HOW WE TEACH | Generalizable Education Research

Medical students’ vs. family physicians’ assessment of practical and logical values of pathophysiology multiple-choice questions

Damir Secic,1 Dzenana Husremovic,2 Eldan Kapur,3 Zaim Jatie,4,5 Nina Hadziahmetovic,2 Benjamin Vojnikovic,6 Almir Fajkic,1 Amir Meholic,7 Lejla Bradic,7 and Amila Hadzic7

1Department of Pathophysiology, Medical Faculty University of Sarajevo, Sarajevo, Bosnia and Herzegovina; 2Department of Psychology, Faculty of Philosophy, University of Sarajevo, Sarajevo, Bosnia and Herzegovina; 3Department of Anatomy, Medical Faculty University of Sarajevo, Sarajevo, Bosnia and Herzegovina; 4Department of Family Medicine, Medical Faculty University of Sarajevo, Sarajevo, Bosnia and Herzegovina; 5Public Institution Health Centre of Sarajevo Canton, Sarajevo, Bosnia and Herzegovina; 6Dean’s Office, Medical Faculty University of Sarajevo, Sarajevo, Bosnia and Herzegovina; and 7BoHeMSA, Bosnian and Herzegovinan Medical Students’ Association, Sarajevo, Bosnia and Herzegovina

Submitted 23 September 2015; accepted in final form 22 November 2016

Medical students vs. family physicians’ assessment of practical and logical values of pathophysiology multiple-choice questions. Adv Physiol Educ 41: 62–68, 2017; doi:10.1152/advan.00145.2015.—Testing strategies can either have a very positive or negative effect on the learning process. The aim of this study was to examine the degree of consistency in evaluating the practicality and logic of questions from a medical school pathophysiology test, between students and family medicine doctors. The study engaged 77 family medicine doctors and 51 students. Ten questions were taken from cardiac pathophysiology and 10 questions from pulmonary pathophysiology, and each question was assessed on the criteria of practicality and logic. A nonparametric Mann-Whitney test was used to test the difference between evaluators. On the criteria of logic, only four out of 20 items were evaluated differently by students in comparison to doctors, two items each from the fields of cardiology and pulmonology. On the criteria of practicality, for six of the 20 items there were statistically significant differences between the students and doctors, with three items each from cardiology and pulmonology. Based on these indicative results, students should be involved in the qualitative assessment of exam questions, which should be performed regularly under a strictly regulated process.

students’ assessment; M.D.s’ assessment; multiple-choice questions; pathophysiology

“Non scholae sed vitae.”
“We do not learn for school, but for life.”
—Latin saying

A high-quality education includes evaluation through testing of acquired knowledge. Rowntree (31) and Ramsden (29) have emphasized in their papers how powerful the impact of testing strategies on the learning process can be, as testing strategies positively or negatively influence the learning process itself. Therefore, assessment is crucial and has a very strong effect on the approach to learning and learning outcomes (24, 28). User feedback is considered to be one of the most important methods of evaluation of any activity, and in terms of learning feedback from students it is considered to be important and even crucial to the evaluation process (27).

If the examination process does not comply with the objectives of teaching and the content of the future work the students will do after graduation, we can say that the test completely loses its meaning. This is explained in the concept of constructive alignment by Biggs (1) and Biggs and Tang (2). Constructive alignment is deeply established in the constructivist theory of teaching. This concept emphasizes that quality and effective teaching involves reconciliation between the expected learning outcomes (intended learning outcomes), teaching and learning activities, and assessment tasks. According to this concept, teachers need to clearly define the goals of their teaching and evaluation criteria of learning to help students align their learning with the professor requirements. According to Biggs and Tang (2), constructively aligned courses encourage students to achieve a higher cognitive level, such as the application (convenience of acquired knowledge in life), connection (a logical link between learning content), and evaluation in a real situation. Although the concept of constructive alignment was originally intended for the preuniversity level and work in the classroom (1), quality assurance agencies in many countries have begun to follow the principle of systematic thinking as a framework for the evaluation of programs in higher education (37). This principle is beginning to be recognized as important at the University of Sarajevo to align higher education with the European established standards. While insisting on the reform of the curriculum in terms of preparation of clear outcomes and the use of interactive teaching methods, a significant change that is expected of the professors is in the evaluation and assessment of student achievements.

