811.111::371.3:61
https://doi.org/10.18485/bells90.2020.1.ch29

Sofija Mićić Kandijaš[*]
Faculty of Medicine
University of Belgrade
Serbia

Danka Sinadinović[**]
Faculty of Medicine
University of Belgrade
Serbia

IMPLICATIONS OF MULTIPLE-CHOICE TESTING IN ENGLISH FOR MEDICAL PURPOSES

Abstract

Following the overall trend at the Faculty of Medicine University of Belgrade, the English language test has been transformed into a multiple-choice format. While the previous testing format checked if students acquired the necessary skills through a variety of exercises, the new format contains only multiple-choice exercises. In this paper we compare and contrast the two formats concerning the complexity of the tasks included and their influence on acquiring various aspects of the language. We concentrate on students' success in the final exam, our expectations and possible further improvements to the test. In accordance with our expectations, the new format proved to be more convenient and it meant better overall results in the final exam. Students score much better in grammar and vocabulary sections of the new format, whereas the reading comprehension section seems to be slightly more demanding than before.

Key words: EMP, ESP assessment, evaluation, multiple-choice testing

1. Introduction

Language testing is described as a social phenomenon, as it has multiple functions and is used not only in the field of education, but also in employment, international mobility, language planning and economic policy making (Fulcher 2010: 1). For this reason, it is not only General Language/English (GL/GE) that is tested, but Language/English for Specific Purposes (LSP/ESP) as well. LSP testing is sometimes described

[*] Faculty of Medicine, University of Belgrade, Dr Subotića 8, 11000 Belgrade, Serbia; e-mail: sofija.micic@mfub.bg.ac.rs
[**] Faculty of Medicine, University of Belgrade, Dr Subotića 8, 11000 Belgrade, Serbia; e-mail: dankas78@gmail.com

as a special case of communicative language testing, being based on contextualised communicative language, and it is thought to have equally good testing practice as other types of language tests (Douglas 2000: 1).

At the Faculty of Medicine, University of Belgrade, LSP testing has been present for three decades and it involves testing English for Medical Purposes (EMP). Over the years the test has evolved and it has been recently transformed into the multiple-choice format. Students' success in final exams have also changed over time, and it was interesting to check whether the introduction of the multiple-choice format had any impact on final test results. We expected it to have a positive influence.

Following some background information on language testing, LSP testing and multiple-choice tests, two LSP test formats at the Faculty of Medicine, University of Belgrade are presented in detail. The research conducted focuses on students' overall success in English Language 1 and English Language 2 final exams through comparing two test formats and their components. Some of the research results came as a surprise – although the overall success is better in the new format. The reading comprehension section seems to be more difficult in the multiple-choice format, leading to lower scores in that part of the exam. The research results are discussed bearing in mind both students' and teachers' needs, and some improvements to the current test format are suggested.

2. Testing in LSP

In language testing, just like in psychological and educational testing, the concept of validity has a central position and is highly valued in test evaluation (Fulcher and Davidson 2007: 14). According to Henning (1987: 170) and Hughes (1989: 22), validity means the appropriateness of a particular test or the fact that a certain test measures exactly what it is supposed to measure. Cronbach and Meehl (1955) identified three types of validity: criterion-oriented validity[1], content validity[2] and construct validity[3];

[1] This type of validity involves the relationship between a certain test and a particular criterion the testers would like to make predictions on (Fulcher and Davidson 2007: 4-5). In the case of English for Medical Purposes (EMP) it would mean using scores on test results to check if our students can manage reading research papers, medical textbooks in various subjects, communicating with patients, etc.

[2] It is necessary that the content of a particular test is a representative sample from the tested field, which is often accomplished by using expert judges (Fulcher and Davidson 2007: 6). In this particular case, this would mean that the texts selected for the tests are typical of the types of texts that are used in 1st and 2nd undergraduate medical courses. According to Carrol (1980: 67) in order to achieve content validity in English for Academic Purposes (EAP) domain, it is necessary to closely examine test takers, analyse their communicative needs and specify test content based on those needs.

