*E. All of the above

9. (QAS = 1.0)

In most cases of sensorineural (inner ear damage) hearing loss:

A. Hearing improves with time
B. Surgery can correct the loss
C. Removal of the cochlea is recommended
D. High-frequency hearing is lost before low-frequency hearing
*E. Amplifying incoming sounds will correct all perceptual problems