EUROPEAN JOURNAL OF TRANSLATIONAL
AND CLINICAL MEDICINE 2019;2(1):22-27

RESEARCH ARTICLE

Statistical methods for the assessment of students' knowledge as illustrated by the Introduction to Internal Medicine exam

Adrian Ireneusz Stefański[1], Natalia Bladowska[2], Aleksandra Więczkowska[2], Bogdan Wojtyniak[3], Agata Ignaszewska-Wyrzykowska[1], Hanna Jasiel-Wojculewicz[1], Jarosław Jendrzejewski[4], Marcin Rutkowski[1], Tomasz Zdrojewski[1]

[1]Department of Prevention and Didactics, Medical University of Gdańsk, Poland
[2]Faculty of Applied Physics and Mathematics, Gdańsk University of Technology, Gdańsk, Poland
[3]National Institute of Public Health-National Institute of Hygiene, Warsaw, Poland
[4]Department of Endocrinology and Internal Medicine, University Clinical Center, Gdańsk, Poland

Abstract

Background: We wanted to develop substantial and statistical methodology for complex assessment of quality of teaching internal medicine in medical university. Our aim was also to check connection between the results obtained during the midterm and final exam. **Materials and methods:** We have compared the results obtained by Polish (n=235) and English Divisions (n=81) students achieved during the midterm exam and multiple-choice final exam. The mean scores were calculated with t-Student test. For further evaluation Wilcoxon tests were used with the Bonferroni correction, The Stuart-Maxwell test was carried out to verify the hypothesis about correlations between results in the midterm and final exam. **Results:** The mean midterm exam score was 84.4% in PD and 72.6% in ED ($p < 0.0001$) and mean final exam score was respectively 72.3% and 55.6% ($p < 0.0001$). Good result of the final exam was obtained by 62% of students who passed well the midterm exam. **Conclusions:** It is crucial to use appropriate tools to grade the quality of tutorship. To evaluate that one should use advance statistical tests. The fact that ED students achieve less points on the exams might have few reasons like a language barrier. Obtaining a good result during midterm exam does not guarantee passing the final exam.

Keywords: exam evaluation • statistical methodology • Iinternal medicine

Citation

Stefański AI, Bladowska N, Więczkowska A, Wojtyniak B, Ignaszewska-Wyrzykowska A, Jasiel-Wojculewicz H, Jendrzejewski J, Rutkowski M, Zdrojewski T, et al. Statistical methods for the assessment of students' knowledge as illustrated by the Introduction to Internal Medicine exam. Eur J Transl Clin Med. 2019;2(1):22-27.
DOI: 10.31373/ejtcm/105493

Corresponding author:
Adrian Ireneusz Stefański, Department of Prevention and Didactics, Medical University of Gdańsk, Poland
e-mail: astefanski@gumed.edu.pl
No external funds.
Available online: www.ejtcm.gumed.edu.pl

Introduction

For several years, increasing attention has been paid to improving the quality of teaching at universities, also in Poland. During this time, several initiatives were aimed at systemic improvement of the teaching model. One of the effects of these efforts was the introduction of the National Qualifications Framework, which includes the development of teaching standards for universities. Currently, the curricula are based on demonstrating the potential for student to acquire and consolidate knowledge, skills and competences related to the particular subject. Despite the unquestionable value of the new standards, they were criticized for listing far too detailed teaching goals, the need to verify skills and competencies, especially at general universities where the basic goal is to acquire knowledge [1-3]. This is less of an issue in higher vocational schools where gaining professional skills and competencies to perform a specific profession are equally important as the acquired knowledge. Medical university curricula are developed taking into account a set of requirements and learning outcomes [4].