[3] Constructs are concepts that are defined in such a way that they have become "operational" and they are measured in a test by being linked to something observable (Fulcher and Davidson 2007: 7).

this idea was widely accepted for several decades. Messick (1989) introduced a revised concept of validity, considering it to be a single concept with several different aspects. Since then, the questions of relevance and usefulness have become most important, which means that a test score needs to show that the knowledge, skills and abilities of a test-taker are justified (Fulcher 2007: 20). Bachman and Palmer (1996: 18) insisted on reliability, construct validity, authenticity, interactiveness and practicality as the most important features of a good test. All tests, language tests included, can be used for several different purposes: achievement, aptitude, diagnosis, placement and proficiency (Fulcher 2010: 21). In *norm-referenced tests* learners/test-takers are compared to each other according to their scores (Fulcher 2010: 31) whereas *criterion-referenced tests* (Glaser 1963) check if an individual test-taker has achieved a pre-specified criterion for a particular context. Language tests are usually designed by teachers who have received training in test design, or by people who specialize in test design (Fulcher and Davidson 2007: 28). The English tests at the Faculty of Medicine, University of Belgrade could be described as criterion-referenced tests with the purpose of checking students' achievement, aptitude and proficiency. They are also *summative* (Fulcher 2010: 3), as they measure proficiency at the end of a period of study (1st year and 2nd year), by which time the learners are expected to have reached a particular standard. These tests are designed by teachers and their validity, usefulness, reliability and practicality are carefully checked and revised.

The first testing method for LSP may have been the English Proficiency Test Battery (Davies 1967), which was based on a construct of General English. The beginning of real LSP testing is probably marked by Temporary Registration Assessment Board (TRAB) examination, which was introduced in 1975 by the British General Medical Council and was intended for evaluating professional and language abilities of physicians trained outside the UK (Rea-Dickens 1987; Douglas 2000: 3). So, LSP testing was introduced in the 1970s (Morrow 1979; Carrol 1980), it was communicative and it used authentic language, unlike GE tests.

An LSP test relies on specific purpose language ability. Its measurement depends on the interaction between the language knowledge of the test-taker and the specificity of the test input. It means enabling the test-taker's strategic competence to engage a specific purpose discourse domain, set a communicative goal appropriate to it, and assemble relevant specific purpose background information and language knowledge to achieve the goal. Test content and test methods result from an analysis of a specific language situation and LSP tests follow the good testing practice of other types of language tests as they rely on contextualized communicative language ability (Douglas 2000: 1). These tests typically involve language for academic purposes and language for occupational/professional purposes (Douglas 2000: 2). Thus, learners'

language knowledge should be at an advanced level so that specific subject content knowledge can be built upon it. Therefore, the purpose is not to test subject-specific knowledge, but test-takers' abilities to manipulate language functions appropriately in different ways (Davies 2001: 143) or their language knowledge and their use of strategic competence in a target language use situation (Douglas 2000: 282; Bucur and Neagu 2015: 899). LSP tests usually measure proficiency, they are criterion-oriented and have a specific target group (Bucur and Neagu 2015: 899). In that sense, they follow practical needs and have a pragmatic effect. Needs analysis, curriculum content and test content are closely related in the successful practice of LSP (Mićić 2007: 309).

Why should we test LSP? Davies (2001: 133) justifies testing LSP by emphasising its connection with content more than with the language itself. He also insists on a practical need and a pragmatic effect of LSP tests as well as the fact that there are different varieties of each language (e.g. Medical English, Legal English, Business English, etc.) (Davies 2001: 134). Robinson (1989: 396) explains that ESP is a type of ELT that is goal-oriented and that students usually study ESP because they need to perform a task in English, not because they particularly like the English language itself. Douglas (2000: 2) describes LSP tests as necessary, reliable, valid and theoretically well-motivated. He prefers LSP tests over general ones for two reasons: (1) in order to represent a specific purpose field test tasks must be authentic; and (2) LSP is very precise owing to the technical language that is used in various fields and that is characterized by specificities that people who work in these particular fields should control (Douglas 2000: 7–8).

