Difficulty, discrimination and cognitive level of Microbiology exam questions in the Faculty of Medicine of Tunis

Difficulté, discrimination et niveaux cognitifs des questions d'examens de microbiologie à la Faculté de Médecine de Tunis

Amene Hermi, Wafa Achour

Centre National de Greffe de Moelle Osseuse, Service des laboratoires, Faculté de Médecine de Tunis

RÉSUMÉ

Introduction: Se basant sur différents indices (difficulté, discrimination) et les niveaux cognitifs de Bloom, nous avons évalué la qualité des questions de Microbiologie de la session principale 2012-2013 à la Faculté de Médecine de Tunis.

Méthodes: Nous avons analysé 70 questions: 16 (examen «A»), 28 («B1») et 26 («B2»), résolues respectivement par 533 (1ère année), 285 et 292 (3ème) étudiants. Pour chaque question, nous avons déterminé les indices de difficulté, discrimination et le plus haut niveau cognitif exploré. Les indices moyens de difficulté et de discrimination pour chaque examen et niveau cognitif, et ceux de discrimination pour chaque niveau de difficulté, ont été déterminés.

Résultats: Les 70 questions étaient de difficulté recommandée (0.58), bonne discrimination (0.31) et exploraient principalement (58.57%) le plus bas niveau cognitif. La difficulté était acceptable en 1ère et 3ème années; la discrimination respectivement marginale (0.27) et bonne (0.33). «A», «B2» et «B1» exploraient respectivement, les bas, bas et hauts et les 3 ordres cognitifs. Chaque niveau cognitif, excepté l'intermédiaire, était de difficulté acceptable. La discrimination était bonne pour tous les ordres cognitifs excepté l'ordre bas de 1ère année (0.27). La discrimination était marginale pour les questions difficiles (0.29) et bonne pour les autres. Comparé à B2, B1 était plus à la portée, plus discriminatif (aucune question de mauvaise discrimination) et explorait tous les niveaux cognitifs.

Conclusion: Bien que notre étude ne permet que des conclusions limitées à des questions spécifiques, elle pourrait servir de modèle pour d'autres, similaires, visant à améliorer la qualité de nos examens.

Mots-clés

Indice de difficulté, l'indice de la discrimination, la taxonomie de Bloom, examens de microbiologie

SUMMARY

Introduction: Based on difficulty and discrimination indices and cognitive levels of Bloom, we assessed in this study the quality of Microbiology exam questions (main session 2012-2013, Faculty of Medicine of Tunis).

Methods: We analyzed 70 questions: 16 (exam"A") given to 533 students (1st year), 28 and 26 (exams"B1" and "B2") given respectively to 285 and 292 students (3rd year). For every question, we determined difficulty and discrimination indices and the highest cognitive level required to resolve it. We calculated mean difficulty and discrimination indices for each exam and cognitive level, and mean indices of discrimination for every difficulty degree.

Results: The 70 questions were of optimum difficulty (0.58), good discrimination (0.31) and explored mainly (58.57%) the lowest cognitive level. For both years, mean indices of difficulty were acceptable, while those of discrimination were good (0.33) and marginal (0.27) for respectively 3rd and 1st year. "A" explored Lower Orders of Cognitive Skills (LOCS), "B2" both Lower and High Orders and "B1" all orders. Mean difficulty indices of every cognitive level were acceptable except for the median one (0.83). Mean discrimination indices were good for all cognitive levels except for LOCS of the 1st year (0.27). Mean indices of discrimination were marginal (0.29) for difficult questions and good for others. Compared to B2, B1 was more attainable and discriminative, free of poor discrimination questions, and explored all cognitive orders.

Conclusion: Our study remains specific to particular questions and generalizations seem difficult. However, it can serve as a guideline to other similar studies.

Key-words

Difficulty index, discrimination index, Bloom's taxonomy, microbiology exams

Adequately measuring medical knowledge is an essential component for evaluation, provision of reliable feedback, and improvement of medical education [1]. Studies worldwide use many approaches to assess exam questions, such as item analysis (based on students responses to different questions) and Bloom's taxonomy of cognitive levels (related to the question itself, identifying it by objective) [2]. The latter is a well-defined and broadly accepted tool for categorizing types of thinking into six different levels: knowledge, comprehension, application, analysis, synthesis, and evaluation. It is used to develop assessment questions at an appropriate cognitive level [3]. In fact, most faculties of medicine would agree that academic success should be measured not just in terms of what students can remember (knowledge), but what students are able to do with their knowledge [3]. This is simply because in their practice, physicians routinely encounter ambiguous, undifferentiated clinical problems that require higher order of thinking, not simply recall of knowledge and skills [4]. It is commonly accepted that the first two levels of Bloom's (knowledge and comprehension) represent Lower Orders of Cognitive Skills (LOCS) that require only a minimum level of understanding. Analysis, synthesis and evaluation are considered as High Orders of Cognitive Skills (HOCS) that require deep conceptual understanding and students often have difficulty performing at. Application represents the median level, the transition between LOCS and HOCS [3]. This highlights the need to more rigorously evaluate medical students by measuring outcomes that reflect higher order processes [1]. However, writing questions that evaluate HOCS can be challenging, and most teachers are not formally trained. This lack of training results in the development of basic recall questions administered to students to assess just their knowledge, whereas the questions should make them think through a series of steps until reach synthesis and evaluation levels [1].