In accordance with the Act of 20 July 2018 on the Law on Higher Education and Science, the Minister of Science and Higher Education in consultation with the Minister of Health issued a regulation defining unified educational standards. The current legal act is the Ministerial Announcement dated 9 January 2018, which specifies in detail the required learning outcomes divided into knowledge and skills in the field of morphological, pre-clinical, behavioral, clinical surgical and non-surgical sciences and the related legal aspects [5-6]. In addition, this regulation highlights the various methods that can be used to evaluate the pre-defined learning outcomes. Students' knowledge can be evaluated by oral or written exams. Types of written exams assessing knowledge include written essays and reports, open-ended questions, multiple-choice tests (single-select questions, multiple-select questions, true/false questions and matching questions).

Multiple-choice (often described as 'objective exams') became a commonly used tool to assess students' knowledge at various levels of their education, including medical universities. This is a quite effective method to check the scope and depth of knowledge acquired by students. Furthermore, such tests also indirectly assess the effectiveness of teaching. Multiple-choice tests are considered the fairest way to verify the level of mastery of the material, because all students are evaluated anonymously, in the same reproducible way, and the result does not depend in any way on the examiner [7]. An additional advantage of this form of assessment is the fact that many students can take this exam simultaneously and yet the results can be an-nounced in a very short time thanks to a properly prepared answer sheet and its computer evaluation, limiting the possibility of human error and bias while checking the test [8]. Addi-tionally, the multiple-choice test format makes it easy to gather data for further evaluation of the test itself, the individual questions or the level of mastery of the material to which the questions referred. Using adequate statistical tools could help to verify which parts of in-structed material need further verification so that teacher could concentrate more on these subjects. The multichoice tests were introduced in the Department of Prevention and Didac-tics of the Medical University of Gdańsk twenty years ago and since then they have been in-creasing the quality of the knowledge assessment of our students [9-10].

Medical didactics is a field that is currently undergoing rapid development. Therefore, proper evaluation methods are needed. Unfortunately, there are few statistical methods in the available literature that would meet the objectives. Teaching a vast subject such as internal medicine is one of the main tasks of all medical schools and a challenge for medical teachers. Therefore, it is necessary to effectively and transparently monitor students' knowledge and achievements. The aim of our work was to develop a methodology for assessing and comparing the results of exams assessing knowledge of third-year students of the Polish and English Divisions acquired during the Introduction to Internal Medicine 2 course.

Materials and methods

Study definitions

The Introduction to Internal Medicine 2 course is included in the third-year curriculum of the Medical Faculty for both the Polish and English Divisions. The course includes 15 hours of seminars and 50 hours of bedside classes. In the second week of the spring semester students write a theoretical (open questions) and a practical midterm exam (whose results were not included in this analysis). The year-long course ends with a multiple-choice final exam.

Theoretical midterm exam

In the academic year 2016/2017, 233 students of the Polish Division (PD) and 80 students of the English Division (ED) wrote the theoretical midterm exam. The theoretical midterm exam conducted in the middle of the course consisted of 7 open questions with identical or similar content and level of difficulty (see

Appendix No. 1). Topics covered the symptomatology of basic cardiovascular, respiratory and digestive diseases discussed in classes so far. The questions were based mainly on the symptoms evident when taking patient's history and performing physical examination during seminars and bedside classes (if relevant patients were available). A maximum of 22 points could be obtained on the midterm exam.

Appendix 1.

- What is a physiological lung sound upon auscultation? List four causes of diminished sound upon auscultation of the lung. / What is a physiological lung sound upon percussion? List four causes of dull sound upon percussion of the lung.
- How can you diagnose dehydration upon physical examination?
- What signs and symptoms could you find upon physical examination in a patient with pneumonia? / What signs and symptoms could you find upon chest examination of a patient with pneumonia?
- In which acquired valve pathologies you might find a diastolic murmur upon auscultation of the heart? / In which acquired valve pathologies you might find systolic murmur upon auscultation of the heart?
- What is Horner syndrome? Write its causes and list its symptoms
- List symptoms of chronic right ventricular failure. / List symptoms of chronic left ventricular failure.
- Draw a graph of: - hectic fever, - Cheyne-Stokes breathing. / Draw a graph of: - intermittent fever, - of Kussmaull's breathing.