In accordance with this, LSP tests do differ from GE tests. According to Douglas (2000: 2), the basic difference lies in the fact that LSP tests are authentic and that there is an interaction between language knowledge and specific purpose content knowledge. He also emphasises background knowledge as the most important defining feature — in order to successfully cope with an LSP test, a test-taker needs to have a knowledge of a particular field, not just the knowledge of a language.

2.1 Multiple-choice tests

Multiple-choice format in testing is becoming more and more popular as a part of large-scale tests as it can be easily scored both by human assessors and machines (Fulcher and Davidson 2007: 27). A multiple-choice item consists of two parts: a *stem* (a statement or a question which is the top part of a multiple-choice item) and usually four *choices* (only one of these is correct and is usually called the *key*, whereas the other choices are incorrect and are called *distractors*). In Kehoe's (1995) guidelines for creating multiple-choice items it is stated that there should be only one problem in each stem and that a stem should be an incomplete statement or a direct question, with no stereotypes.

As writing such test items is rather difficult, it is necessary that each of them is carefully reviewed. It should be checked whether the item tests what it should test (*content review*), that distractors are by no means correct (*key check*), whether any of the items is subjective or may cause offence (*bias sensitivity check*), and if there are any spelling or other mistakes in any of the items (*editorial review*) (Fulcher and Davidson 2007: 118; Alderson et al. 1995: 63).

In designing multiple-choice tests, it is essential that all the correct answers must be truly correct and that each choice should fit equally well into the stem (Alderson et al. 1995: 47–51). It is also recommendable that multiple-choice items are presented in context, whenever possible.

3. EMP tests

In EMP testing, choosing appropriate topics can be more difficult than in general English testing. For example, a test of reading comprehension for 2nd year students could not be based on the textbook description of the anatomy of the heart since they would be able to give answers without reading the text. The text could not serve for testing even if it contained more content-based medical knowledge than the students' mother tongue knowledge. Nevertheless, EMP testing is based on the same principles as general English testing. Most teachers construct achievement tests, drawing their content from the course (Mićić 2007: 309). If the course content is appropriate, the test content is also adequate. What is also important is that there should be a difference between tests for medical students and tests for doctors. In the latter case, specific language skills needed in professional work should be tested, taking external criteria and the target situation into consideration.

There are four periods in the 30-year history of testing English at the Belgrade University Faculty of Medicine: 1987-2000, 2000-2004, 2004-2015, 2015. The first two cover the period before huge educational reforms (Bologna 2004).

The first phase was characterized by the grammar-translation method with the emphasis on translation from English into Serbian and vice versa. The concept applied was based on the belief that students would need primarily reading and comprehension of medical texts of the general type (the first two years of studies). The exam, numerically marked, consisted of a written and an oral component. In the written component, students were supposed to translate an unknown medical text from English into Serbian, as well as four sentences from Serbian into English (testing grammar structures). Students were allowed to use a dictionary. In the oral component, translation from English into Serbian was also tested by reading a paragraph from a familiar text. Eventually, it turned out that the expectations were unrealistic since students' knowledge of English varied and they struggled with translating from Serbian

into English. In order to translate well, it is not sufficient to possess the knowledge of technical terminology or to be familiar with grammatical categories. Providing one and only meaning of a single word (even the most frequent one) was proven to be one of the main obstacles for students to understand medical texts (Mićić 2009: 98).

The second phase was characterized by the modification of the first type of test. The first part of the test was based on grammar relevant for language use in medicine. It consisted of a cloze test, sentence reformulation, combining sentences, multiple choice, etc. The second part remained the same: translation from English into Serbian and from Serbian into English, but in a more concise format. The text was rather short and four sentences turned into a coherent text. The oral part was also changed: instead of reading and translating, we introduced discussions on relevant medical topics. However, the difficulties with translation remained.