According to Leach et al. (20) and McAllister et al. (26), the involvement of students is necessary to encourage them to influence their learning through involvement in the design of evaluations. Grades are negotiated in close personal and working relationships between students and assessors, where the evaluator is the student’s permanent clinical educator. Govarts et al. (11) believe that the trust and acceptance of the system of assessment of the evaluator and those who are evaluated are crucial for assessment. As noted by both Boud...
(4) and McAllister et al. (26), the absence of student engagement can impact both learning associated with the assessment and the validity and reliability of the evaluation. To maximize the positive potential of testing of knowledge, it is necessary to constantly bear in mind the strategic goals (35). Since long-term knowledge is based on understanding, it is essential that these questions are not looking for simple reproduction or recall but are assessing knowledge at higher levels (10, 16, 17, 34). Questions should be designed so that the student uses logic or, according to Bloom’s taxonomy, comprehension, analysis, synthesis, evaluation, and creation (3). The strategic objectives should focus on the relevance of the acquired knowledge, as utilization of knowledge in practice is the basis of training future professionals. However, authors such as Jozefovicz et al. (17) and Vanderbilt et al. (34) point out that teachers, due to the lack of formal training, often engage in inadequate monitoring of the quality of their work in the area of knowledge testing. They tend to prepare tests that do not meet the expected criteria. However, it is indeed very important for faculty to be able to create exams that can accurately measure student competency with an acceptable score distribution (18). And yet, there are not very many studies about how to improve the quality of multiple-choice question (MCQ) exams at medical schools, resulting in the fact that in-house examinations are variable and generally of poor quality (17).

Test practicality refers to whether the content of the question measures practical use of learned concepts and is directly linked with the ability to use such knowledge in future practice. Logic applies to how the content is presented so that the students can fully understand the whole issue. It is important to note that the information should be relevant to prevent loss of knowledge (16).

The Bologna system of study has brought about a number of significant changes in teaching practice, particularly in the area of evaluation and assessment of students’ achievements. Teachers were introduced to the idea that in the evaluation of knowledge, subjective factors need to be reduced. As a result, the following tests were far more objective and impartial than their preceding counterparts. In the evaluation of our medical students, the most commonly used standard is the MCQ. Designing these questions is very challenging and difficult due to the fact that this format of questioning may be a limiting factor if the teacher does not have enough time or expertise to prepare high-quality, appropriate questions. In the preparation of tests of knowledge, teachers do not have a lot of support even in the textbooks used for teaching. We have observed that medical textbooks often present theoretical information that is unrelated to the practical context as well as information representing specialist and subspecialist levels of knowledge, which renders such information irrelevant to the initial and future work of general practitioners.

According to Jozefovicz et al. (17) most of the in-house medical exams are written and composed at the last minute. Additionally, the author claims that many excellent teachers evaluate their students through the use of exams of questionable quality. This problem is enlarged because the faculty teach and examine only a part of the curriculum, but it is clear that teachers with knowledge evaluation training define better questions.

The central hypothesis of our study is that there are no significant differences in assessing the practicality and logic of the test questions from the pathophysiology curriculum when comparing medical students and family medicine doctors. The aim of this study was to determine the degree of agreement between the evaluation by experts, in our case family medicine doctors and the students as assessors at the undergraduate level, and to examine how the questions are practical and logical in relation to the evaluation of knowledge of medical students in the subject of pathophysiology, more specifically, in the fields of pulmonology and cardiology.