Multiple-choice test

Two hundred and thirty-five students of the Polish Division and 81 students of the English Division took the final exam. It was a final test written by the same students that wrote the midterm exam (changes in number of students between both exams were due to illness-related absence). The multiple-choice test consisted of 100 questions of various types (single-select questions, multiple-select questions, questions with negation, premise-conclusion questions). Question content was the same for Polish and English Division students. Each question contained 5 answers to choose from. Students had 100 minutes to write the exam. Three versions of the test were prepared for each group, differing only in the order of the questions. The questions checked the knowledge acquired during the Introduction to Internal Medicine course in the field of cardiology (21 questions), pulmonology (8 questions), endocrinology (8 questions), gastroenterology (8 questions), nephrology (13 questions), hematology (6 questions) hypertension and diabetes (13 questions) and physical examination (23 questions).

Statistical analysis

The statistical analysis of the theoretical midterm exam results was performed with the t-Student test for independent variables. The mean scores of the multiple-choice test results obtained by students of the Polish and English Divisions were calculated and subjected to statistical analysis with the t-Student test.

Levels of questions regarding different areas of internal medicine were assessed by treating every student as an individual statistical unit and the thematic category in which he or she obtained the best result, i.e. the highest score, was determined and the number of the students' best results was summed up for each category. The analysis of related variables was performed using non-parametric Friedman's test and Wilcoxon's test with the Bonferroni correction; (P-value <0.05). A total of 28 Wilcoxon tests, using the Bonferroni correction that involves multiple repetitions in the same pool of data sets, were performed successively within the categories of questions from the final exam.

To check whether the students who did well on the theoretical midterm exam also passed the final exam, their exam results were divided into score subgroups: every 5% for Polish students and every 10% for ED students (due to much lower number of ED students). Since there was no linear relationship between the results of the midterm and the final exam, the students of both divisions were pooled into one group and the analysis was repeated following the calcu-lation of Pearson's correlation coefficient. Then, the results of the midterm exam and the final exam were divided into two levels: level 0 (0–59% of correct answers – exam failed) and level 1 (60–100 % of correct answers – exam passed). McNemar's test was performed (P-value <0.05), which rejected the null hypothesis that the proportions of passing scores is equal for both exams. In the next analysis, the results of the midterm and the final exam were divided into three levels: level 0 (0–59% of correct answers – exam failed), level 1 (60–69% of correct answers – exam passed poorly) and level 2 (70–100% of correct answers – exam passed well). The Stuart-Maxwell test was carried out and the P-value was <0.05.

Results

The mean midterm exam score was 84.4% among the PD students and 72.6% among the ED students (p <0.0001). The mean final exam score was 72.3% (71.4, 73.2) PD students and 55.6% (53.4, 57.8) among the ED students (p <0.0001). The exam was failed by 5 PD students (2.1%) and 32 ED Students (40%). The specific topics in which the students obtained the highest scores (percentage of students who had the highest results in this area) were hematology and gastroenterology (88% and 86%, respectively), followed by hypertension/diabetes and endocrinology (49% each), pulmonology, cardiology and nephrology (42%, 39% and 23%, respectively) and physical examination (the lowest score of 12%, see Table 1). The Wilcoxon's tests with the Bonferroni correction showed that categories of questions from endocrinology, pulmonology, cardiology and physical examination were answered similarly. After dividing the students into two groups depending on the score obtained (level 0: 0–59% of correct answers – exam failed and level 1: 60–100% of correct answers – exam passed), it was shown that among those who failed the midterm exam (overall 53 examinees in both divisions), 36 students (68%) did not pass the final exam, but only every third student who failed the midterm exam had positive grade on the final exam (17 students, 32%). Of all those who passed the midterm exam, 136 students (62% of examinees) succeed on the final exam (Table 2).