In the third phase, after thorough reforms, the exam was transformed into a written test, descriptively marked. The grammar test remained the same, but the second part was transformed into a reading comprehension text. It turned out that it was more important for students to understand the text in English and that the translation was not a necessary proof of the ability to understand the texts.

The grammar section of English Language 1 test in this phase contained 55 points and tested several important items: articles, singular/plural of nouns, prefixes/suffixes, adjectives/adverbs, tenses. For this purpose, three objective-type items were used: gap-filling[4] (Example 1), multiple-choice items (Example 2) and short-answer questions.

Example 1:
Insert A/AN or THE where necessary:
_____ pineal gland is _____ tiny, cone-shaped structure within ____ brain. Its function is _____ secretion of the hormone _____ melatonin.

Example 2:
Provide the meaning of the **UNDERLINED** PREFIXES and SUFFIXES:
1. RENAL
a) rib b) kidney c) liver
2. NEURALGIA
a) algae b) against c) pain

[4] Gap-filling in testing refers to a short passage from which some words or phrases have been deleted and the test-taker is supposed to provide them, while not violating the grammar of the passage. In this way a test writer can choose various aspects of the language that they wish to test. Both authentic texts and specially written passages can be used to this purpose. There should be only one correct answer for each gap, but this is often very difficult to achieve (Alderson et al. 1995: 54-55).

The grammar section of English Language 2 was also worth 55 points and apart from the previously mentioned items, it tested several more: connectors, prepositions, modal verbs, participles/infinitives, indirect speech, if-clauses. Objective-type items used for testing were cloze test[5] (Example 3) and short-answer questions (Example 4).

Example 3:
The patient is a 32 - _____ - old man. When he visited his GP six months _____, he _____ (COMPLAIN) of headaches which _____ (TROUBLE) him for three months. On examination he _____ (FIND) to have a blood pressure of 180 _____ 120. Urinalysis was normal, as well as ECG and chest X –rays. He _____ (TAKE) a beta blocker ever since, but his blood pressure _____ (BE) still slightly elevated. Urinalysis now shows albumen, his haemoglobin is 12.9 and blood urea is greatly raised. For this reason, his doctor _____ (ARRANGE) an urgent admission _____ investigation and treatment _____ chronic renal _____.

Example 4:
Rewrite the sentences using the correct MODAL verb (can, could, must, should, might, may, mustn't...):

1. You were supposed to take the pills regularly, but you didn't.
.....
2. The patient has difficulty in breathing, so she is forbidden to smoke.
.....
3. He is able to hear well after the operation.
.....

Both in the English Language 1 and English Language 2 tests of this phase, the reading comprehension section was worth 45 points and it tested vocabulary and grammar in context. The objective-type item used for this purpose was a multiple-choice test containing nine items with four choices each. They were mostly affirmatively phrased, but there were false affirmatives and false negatives as well, which proved to be most difficult for students.

In the last phase, in accordance with modern trends, the test was entirely turned into a multiple-choice format. The English Language 1 test now consists of four parts: grammar section (20 points), language in use (20 points), a vocabulary section (30

[5] Cloze test usually consists of a passage in which every nth word is deleted and many authors believe this kind of a test measures overall language ability and language proficiency (Aitken 1977; Oller 1979; Madsen 1983: 47). This test is different from "fill in the blank" exercise which consists of isolated sentences as it usually uses a longer passage and is contextualised (Keshavarz and Selimi 2007: 82).

points) and reading comprehension (30 points). Items tested in the grammar section are those that were tested in the Grammar section in the previous phase, including quantifiers (Example 5). In the language in use section, tenses (active and passive) are tested (Example 6), whereas the vocabulary section tests specific medical terms and medical terms in context (Example 7). The reading comprehension section tests vocabulary and grammar in context and the objective-type items used for this purpose are dichotomous items[6] (Example 8) and circling the most precise answer (Example 9).