**METHODS**

**Educational context.** All items in the knowledge tests for cardiology and pulmonology were designed as Pick-N-type MCQs with five potential answers. The grading was binary, meaning that the answer was marked as “correct” if the students picked all required answer options. According to Lord (22), to maximize the discrimination item function, the required level of difficulty should be adjusted for the probability of guessing. Since the design of the questions in these particular tests was such that the probability of guessing correctly was 1.6%, there was no need to adjust for guessing in determining the optimal difficulty of questions.

The main objective of the undergraduate part of the medical schools in our country is that medical students as well as family/general medicine doctors are trained to work in primary care. After passing the state examination, they are licensed to work independently as family/general physicians in primary care. Accordingly, questions from the tests mentioned above are designed to examine their knowledge in the field of family/general medicine rather than as a specialist or subspecialist.

Who is supposed to intelligently assess the quality of test questions in certain fields? Qualitative analysis of questions is usually done by staff that are experts in the field and experienced in testing. In our case, because it is a test of an undergraduate course, the evaluation was carried out by family medicine doctors, most of whom are family medicine specialists. The main reason for choosing them is because the medical students’ work, upon completion of 6 yr of study, will be closest to the work of family medicine doctors. Therefore, we believe that family medicine doctors can assess the practicality of questions fairly well because of their experience in daily practice. Also, according to Harris et al. (16), the postgraduate program of family medicine covers the widest range of medical disciplines. If the questions regarding cardiology were to be checked by internal medicine subspecialists (cardiologists) or internal medicine specialists, the questions would have been observed from the perspective of specialists, not family doctors.

**Participants.** Our research was designed as a cross-sectional study and was conducted at the Public Institution Health Center of Sarajevo Canton and Medical Faculty University of Sarajevo during the 2013–2014 academic year. The research included two subsamples: doctors and students.

The study engaged 79 of 204 family doctors from the Health Center of Sarajevo Canton. These family medicine doctors work daily in primary care, covering all medical fields. Their experience qualifies them as the most appropriate evaluation experts for the practicality and logic of test questions in undergraduate course exams. Prior to the start of the study, the purpose and goals of the research were explained to all of the family medicine physicians, after which some doctors volunteered to participate.

Fourth-year students during the 2013–2014 academic school year that had taken a pathophysiology exam during the prior academic year of 2012–2013 as third-year students were invited to participate in the study. The objective of the research was explained. Students then volunteered for the study, and a survey of 52 of the 78 medical students was conducted.
Participation of family medicine doctors and medical students was approved by the Governing Office of the Health Center of Sarajevo Canton and the Review Board of the Bosnian and Herzegovinian Medical Student Association.

Data collection. Of the total number of completed questionnaires, we eliminated only three surveys, two samples done by doctors and one sample done by a student due to failure to follow instructions and irregular data. Therefore, the usable number of questionnaires was 77 filled out by family medicine doctors and 51 completed by students. Certain questions remained unanswered by both doctors and students alike.

On the questionnaire there was space provided for doctors and students to write their names, but it was not obligatory. Most of the family medicine doctors, 56 of 77 (72.7%), wrote their names, seven (9.1%) wrote only their initials, and 14 (18.2%) doctors did not write their names because they wanted to remain anonymous. Of the 56 family medicine doctors with their name written down, there were 46 (82.1%) women and 10 (17.9%) men. Almost the same proportion of students (22 of 51) wrote their initials (43.1%) and their full name (21 of 51; 41.2%), and only eight (15.7%) of the students wanted to remain anonymous. Of the 21 students with a full name written, there were five men (23.8%) and 16 women (76.2%). As to the large number of women in the study, this is a correct reflection of both of the population groups. Of 204 family medicine doctors in the Health Centre of Sarajevo Canton, 173 are women (84.8%), and 59 of 78 (75.6%) medical students are women.

All doctors and students independently evaluated the given questions, and they were also assigned their own identification number that was used during the study. Students and doctors evaluated a Pick-N-type MCQ type exam marked with the correct answers, with instructions to evaluate all of the questions as accurately and conscientiously, to the best of their ability.