Analysis performed after dividing the students into three score levels (level 0: 0–59% of correct answers – exam failed, level 1: 60–69% of correct answers – exam passed poorly and level 2: 70–100% of correct answers – exam passed well), showed that good result of the final exam was obtained by 23 students (11%) who failed the midterm exam and 136 students (62%) who passed well the midterm exam (Table 3). Additionally, we analyzed if the PD students' final exam results were correlated with in terms of admission to the University. To clarify, we compared the students who were admitted to our University during primary enrollment (exceeded the required number of points scored on the secondary school exit exams) and those who were qualified during additional enrolment period (students who did not earn the points required for primary enrolment, they fully participate in all the classes but have to pay for tuition). The mean exam scores among those 2 groups were 72.5% and 70.1%, respectively (p = 0.04).

Table 1. Number of students whose highest scores were in the respective topics

Category	Quantity
Cardiology	39
Diabetes and hypertension	49
Endocrinology	49
Gastroenterology	86
Physical examination	12
Hematology	88
Nephrology	23
Pulmonology	42

Table 2. The results of the midterm exam and final multiple-choice test

		Final test	
		0	1
MIDTERM	0	**36** (68%)	**17** (32%)
	1	**83** (38%)	**136** (62%)
TOTAL		**119**	**153**

0 (0–59% of correct answers – exam failed),
1 (60–100 % of correct answers – exam passed)

Discussion

It is difficult to find well-designed and effective statistical methods in the available literature that can be used for evaluation of exam results beyond just comparing mean values8. Simple statistical methods allow establishing which group of students achieved better learning outcomes and whether the difference between the data is statistically significant. Howev-

Table 3. The results of the midterm exam and final multiple-choice test

		Final test		
		0	1	2
MIDTERM	0	**10** (44%)	**7** (30%)	**6** (26%)
	1	**10** (33%)	**9** (30%)	**11** (37%)
	2	**23** (11%)	**60** (27%)	**136** (62%)
TOTAL		**43**	**76**	**153**

0 (0–59% of correct answers – exam failed),
1 (60–69% of correct answers – exam passed poorly),
2 (70–100% of correct answers – exam passed well).

er they cannot for example determine whether the students who achieved good results on a written midterm exam also obtained high scores on the final multiple-choice test or to find which topics on the exam were the least and the most difficult. That is why we decided to development our own method for statistical evaluation in order to assess the results from various types of exams and to identify the topics on which greater emphasis should be placed in the teaching process.

Our analysis also compared the results obtained by students of the Polish and English Divisions who answered the same questions. In both the theoretical midterm exam and final multiple-choice test, students of the Polish Division obtained statistically significantly higher scores (84.4% vs 72.6% and 72.3% vs. 55.6%, respectively; p <0.0001 for both). This may be due to various factors, with the language barrier being probably the most important. PD students are taught and examined in Polish which is their native language, whereas the ED students have all their classes in English which for majority of them is their second or even third language. Therefore, it takes more time for them to study the same material. It is noteworthy that English is also the second language of the teachers who wrote the exam questions, which might have caused additional problems in understanding the meaning of some of the exam questions. This issue has been known for quite a long time. In 1975 Massam reported that 60% of foreign medical school graduates who applied to work in United Kingdom, failed the English language and professional competence test [11]. This shows that there might also be some differences in teaching standards between the various countries around the world. Obviously all medical teaching and tests are prepared based on the medico-legal demands of the country where the teaching takes place.

The results of students admitted to the University in primary and additional enrolment (explained above) were also evaluated and the difference in scores between the two groups was statistically significant (72.5% vs. 70.1%; p = 0.04). This might be explained by the fact that students who were admitted during the additional enrollment round might have had some shortages in knowledge from the beginning (since they have had worse results on the secondary school exit exams).

The analysis of student's results achieved during midterm and final exams showed that obtaining a good result on the theoretical midterm exam does not guarantee a high score on the final test exam. Perhaps some students who received a high score on the midterm exam did not study enough before the final exam. In contrast, some of the students who failed the midterm exam might have been more motivated to better prepare for the final exam.

It is very important to use appropriate tools to check the quality of teaching. To evaluate didactics in such a complex subjects like internal medicine one should not use simple statistical tests but those that include the fact that the same student answered questions from different fields of internal medicine (related variables). The most important achievement of our work is the development of statistical methods that enable the assessment of complex parameters describing student results achieved on multiple-choice tests or open-question exams. The prepared tools can be used to analyze various forms of the assessment of students' knowledge regardless of the course being taught.