Example 5:

They did several laboratory _____ to prove that he hadn't made a mistake.

a) analysis b) analyses c) analizes

Example 6:

Mr. Johnson (1) _____ a retired bank clerk. He (2) _____ unwell and he (3) _____in bad with a cough for almost a week when he finally (4) _____ in his GP three days ago. On the occasion, a lower respiratory tract infection (5) _____ and the doctor (6) _____ erythromycin. Mr. Johnson (7) _____ the antibiotic since then, but he (8)_____ (FEEL) worse than three days ago. Tomorrow morning Mr. Johnson (9) _____ to hospital as his doctor (10) _____ to run some more tests.

1. a) was	b) is	c) is being
2. a) had been feeling	b) has been feeling	c) is feeling
3. a) has been	b) was	c) had been
4. a) was called	b) has called	c) called
5. a) was diagnosed	b) diagnosed	c) is diagnosed
6. a) prescribes	b) prescribed	c) has prescribed
7. a) is taking	b) had been taking	c) has been taking
8. a) is feeling	b) felt	c) was feeling
9. a) will admit	b) has been admitted	c) will be admitted
10. a) is wanting	b) wants	c) wanted

[6] These are True/False or Yes/No items. Alderson et al. (1995: 51) find them rather unsatisfactory because there is 50% possibility of getting the answer right simply by chance. However, they believe this possibility can be somewhat reduced by introducing the third category, such as Not Given or Does not say and they agree this type of test might by a good choice in testing reading comprehension. In our Reading comprehension sections, there are always three items - True/False/Not Given. In English language 1, students are familiar with the text the excerpt is taken from, whereas in English language 2 they get a passage they have never seen before.

Example 7:

The thyroid gland is one of the most significant of the (1) _____ glands. It lies in the front part of the throat, along the (2) _____, and weighs between 20 and 25 grams. The thyroid gland (3) _____ the thyroid hormones thyroxine and calcitonin which primarily influence the metabolic (4) _____. Iodine is essential for the production of the thyroid hormones. Iodine (5) _____, most common in inland and mountainous areas, can predispose to goitre (enlarged thyroid gland).

a) renal	b) endocrine	c) target
a) alveoli	b) lungs	c) windpipe
a) eliminates	b) secrets	c) secretes
a) rate	b) radius	c) ratio
a) proficiency	b) deficiency	c) deficit

Example 8:

Urinary tract infections are a significant cause of illness and a major factor in the development of chronic renal failure. Females are more susceptible to urinary tract infections then are males because the urethra is shorter in females and the urethral and anal openings are closer together. The incidence of infection increases directly with sexual activity and aging in both sexes. Some of the most common urinary tract infection symptoms are: severe pain, a frequent need to urinate, inability to delay urination, cloudy, bloody or smelly urine.

Urinary tract infections do not affect the development of chronic renal failure.

a) TRUE b) FALSE c) NOT GIVEN

Women suffer from urinary tract infections more often than men.

a) TRUE b) FALSE c) NOT GIVEN

Shorter urethra in women allows the bacteria to reach the bladder quickly.

a) TRUE b) FALSE c) NOT GIVEN

Elderly people are more prone to urinary tract infections than young people.

a) TRUE b) FALSE c) NOT GIVEN

Most common urinary tract infection symptoms are: frequent urination, mild pain, urgent need to urinate, changed urine.

a) TRUE b) FALSE c) NOT GIVEN

Example 9:

Smallpox is caused by the variola virus and is most often transmitted by inhaling the virus. It has an incubation period of between 7 and 17 days, after which symptoms begin to appear. The initial symptoms are flu-like. A significant feature of the disease is the development of blisters on the upper part of the body, which eventually scab over and leave scars when the scabs fall off. Around 30 per cent of those infected with smallpox die, usually within two weeks

of symptoms appearing. The first attempts to control the disease used a technique known as variolation. Dried scab tissue from victims of smallpox was used to deliberately infect young people. Of those infected by variolation, one per cent died, far fewer than the 30 per cent killed by infection in the normal way. Despite the risk, variolation was still used in some remote communities until relatively recently.