In the study, we prepared an instrument that consisted of two parts: a group of pulmonology questions and a group of questions based on cardiology. Beside each question, participants had to note to what extent were the questions logical and practical. We used a scale from 1 to 5, where 1 meant “not at all” and 5 meant “completely”. Practicality was defined as whether the question contained practical importance for the future of medical doctors who work in general practice/family medicine, meaning in primary health care. Logical was defined as whether the questions were understandable and without ambiguity.

The full test of the special pathophysiology course had 70 questions based on the specified organ systems. We took all of the questions on cardiology (10 questions) and on pulmonology (10 questions), regardless of their difficulty index. We decided not to put a larger number of questions in the questionnaire, because there could have been a decline in the assessment quality by family medicine doctors and students. The two criteria of practicality and logic were used for both cardiology and pulmonology questions.

Data analysis. In addition to analysis based on the main survey, item analysis was conducted based on the results of tests of knowledge of 51 students who took the test (the one we used) and who participated in the assessment of practicality and logic. We established difficulty indices for each question and the linkage with the overall dynamic figure of the test using the point biserial correlation coefficient.

As for data processing and analysis, SPSS version 21 (SPSS, Chicago, IL) was used. For the testing of distribution normality, the Kolmogorov-Smirnov (K-S) test was used. Finally, the Mann-Whitney test was used for the test of statistical significance of differences in grades. The level of statistical significance was set at $P \leq 0.05$. Item analysis was performed by calculating the difficulty index for each question and item/total correlation by using the point biserial coefficient.

## RESULTS

### Item analysis.

As a first step in the item analysis the test statistic was performed, and the results are given in Table 1.

According to the average item means for both tests and the average item/total correlations, the test was moderately difficult to easy, and the items were moderately correlated with the total score on the test. Table 2 shows the item statistics. As seen in the table, no questions were identified as very difficult ($P < 0.30$). On the cardiology test the easiest question was question C5, and the most difficult was question C8, which, unexpectedly, also had the lowest correlation with the total score. On the pulmonology test the easiest question was P1, and the most difficult was P6. When question difficulties and correlations were cross-tabulated in a p-r diagram it was noticeable that there were no questions with low correlations ($r < 0.1$) and no difficult ones ($P < 0.30$). According to the item statistics, all questions correspond with the general achievement, and correlations are higher for moderately difficult items, which is expected since they have the largest variance.

Based on the average value of the results of the individual questions, we can conclude that the questions in cardiology and pulmonology were rated relatively high on the criteria of logic and practicality (Table 3). This indicates that the design of the test was taken into account, the questions were meaningful, and the questions were incorporated in a practical context to be able to measure higher levels of knowledge, enabling the under-

<table>
<thead>
<tr>
<th>Question</th>
<th>$P$ Value</th>
<th>$r_{bt}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.80</td>
<td>0.36</td>
</tr>
<tr>
<td>C2</td>
<td>0.78</td>
<td>0.53</td>
</tr>
<tr>
<td>C3</td>
<td>0.62</td>
<td>0.44</td>
</tr>
<tr>
<td>C4</td>
<td>0.84</td>
<td>0.51</td>
</tr>
<tr>
<td>C5</td>
<td>0.88</td>
<td>0.54</td>
</tr>
<tr>
<td>C6</td>
<td>0.74</td>
<td>0.46</td>
</tr>
<tr>
<td>C7</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>C8</td>
<td>0.44</td>
<td>0.16</td>
</tr>
<tr>
<td>C9</td>
<td>0.54</td>
<td>0.60</td>
</tr>
<tr>
<td>C10</td>
<td>0.86</td>
<td>0.20</td>
</tr>
<tr>
<td>P1</td>
<td>0.94</td>
<td>0.13</td>
</tr>
<tr>
<td>P2</td>
<td>0.57</td>
<td>0.37</td>
</tr>
<tr>
<td>P3</td>
<td>0.59</td>
<td>0.28</td>
</tr>
<tr>
<td>P4</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>P5</td>
<td>0.92</td>
<td>0.32</td>
</tr>
<tr>
<td>P6</td>
<td>0.49</td>
<td>0.31</td>
</tr>
<tr>
<td>P7</td>
<td>0.69</td>
<td>0.49</td>
</tr>
<tr>
<td>P8</td>
<td>0.71</td>
<td>0.50</td>
</tr>
<tr>
<td>P9</td>
<td>0.82</td>
<td>0.59</td>
</tr>
<tr>
<td>P10</td>
<td>0.61</td>
<td>0.42</td>
</tr>
</tbody>
</table>