Limitations of the study

This study compared the results obtained by Polish and English Division medical students during midterm and final exams. PD students answered the questions in Polish which is their native language, as opposed to ED students whose exams were written in English which is the second or even third language for majority of those students and the second language for the majority of the teachers. This issue might be a very important reason for the results that we obtained. Authors also understand that statistical methods described above might not be applicable to every type of the examination but could be used to asses similar types of exams.

Conclusions

It is important to use appropriate statistical tools to evaluate the quality of tutorship and to discover areas of presented material which needs more attention put both by students and teachers. There might be several reasons for the finding that ED obtained worse results on the exams, however the language barrier seems to be the main issue. The students who performed well on the midterm exam were less likely to fail the final exam.

References

1. Biały K. Przemiany współczesnego uniwersytetu od idei von Humboldta do modelu uczelni przedsiębiorczej [Polish only]. Wydawnictwo Uniwersytetu Łódzkiego; 2011. 34–35 p.
2. Czerepaniak-Walczak M. Autonomia w kolorze sepii w inkrustowanej ramie KRK. O procedurach i treściach zmiany w edukacji akademickiej [Polish only]. In: Fabryki dyplomów czy universitas. Kraków: Oficyna Wydawnicza 'Impuls'; 2013. p. 37.
3. Pilarczyk PM. National Qualifications Framework and the reconstruction of higher education in Poland [Krajowe Ramy Kwalifikacji a przebudowa szkolnictwa wyższego w Polsce / Polish only]. Poznańskie Zesz Humanist. 2015;XXV:27–37.
4. Ustawa z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce [Internet]. Dziennik Ustaw. Kancelaria Sejmu RP; 2018 [cited 2019 May 6]. p. poz. 16668, Art. 68. Available from: http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20180001668
5. Obwieszczenie Ministra Nauki i Szkolnictwa Wyższego z dnia 9 stycznia 2018 r. w sprawie ogłoszenia jednolitego tekstu rozporządzenia Ministra Nauki i Szkolnictwa Wyższego w sprawie standardów kształcenia dla kierunków studiów: lekarskiego, lekarsko-dentys [Internet]. Dziennik Ustaw. 2018 [cited 2019 May 6]. p. poz. 345. Available from: http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20180000345
6. Rozporządzenie Ministra Nauki i Szkolictwa Wyższego z dnia 9 maja 2012 r. w sprawie standardów kształcenia dla kierunków studiów: lekarskiego, lekarsko-dentystycznego, farmacji, pielęgniarstwa i położnictwa [Polish only] [Internet]. Dziennik Ustaw. 2012 [cited 2019 May 6]. p. poz. 631. Available from: http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20120000631
7. Osterlind SJ. Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats. Osterlind SJ, editor. J Educ Meas. 1999;36(3):267–70.
8. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33(6):447–58.
9. Jasiel-Wojculewicz H, Zdrojewski T, Rutkowski M, Chwojnicki K, Bandosz P, Kaczmarek J, et al. Zastosowanie testów w ocenie nauczania Propedeutyki Interny:10 lat doświadczeń własnych w Akademii Medycznej w Gdańsku. Cz. 2. [Polish only]. Pol Arch Med Wewnętrznej. 2006;65(3):282–7.
10. Zdrojewski T, Jasiel-Wojculewicz H, Chwojnicki K, Szpakowski, P. Częstochowska, E. Wyrzykowski B, Miotk M. Zastosowanie testów w ocenie nauczania Propedeutyki Interny: 10 lat doświadczeń własnych w Akademii Medycznej w Gdańsku Cz. 1. [Polish only]. Pol Arch Med Wewnętrznej. 2006;65(3):276–91.
11. Massam A. First exam in UK for graduates of foreign medical schools fails 60%. Can Med Assoc J. 1975;113(5):468–9.