1. **The variola virus is:**
a) caused by smallpox. b) an airborne virus. c) a flu symptom.

2. Smallpox is characterized by scars which appear:
a) as a result of blisters that scab over.
b) in the incubation period.
c) after the patient has died.

3. Smallpox:
a) is always fatal. b) causes an instant death. c) may be fatal.

4. Variolation technique involved:
a)infecting people intentionally.
b) infecting people accidentally.
c) analyzing dried scab tissue.

5. The number of people who died from variolation:
a)equals the number of people who died from smallpox.
b) was the highest in distant communities.
c) was significantly lower than the number of people who died from smallpox.

The English Language 2 test now consists of three parts: the grammar section (40 points), vocabulary section (30 points) and reading comprehension (30 points). The same objective-type items are used to test all the items in these sections. The only difference refers to the reading comprehension section, where students are not familiar with the reading passages. However, these passages always belong to one of the fields that have been done in class and they mostly contain the vocabulary we discussed in detail so it cannot be said they are completely new to the students.

4. Research

For the purpose of checking the impact of the new format test on students' success in final exams we compared the old format test (phase 3) and the multiple-choice test (current phase) in students' exam papers from two different academic years.

First of all, final exam results from June 2013 (when the old format test was used) and June 2018 (when the new multiple-choice format was applied) for English Language 1 and English Language 2 were compared for general insight into students' success in these exams. Comparing the number of students who took the exams, it can be noticed that in June 2013 a greater number of students took the exam than in June 2018: in English Language 1 there were 528 students in June 2013 and only 228 in June 2018, whereas in English Language 2 there were 487 students who took the exam in June 2013 and only 209 students who did so in June 2018. The fact that a significantly larger number of students took the exams in June 2013 is related to certain organisational problems at the Faculty of Medicine, where it is sometimes possible to have several exams scheduled at exactly the same time. Therefore, we do not think this fact is important for our research, as it is by no means a result of introducing a new test format.

In the case of the English Language 1 test, the percentage of students who failed the exam showed only a slight difference between the two tests: 9.5% in June 2013 and 9% in June 2018. However, the percentage of students who failed the English language 2 exam revealed a more significant difference, as there were 14% of those who failed the old format test and only 6% of those who did not manage to pass the multiple-choice test. The results are presented in Table 1.

	June 2013 English language 1	**June 2018 English language 1**	**June 2013 English language 2**	**June 2018 English language 2**
no. of students who took the exam	528	228	487	236
no. of students who passed the exam	478	209	419	223
no. of students who failed the exam	50 (9.5%)	19 (9%)	68 (14%)	13 (6%)

Table 1. Students' overall success in the exam

So, students' overall success in the final exam is slightly better in the new format. The reason for a significantly reduced percentage of students who failed the English Language 2 exam after introducing the multiple-choice format probably lies in the objective-type items used in grammar section of the old format test. Cloze test and short-answer questions were too difficult for students as they were supposed to

provide all the answers for themselves. There were no items to choose from and they had to write all the answers, so they could not rely on recognizing the correct answer. Spelling also often got in their way of providing the answers in these sections of the test and they lost some important points.

Test results from June 2013 and June 2018 were also compared in relation to particular parts of the exam and their complexity. To this purpose a corpus of 200 test papers was used: 100 papers from June 2013 (50 English Language 1 papers and 50 English Language 2 papers) and 100 papers from June 2018 (50 English Language 1 papers and 50 English Language 2 papers).

When it comes to the grammar section (grammar and vocabulary) of the English Language 1 test, students scored slightly better in the multiple-choice format, but the difference is not particularly significant: they had 79% of correct answers in this section in the multiple-choice format compared to 73% in the old format test. However, comparing the results in reading comprehension sections of the two formats, a significant drop can be noticed in the number of points the students achieved in the multiple-choice test. Surprisingly enough, they scored much better in the old format test gaining 81% of correct answers compared to only 66% in the new format test.