The other approach to assess the quality of test questions is item analysis which is the process of collecting, summarizing and using information from students' responses to revise and then modify or remove specific questions [5, 6]. It uses many statistic tools such as difficulty and discrimination indices. Difficulty index appreciate the level of difficulty of the question. It ranges from 0 to +1. The higher is the value, the easier is the question [6]. Discrimination index is the measure of a question's ability to discriminate between higher and lower-performing students based upon correlation between the question score and the overall measure of performance. It ranges from -1 to +1 and negative indices are clear indicators of a problematic question [6]. Based on difficulty and discrimination indices as well as cognitive levels of Bloom, we aimed in this study, to assess for the first time, the exam questions proposed by the Microbiology department of the Faculty of Medicine of Tunis (FMT) in the academic year 2012-2013.

METHODS

In every main session, the Microbiology department of the FMT provides questions contributing in 3 multidisciplinary examinations: One given to the 1st year, the "disease factors" paper, and two given to the 3rd year, the certificate of "infectious diseases" (one to each mid promotion).

In this study, we focused in overall 70 questions proposed in 2012-2013 by the department, gathered from 3 examinations:

• 16 Short Answers Questions (SAQs) in the "disease factors" paper, passed by 533 students (Exam A)

• 28 questions (SAQs, vignettes and essay questions) in "infectious diseases" paper passed by 285 students in the first semester (Exam B1).

• 26 questions (SAQs, vignettes, essay questions) in "infectious diseases" paper passed by 292 students in the second semester (Exam B2).

For each exam, we were provided with a hard copy and an Excel file containing the marks obtained by every student on each of the questions. First, to obtain the difficulty and discrimination indices for each question, we used "AnItem.xls", a specific Excel file used for such purposes at the assessment office of the Faculty of Medicine of Montreal University since 2000 and free downloaded online. It gives us those two indices directly after a brief and easy adaptation of the Excel file of marks. All the other calculations of the different mean indices were done with ordinary excel functions.

"AnItem.xls" calculates the difficulty index was as follows: the mean mark obtained by all candidates attempting the item divided by the maximum mark available on the item [7]. We used these intervals to classify the index values obtained: "_> 0.7 – Easy, _0.3-0.7 – Acceptable, _ < 0.3 – Difficult, _ 0.5-0.6 – Optimum [5, 8, 9].

Discrimination index is calculated by "AnItem.xls" using the corrected item-total correlation. It is a Pearson correlation between the sum of marks of the examinees on the item and the sum of their scores on all the other items [10]. We used this classification to interpret the index values: "≥ 0.4 – Excellent, 0.3-0.39 – Good, 0.2-0.29 – Marginal, ≤ 0.19 – Poor (subject of rejection) [2].

After determining the difficulty and discrimination indices for each question with "AnItem.xls", we calculated the mean indices for the overall microbiology part in each exam. We also discussed each question to determine the highest Bloom's cognitive level required for its solution.

Finally, to determine the relation between the 3 elements of analysis (difficulty, discrimination and cognitive level), we calculated for each exam the mean indices of difficulty and discrimination for the LOCS, HOCS and the median level (application) as well as the mean indices of discrimination of the different degrees of difficulty.

RESULTS

The overall 70 questions of Microbiology analyzed were of optimum difficulty (mean index= 0.58), good discrimination (mean index= 0.31), and explored mainly the lowest level, knowledge (58.57%).

The mean index of difficulty was acceptable in both years and particularly optimum (0.56) in the 3rd year. The majority of questions in the 3 examinations were of acceptable difficulty: 53.57% in B1, 69.23% in B2, and 62.5% in A. The latter was the only that didn't contain any difficult question (Table 1).

The mean index of discrimination was good for the 3rd year (0.33) whereas it was marginal for the 1st one (0.27). The majority of questions were of good discrimination in the 3rd year (50% in B1, 42.30% in B2), and of marginal discrimination in the 1st (50%) (Table 1). None of the questions of B1 was of poor discrimination, meaning that none was to be rejected. The Exam A didn't contain any question of excellent discrimination (Table 1).