C1–C10, cardiology questions; P1–P10, pulmonology questions.
standing of the problem in a practical situation and implicitly increasing the use of acquired knowledge.

The K-S test confirmed the distribution of the results of all questions, with overall scales included, and differed substantially from the normal distribution. The values of the K-S test are within a range from 2.18 to 4.90 (P < 0.00) for the individual questions. The average scores on individual questions ranged from 3.1 to 4.7 on a scale of 1 to 5. Because the distribution of results on all questions deviated significantly from normal, to analyze the differences between the estimates of students and doctors, we used nonparametric statistics. The Mann-Whitney U-test was used to assess the significance of differences of two independent samples.

A total of four items were evaluated differently by students and doctors on the criterion of “logic.” Of these, two items are in the cardiology group (items 3 and 7), and two are in the pulmonology group (items 3 and 5). Two items were evaluated lower by doctors, and two items were evaluated lower by the students. For six items there were statistically significant differences in the estimates between students and doctors on the criterion of “practicality.” Of these six items, three items were in cardiology (items 6, 14, and 16), and three items were in pulmonology (items 4, 10, and 20). Three items were evaluated lower by doctors (item 14 in cardiology and items 4 and 10 in pulmonology), and three items were evaluated lower by the students (items 6 and 16 in cardiology and item 20 in pulmonology).

Once the results of the tested differences were obtained, a group of three experts (professors from the Medical Faculty University of Sarajevo and general medicine doctors) did a qualitative analysis of the questions, from which significant differences were found.

The main reasons that the experts identified to explain these differences in assessment are related to the structure and technique of implementation of the teaching process. The questions students have evaluated as more logical are the ones that can be found in textbooks. These questions are structured in a way so that they reflect book knowledge, whereas experienced doctors recognized them as less logical. Another reason the students evaluated some questions with a lower grade is the fact that the syllabus is not fully compliant in terms of the connectivity of the syllabus of each subject and their position in the individual semesters. For example, within the subject of pathophysiology, students work with topics that include elements of pharmacology and clinical pharmacology, but these subjects have not been taught yet. The third important reason is the fact that students in the fourth year still do not have enough practical training (work with patients) to deal with issues that can be found in textbooks. These questions are structured in a way so that they reflect book knowledge, whereas experienced doctors recognized them as less logical. Another reason the students evaluated some questions with a lower grade is the fact that the syllabus is not fully compliant in terms of the connectivity of the syllabus of each subject and their position in the individual semesters. For example, within the subject of pathophysiology, students work with topics that include elements of pharmacology and clinical pharmacology, but these subjects have not been taught yet. The third important reason is the fact that students in the fourth year still do not have enough practical training (work with patients) to deal with issues that doctors in their everyday work constantly encounter.

**DISCUSSION**

It is very common for students to complain on a postexam consultation, saying that the test questions were vague and that the exam calls for the memorization of insignificant details and lists of data, with little practical content. On the other hand, many teachers believe that students should not evaluate the quality and content of tests on the merit of their logic and practicality because they are not competent enough. This was precisely the reason for conducting research and writing a paper that showed that student exam question quality assessments differ very little from the rating of family medicine doctors. This study, although conducted on a small scale and on only one subject, indicated that the students’ assessments are compatible with the assessments of physicians as practitioners who are experts in identifying what is practical and logical for future work.