Inspecting English Language 2 test results we got slightly different results. In relation to the grammar section (grammar and vocabulary), students scored significantly better in the multiple-choice test – they got 80% of correct answers in comparison with only 55% in the old test format. However, when it comes to the reading comprehension section, students remained more successful in the old test format although the difference was not as conspicuous as in English Language 1: they had 87% of correct answers in the old format test compared to 74% in the new format. The results are presented in Tables 2 and 3, respectively.

ENGLISH LANGUAGE 1	**Grammar section (grammar + vocabulary)**	**Reading comprehension**
JUNE 2013	36.6/55 73%	36.5/45 81%
JUNE 2018	55.2/70 79%	19.8/30 66%

Table 2. English language 1 test parts and results

ENGLISH LANGUAGE 2	Grammar section (grammar + vocabulary)	Reading comprehension
JUNE 2013	30/55 55%	39/45 87%
JUNE 2018	56/70 80%	22/30 74%

Table 3. English Language 2 test parts and results

Students obviously score better in the grammar and vocabulary sections of the multiple-choice format than in the grammar section of the old format, and this is especially true of the English Language 2 test. We have already mentioned that objective-type items used in the old format test were much more demanding for students than multiple-choice items, especially in English Language 2. Besides, grammar tested in English Language 2 is more demanding than the items tested in English Language 1; it includes reported speech, IF-clauses, modal verbs and differences between participles and infinitives. Although these same items are tested in the new format as well, it is expected that students will score better in the multiple-choice format as it is easier for them to "recognize" the correct answer than to provide it on their own.

Surprisingly, however, students' scores in the reading comprehension section of the new format, both in English Language 1 and English Language 2, are lower than in the old format. A possible reason for this may be the introduction of dichotomous answers (True/False/Not Given) to the first part of the reading comprehension section. We have noticed that students often have a problem with differentiating between False and Not Given — they rarely opt for Not Given as they believe that something should not be mentioned in the passage at all in order to choose the Not Given option. However, this option is used for pieces of information that are mentioned in the text, but not with all the details. This part of the reading comprehension section contains five questions and is worth 15 points, but quite often students have a very low score in this section or they even skip it entirely and earn zero points. The second part of the new format reading comprehension is similar to the old format, but it is less complex, so we do not believe that is the reason for a drop in the number of points students earn in the reading comprehension section.

From a teacher's perspective, the new format is more convenient for evaluation, but there is a dilemma concerning the real scope of students' knowledge when tested with multiple-choice tests. We need to ask ourselves how much can be attributed to chance alone when doing a multiple-choice test and how much real knowledge students reproduce during the test. In other words, is it possible that students' knowledge

remains exclusively passive when tested this way? We believe that students are able to acquire certain aspects of the language equally successfully in both formats. In both test types the accent is put on reading and writing skills, as speaking and listening skills are not tested directly.

From the students' perspective, the new test format is generally easier and more practical. They feel more secure as they do not need to spell words for themselves or be absolutely sure about any answer — it is "enough" to look at the given options and choose the correct answer. However, they find it difficult to concentrate on completing the multiple-choice test appropriately — they tend to forget they are not supposed to circle more than one option, they make corrections in their final answers (although they are told not to do so) or they choose several wrong options in a line as they were not concentrated enough.

Generally speaking, the new format is more convenient for both teachers and students, but it is necessary to inspect if (and how much) it affects the overall level of knowledge students achieve in these courses and if it means they will be less precise in using the language.

6. Conclusion

In conclusion, LSP needs to be tested and it should be done in a different way from GE. LSP is specific insofar as it is usually authentic; it is much more precise than GE; it is goal-oriented, and it relies on background knowledge. Specific test situations are used for deriving test content and test methods. Subject-specific knowledge is not what matters in this field, but rather the learners' ability to manipulate language functions in different situations and in various ways.