Table 1: Difficulty, discrimination and cognitive level of exams questions

		B1	B2	A
	N.questions	28	26	16
Cognitive level	Knowledge	12	16	13
	Comprehension	3	2	3
	Application	2	0	0
	Analysis	1	2	0
	Synthesis	5	5	0
	Evaluation	5	1	0
Difficulty	Difficult	1	5	0
	Acceptable	15	18	10
	Easy	12	3	6
Discrimination	Poor	0	3	2
	Marginal	7	7	8
	Good	14	11	6
	Excellent	7	5	0

The majority of questions in the 3 examinations were of LOCS (53.57% in B1, 64.23% in B2 and 100% in A) and more precisely of the lowest one (knowledge). The questions of Exam B2 were of both LOCS and HOCS, while those of B1 were of the 3 orders (LOCS, HOCS and application) (Table 1). All mean difficulty indices calculated for the 3 cognitive orders in all exams were acceptable except for the median level where it was found easy (0.83) (Table 2). All mean discrimination indices calculated for the 3 cognitive orders in the 3 exams were good except for the LOCS of the 1st level which was marginal (0.27) (Table 2). Mean indices of discrimination calculated for each degree of difficulty were marginal for difficult questions (0.29), and good for both acceptable difficulty (0.32) and easy questions (0.34) (Table 3).

We found that B1 and B2, delivered to the same year students, had many similar points. Both had a majority of questions of acceptable difficulty, good discrimination and of LOCS. Nonetheless, B1 had stronger characteristics. In fact, it was more attainable (mean difficulty index 0.65 versus 0.46), more discriminative (mean index 0.34 versus 0.32), didn't contain any question of poor discrimination, and had questions of all cognitive orders (while B2 questions were of LOCS and HOCS).

Table 3: Mean discrimination indices by difficulty degrees

	B1	B2	A
Difficult questions	0.21	0.31	
Acceptable difficulty	0.34	0.32	0.26
Easy questions	0.35	0.29	0.27

DISCUSSION

High quality questions are important for medical student assessments. This is even more essential for summative assessments where medical student performance influences their progression within the program and results in a final grade for a course [1]. The three parameters used in this study are obviously of great interest in determining the overall quality of exam questions.

We found that the majority of questions were of acceptable difficulty (61.43%). The analysis of 50 MCQs of physiology paper of the 1st year students of Dental College in Karachi found that 78% of the questions were of average difficulty with a mean index of 0.54 [6]. Another study in the Municipal Medical College in India, conducted on 50 MCQs of MMBS curriculum, found that 80% of the questions were of acceptable difficulty [8]. Another study done in a Faculty of Medicine in Iran, analyzing 1496 MCQs, found that 47.8% of the questions were of acceptable difficulty, 44.58% were easy, and 7.62% were difficult [9]. Easy questions (representing 30 % in our study) should usually be placed either at the start of the test as "warm-up" questions or removed altogether. The difficult items (that represented 8.75% in our series) should be reviewed for possible confusing language or areas of controversy [6].

The majority of questions were of good discrimination in the 3rd year (50% in B1, 42.30% in B2), and of marginal discrimination in the 1st (50%). A study analyzing 150 MCQS of the Royal College of General Practitioners International Examination in Oman, in 2008, showed that only 18.66% (22/150) of the questions were of good discrimination whereas the majority (49.33%, 74/150 MCQs) were of poor discrimination [11]. Even if questions of poor and marginal discrimination were a minority in our study (38.57%), they still should be referred back to experts for revision to improve their standard. The different causes to such a discriminative power may be a miskey, an ambiguous wording, grey areas of opinion, areas of controversy, guessing, information not covered or inadequately covered, or too easy/ too difficult questions [5]. In fact, the mean discrimination index

Table 2 : Mean difficulty and discrimination indices by cognitive level

	Mean difficulty index			Mean discrimination index		
	B1	B2	A	B1	B2	A
Total	0.65	0.46	0.63	0.34	0.32	0.27
LOCs	0.62	0.50	0.63	0.33	0.32	0.27
Median level	0.83			0.39		
HOCs	0.66	0.38		0.35	0.32	

calculated for all difficult items in the study was the worst (0.29).