The curriculum of pathophysiology at our faculty is composed of both a theoretical and a practical part. During the theoretical classroom lectures, especially during seminars, and

<table>
<thead>
<tr>
<th>Question No.</th>
<th>Item No.</th>
<th>Cardiology</th>
<th>Pulmonology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>1</td>
<td>L 4 (3.25–5)</td>
<td>4 (3.25–5)</td>
</tr>
<tr>
<td>2.</td>
<td>2</td>
<td>P 5 (4–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>3.</td>
<td>3</td>
<td>L 5 (5–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>4.</td>
<td>4</td>
<td>P 5 (5–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>5.</td>
<td>5</td>
<td>L 4 (4–5)</td>
<td>4 (3–5)</td>
</tr>
<tr>
<td>6.</td>
<td>6</td>
<td>P 5 (4–5)</td>
<td>4 (3–5)</td>
</tr>
<tr>
<td>7.</td>
<td>7</td>
<td>L 4 (3–5)</td>
<td>5 (3–5)</td>
</tr>
<tr>
<td>8.</td>
<td>8</td>
<td>P 4 (3–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>9.</td>
<td>9</td>
<td>L 5 (4–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>10.</td>
<td>10</td>
<td>P 5 (4–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>11.</td>
<td>11</td>
<td>L 4 (3–5)</td>
<td>4 (3–5)</td>
</tr>
<tr>
<td>12.</td>
<td>12</td>
<td>P 5 (4–5)</td>
<td>4 (3–5)</td>
</tr>
<tr>
<td>13.</td>
<td>13</td>
<td>L 4 (3–5)</td>
<td>5 (3.75–5)</td>
</tr>
<tr>
<td>14.</td>
<td>14</td>
<td>P 5 (4–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>15.</td>
<td>15</td>
<td>L 5 (4–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>16.</td>
<td>16</td>
<td>P 5 (5–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>17.</td>
<td>17</td>
<td>L 4 (3–5)</td>
<td>4 (4–5)</td>
</tr>
<tr>
<td>18.</td>
<td>18</td>
<td>P 4 (3–5)</td>
<td>4.5 (4–5)</td>
</tr>
<tr>
<td>19.</td>
<td>19</td>
<td>L 5 (4–5)</td>
<td>5 (4–5)</td>
</tr>
<tr>
<td>20.</td>
<td>20</td>
<td>P 5 (4–5)</td>
<td>5 (4.75–5)</td>
</tr>
</tbody>
</table>

P, practicality; L, logic. *Statistically significant.
During the hands-on student practical work, we talk about pathophysiological mechanisms of the conditions and diseases with supporting explanations to achieve a constructive alignment with the goals and assessment of our department. During the design of the study, methods and literature principles of constructive alignment with their future practice were taken into account. With respect to the adequacy of the assessments being consistent with the opinions of family doctors, students were given the appropriate materials with very specific areas to learn from, which allowed them to perform appropriate assessments in relation to the doctors of family medicine. At our institution, the subject of family medicine is in the sixth year of study, and during the formation of the curriculum of all subjects, teachers of family medicine and the teachers of all of the other departments provide comments and suggestions. In this way, the assessment of students and family doctors is aligned. As the survey results show, there is still room for improving the design of the course in accordance with Biggs’ theory of constructive alignment, starting from a more precise definition of the objectives and outcomes through the teaching and learning methods so that the students can understand the practicality and logic of the assessment questions in the future and also comply with the perceptions of practitioners. The questions that showed important differences indicate the need to further improve the teaching process. The teaching process should be structured in such a way that it is more focused on explaining the importance of learning topics for future work and also leaves enough space for student involvement, including question asking, to obtain better explanations of the facts to enhance and make classes more useful.