At the Faculty of Medicine, University of Belgrade, English for Medical (Academic) Purposes has been tested for three decades and the test has undergone considerable changes over time. From the grammar-translation method, through a combined written and oral test and a written test that contained several objective-type items, the current test has the form of a multiple-choice test.

The old format test and the multiple-choice test were compared in order to check the impact of the changed test format on students' overall success in the final exams of English Language 1 and English Language 2. We also looked into the particular parts of the exam and their complexity and we came to several conclusions: students' success in the final exam is slightly better in the new format; both test formats test the same language items, but the tasks in the Grammar section of the old test format proved to be more difficult than in the multiple-choice format, owing to objective-type items used in the old test; students' scores in the reading comprehension section of the new format are lower than in the old format probably due to the introduction of dichotomous answers;

the multiple-choice test format is more convenient for both teachers and students; it is not clear if we can be sure about the real scope of knowledge when tested this way.

A future study could be dedicated to improving the current multiple-choice format and making it more reliable and more informative concerning students' proficiency and abilities at the end of the course. We would like to revise the current form of test to a degree, by changing the reading comprehension section. Instead of dichotomous answers we could opt for circling the most precise answer, offering two shorter reading comprehension texts. This would make it easier for students who find dichotomous answers confusing and it would probably improve their overall score.

References

Aitken, K. G. (1977). Using Cloze Procedure on an Overall Language Proficiency Test. *TESOL Quarterly*, 11(1), 59–67.

Alderson, J. C. et al. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F. and A. S. Palmer (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bucur, N. F. and C. Neagu (2015). The Limits of ESP Tests. *Challenges of the Knowledge Society*, 5(1), 898–901.

Carroll, B. J. (1980). Specifications for an English language testing service. In: Alderson, J. C. and A. Hughes (eds.), *Issues in Language Testing. ELT Documents 111*, London: British Council, 66–110.

Cronbach, L. J. and P. E. Meehl (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Davies, A. (1967). The English proficiency of overseas students. *British Journal of Educational Psychology*, 37(2), 107–122.

Davies, A. (2001). The Logic of Testing Languages for Specific Purposes. *Language Testing*, 18(2), 133–147.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Fulcher, G. and F. Davidson (2007). *Language Testing and Assessment*. Abingdon: Routledge.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, 18, 519–521.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, MA: Newbury House.

Hughes, A. (1989) *Testing for Language Teachers.* 1st edition. Cambridge: Cambridge University Press.

Kehoe, J. (1995). Writing multiple-choice test items. *Practical Assessment, Research and Evaluation*, 4, 9. (4 January 2006) <http://PAREonline.net/getvn.asp?v=4andn=9>.

Keshavarz, M. H. and H. Selimi (2007). Collocational Competence and Cloze Test Performance: A Study of Iranian EFL Learners. *International Journal of Applied Linguistics*, 17(1), 81–92.

Madsen, H. S. (1983). *Techniques in Testing.* New York: Oxford University Press.

Messick, S. (1989). Validity. In: Linn, R. L. (ed.), *Educational Measurement*, New York: Macmillan/American Council on Education, 13–103.

Mićić, S. (2007). Novine u testiranju engleskog jezika za studente medicine. *Primenjena lingvistika,* 8*,* 307–316.

Mićić, S. (2009). *Studije o jeziku medicine u engleskom i srpskom*. Beograd: Beogradska knjiga.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In: Brumfit, C. K. and K. Johnson (eds.), *The Communicative Approach to Language Teaching*, Oxford: Oxford University Press, 143–159.

Oller, J. W. (1979). *Language Tests at School.* London: Longman.

Rea-Dickens, P. (1987). Testing Doctors' Written Communicative Competence: An Experimental Technique in English for Specialist Purposes. *Quantitative Linguistics*, 34, 185–218.

Robinson, P. C. (1989). An overview of English for Specific Purposes. In: Coleman, H. (ed.), *Working with language: a multidisciplinary consideration of language use in work contexts*, Berlin: Mouton de Gruyter, 395–427.