The great majority of questions in the study were found to be of LOCS (58.57%). Most exams developers would agree that students should be taught and tested for higher-cognitive skills. This disagreement between convictions and results is often found in many institutions. In fact, when teachers rank their own exam questions using Bloom's, they often realize that the majority of their test questions are at the LOCS [3]. For example, at a national meeting for undergraduate biology education, 97% of the teachers at Faculty of Washington who attended and received a formal lecture on using Bloom's to rank exam questions agreed that only 25% of their exam questions tested for HOCS [3]. Therefore, most of the time we may not be testing or providing students with enough practice at using content and science process skills at higher cognitive levels, even though our goals are that they master the material at all levels. In this setting, it is important to underline that no one of the questions analyzed in this study was a multiple choice question (MCQ). They were SAQs, essay questions or vignettes. Although MCQs are considered as the most practical assessment method in medical education, they are often accused of focusing on the knowledge recall [12]. Contrariwise, vignettes are known to assess HOCs rather than simple recall [1]. So an explanation for the discrepancy found between goals and facts, may be that teachers have not been given the tools and guidelines that would help them to better develop questions at HOCS [3]. For example, a not well-constructed vignette can become a pseudovignette, which is defined as a "clinical" vignette ending with a recall question (declarative knowledge). Therefore, all of the information provided in it is not necessary to answer the question. Pseudovignettes can be prevalent within an examination that is written by a teacher who has not received training or support in test development. It is for sure time consuming and challenging to write a true clinical vignette asking clinical or even basic science questions; however, this can lead to better test questions and assessment of students' knowledge at a higher level [1]. So a particular attention should be done to Bloom's cognitive levels when preparing exam questions because of its importance in accurately gauging student performance. Furthermore, the cognitive challenge of exam questions strongly influences students' study strategies [3]. In fact, if learning activities focus on concepts requiring HOCS but exams test only on LOCS, students quickly learn that they do not need to put forth the effort to learn the material at a high level. Similarly, if teachers discuss simple subjects in courses but test at a higher cognitive level, students often perform poorly on examinations because they have not been given enough practice developing a deep conceptual understanding of the material [3].

The LOCS were found to produce marginal discrimination and Exam A, comprising 100% of LOCs questions, was the only exam that had marginal discrimination while the others assessing higher levels were found to have good discrimination. In fact, while the majority of students can respond to questions of simple recall, it will not be the same for questions assessing HOCS.

B1 was found to have stronger characteristics than B2 (cognitive level, difficulty and discrimination) in spite the fact that the same teachers of the Microbiology department asked questions in the two semesters. In order to ensure equity between all students of a same level, exams delivered in the two semesters in the same certificate should be as much as possible close in terms of questions quality. Nonetheless, a perfect equity is obviously impossible because indices depend also of the group tested itself.

We acknowledge that our study is limited to the microbiology questions that were administered during the academic year 2012-2013. The findings and conclusions that we draw are related only to this particular year and it will be difficult to make generalizations. However, in the future, we highly recommend that the Microbiology department as well as the other departments of the FMT, and even other medical schools of our country, take the time and effort to evaluate their current assessments to ensure a higher quality exam questions.

CONCLUSION

The findings of the present study highlight the importance of both item analysis and Bloom's taxonomy in the assessment of the quality of exam questions in the purpose to adjust and improve them to guarantee a higher level of assessment.

References

1. Vanderbilt A, Feldman M, Wood I. Assessment in undergraduate medical education: a review of course exams. Med Educ Online 2013; 18:1-5.
2. Shafizan S. Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. IJER 2013; 1:1-14.
3. Crowe A, Dirks C, Pat WP. Biology in Bloom: implanting Bloom's taxonomy to enhance student learning in biology. CBE Life Sci Educ 2008; 7: 368-381.
4. Plack M, Driscoll M, Marquez M, Cuppernull L, Maring J, Greenberg L. Assessing reflective writing on a pediatric clerkship by using a modified bloom's taxonomy. Ambul pediatr 2007; 7: 285-91.
5. Karelia B, Pillai A, Vegada B. The levels of difficulty and discrimination indices and relationship between them in four-response type multiple choice questions of pharmacology summative tests of year II M.B.B.S students. IeJSME 2013; 7: 15-20.
6. Hingorjo MR, Farhan J. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. J Pass Med Assoc 2012; 62: 142-7.
7. MacAlpine M. A summary of methods of item analysis. CAA Centre Bluepaper 2, University of Luton 2002.
8. Kartik A, Neeraj R. Itemized analysis of questions of multiple choice question (MCQ) exam. ISJR 2013; 2: 279-80.
9. Tabatabaee M, Bahreyni TM, Derakhshan A, Dalloee M, Gholami H. Analytic assessment of multiple-choice tests. IJME 2003; 2:87-91.
10. Henrysson S. Correction of item-total correlations in item analysis. Psyckometrika 1963; 28: 211-218.
11. Theodorsson T, El Shafie K, Al Wardy N, Al Maherzi A, Al Shafaee M. Assessment of Family Doctors in Oman: getting the questions right preliminary findings of a performance analysis of multiple choice questions. Int J Med Educ 2010; 1: 2-5.
12. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC med educ 2007; 7: 1-7.