We are aiming to explain the differences between student and doctor assessments of the questions. For item no. 3 in cardiology, a possible explanation for the slightly higher logic rating from the doctors is that doctors are likely to regularly be in a situation to watch patients with chronic stress who then have coronary artery spasms or functional stenosis, whose concepts are contained in the given question. Regarding item no. 7 from cardiology, concerning logic, students gave in this case a slightly higher score in relation to the doctors. The possible explanation is that the students have recently had details of this situation (stenosis, insufficiency, pressure, and volume overload) presented in physiology classes. Item nos. 6, 14, and 16 in cardiology have different marks for practicality. On two items, doctors gave higher grades. On one item, students gave higher grades. We expect these to be different because doctors have a lot more practical experience than students. On item no. 3 from the field of pulmonology, the higher grade from the logic point of view was given by doctors. The item is related to hypercapnia and its causality, which was recently presented to the students in physiology and biochemistry, and is not very common in primary health care. In the pulmonology part, in the assessment of logic on item no. 5, which refers to endobronchial edema, doctors gave a slightly higher score because it is a state they quite often encounter, particularly in patients with chronic obstructive pulmonary disease and heart failure. For the practicality of other items regarding pulmonology, students gave higher grades for item nos. 4 and 10, but on item no. 20 the higher grade was the one given by the doctors. Still, we believe that the grades given by the doctors of family medicine are more relevant because of the lack of involvement of medical students in practical work. However, when comparing results with item statistics for the cardiology test, it can be concluded that all questions from which differences occur are moderately high to high item/total correlations, except for question 8. Question 8 is the most difficult question on the cardiology test \(P = 0.44\) and has the lowest correlation with the test results \(r = 0.16\). This question was significantly assessed by students as less practical than doctors, which may be the result of the question difficulty and also due to the fact that the question was not constructively aligned with the goals and methods of our teaching. Also, when comparing the relative ratio of the items for which there is a significant difference and the items in which there are no significant differences, in the category of easy questions (questions for which \(P \geq 0.7\)), the ratio is 1:5 in favor of items without distinction, whereas in the category of moderately difficult questions that ratio is 1:1.7 in favor of items without distinction. In pulmonology, the difference with easy questions is even larger at 1:7, which confirms that there is a greater degree of agreement between doctors and students, whereas in the group of moderately difficult questions that ratio is 1:2. In both the fields of cardiology and pulmonology, according to the point biserial correlation between groups of average \(r = 0.1 – 0.3\) and good questions \(r > 0.3\), there is the same percentage. For 75% of the questions there are no differences between the assessments of medical students and family medicine doctors. These results provide grounds for improving such relations and future activities to constructively complement the objectives, teaching, and testing, especially in the category of moderately difficult questions. The explanation for the relatively high ratings of our questions may be due to the fact that diseases in cardiology and pulmonology are the most common ones in primary practice, as confirmed by Elnicki et al. (9). Keyword phrases rated highest (top 10%) where, out of the best 31, 21 of them were in the field of cardiology and pulmonology for relevance to the curriculum and the importance for knowledge with average scores of 4.3 to 5.0. According to Jozefowicz et al. (17), faculty spend a lot of time preparing lessons but spend little or no time reviewing the questions prepared for exams, resulting in a relatively low quality of in-house tests. One of the reasons given for this is to avoid criticizing colleagues. There is a lack of uniform standards and guidelines directed toward the quality of in-house exams. We believe that there is a need to conduct regular pre- and posttest analysis, use questionnaires about the quality of teaching and examinations, and obtain regular and special exam question assessments from the students as well as from faculty that were not involved in making the test constructions, family medicine doctors, and other experts. As noted by McAllister et al. (26), student involvement in the design of exam evaluation is very useful, with a strong congruence of student opinions and those of clinical educators and experts, which enabled the researchers to go forward with confidence in the design of the assessment tool, highlighting the efficiency of inclusion of students in the design of the assessment. In addition, according to Bowden and Marton’s (5) concept of assessment, the goal is that students are constantly engaged in the important aspects of these problems.

Vanderbilt et al. (34) recommend that medical schools should review their examination questions to help prepare their students for the United States Medical Licensing Examination (USMLE), with an emphasis on evaluation of questions about...
applicable knowledge. Each medical school/university (possibly at the state level) should regulate their test procedures in detail, have specific regulation with an emphasis on higher levels of knowledge (long-term knowledge), and use questions with practical applications. According to Hamdy et al. (15), the highest correlation was between scores on USMLE Part II and Part III and the lowest among USMLE Part I and supervisor ratings during residency $r = 0.22$. This low correlation speaks in favor of a necessary separate assessment, which should be done regularly and enforced by legislation. Various forms of written examinations have a low correlation with the observed structured clinical exams (30). This is the reason for further evaluation of these methods, with an emphasis on audit of university written exams.

The validation process for MCQs of the Swiss licensing examination checks the content and significance of each question in accordance with general practice and the Swiss Catalogue of Learning Objectives for Undergraduate Medical Training. Any revisions are carried out by interfaculty and multidisciplinary groups of clinicians and representatives of general practitioners (13). The above-mentioned process should also be applied to the university exams with student engagement in those groups. There are not many studies (19, 21) that compare the feedback of students and faculty, and those studies that do exist are related mainly to lectures. Since it is difficult to interview majority of the current faculty, an alternative approach is to combine students, faculty, family medicine doctors, and other experts to enhance student learning.

The study by Harris et al. (16) analyzes test questions on an index of difficulty (ID item difficulty) $\leq 0.6$ (relatively simple questions), and their results suggest that family medicine postgraduate students evaluate questions accurately, thus ensuring that relevant and nonproblematic questions can be used to test undergraduate students. A much bigger problem is the questions with a very low index, indicating high difficulty, which are often found in tests.

The high concordance we found between the grades of medical students and family medicine doctors suggests the possibility of quality assessment of the given questions by medical students under the supervision of family medicine doctors. Historically, in general, professors and staff as cited by Wallach et al. (36) invest a lot of time in the preparation of lectures, seminars, and practical lectures, but little time is allocated to the process of creating test questions, which is very important for learning outcomes. It takes a lot more time to create adequate examination questions, since examination questions directly determine the way of learning, access to lectures, seminars, and practical lectures. Regardless of the presence of numerous university programs and references on how to design appropriate test questions (6, 12, 14) many university in-house exams are of low quality and even violate the most basic rules of writing test questions (3, 8). Consequently, there is a need to continuously regulate and control the quality of testing procedures.

As noted by MacLellan (23), it is very important to recognize the central role of students in assessment. According to Cross (7), there are three essential conditions for excellence: high expectations, participation and involvement of students, and assessment and feedback. Since we are committed to student-centered learning, students need to play a very important role in assessment and feedback, which is an integral part of teaching and learning (33). As mentioned by Sadler (32) and Taras (33), it is necessary to involve students as active participants and protagonists in the assessment process. Although assessment and feedback play a central role in learning, student participation in assessment is still quite rare in higher education (25, 33).

We acknowledge that our study is limited to examination questions presented to third-year students of the Medical Faculty University of Sarajevo. The results and conclusions that we reached are based on the findings specific to our test procedure, and it may be difficult to generalize these findings to other medical schools. For the given topic, the assessments given by the students and general medicine doctors may be subjective. There is also a small sample size and study duration, because the total duration of each course of cardiac pathophysiology and pulmonary pathophysiology is 2 wk. Also, this analysis supports further studies of compliance of test question assessment by students and family medicine doctors.

However, medical schools can use our methods to examine the degree of concordance between the grades given by experts and medical students to raise the quality of testing procedures. In conclusion, based on these indicators, as well as the results of this study, students in coordination with other experts should be involved in the procedures of quality assessment of test questions, which should be a regular, strictly regulated process that is constantly evaluated.

ACKNOWLEDGMENTS

We thank the doctors of the Health Center of Sarajevo Canton and the fourth-year students of the Medical Faculty University of Sarajevo who participated in the preceding study. We thank Adem Cemerci, current medical student at the Medical Faculty University of Sarajevo and University of Delaware alumnus, for linguistic and academic contributions to this article.

GRANTS

This study received research funding from the Ministry of Education and Science of the Federation of Bosnia and Herzegovina.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS


REFERENCES


