

HOW TESTING ONLINE MAKES EDUCATION MODERN AND EFFECTIVE

Lucie Němečková

Institute of Lifelong Learning, University of Žilina, Slovakia

Resumé: Cieľom autorky článku je poukázať na to, že modernizácia v testovaní vysokoškolských študentov zefektívňuje vyučovanie a pomáha učiteľom zlepšiť ho a urobiť pre študentov zaujímavejší. Sekcia cudzích jazykov Žilinskej univerzity nedávno zmenila spôsob testovania a hodnotenia študentov s dôrazom na položkovú analýzu a diagnostiku ľahkých, optimálnych a ťažkých položiek testu. Podstatná časť výskumu je venovaná výsledkom jednotlivých študentov, nazvaným „študentský zisk“ odrážajúci vedomosti každého študenta, ktoré získal počas semestra.

Kľúčové slová: testovanie online, položková analýza, študentov zisk, spätná väzba

Abstract: The main aim of this paper is to prove that the modernization of testing university students makes the learning process more effective and helps teachers to improve and make this process more interesting for students. The Department of Foreign Languages from University of Žilina has recently changed its way of testing and evaluating students capitalizing on the impact on test item analysis and diagnosing easy, optimal and difficult tasks. The most significant part of this piece of research deals with a marker of individual students' achievement called "student gain" reflecting the knowledge students gain during a semester.

Keywords: testing online, item analysis, student gain, feedback

Introduction

Language education is provided by Department of Foreign Languages at certain faculties of University of Žilina. Both the department and faculty representatives aim to put the effectiveness of language learning at the center of attention. Since the university faculties are technically oriented, the main role of language lessons is to teach students technical terms and how to deal with various technical topics and case studies. They will need it for communication, presentation, defining and solving problems during internships, Erasmus exchange trips and primarily in future employment as well as to gain professional experience. The need of professional vocabulary, fluent and correct spoken language is important and because of that enhancing and reinforcing technical language learning requires implementing modern and

effective strategies and preparing for lessons effectively. We must be highly precise when preparing study materials, choosing appropriate teaching methods and last but not least preparing tests, which we consider as one of the main indicators of successful education process.

Tests reflect relevant information about the best student results possible and that is why we are considering making the assessment of students a crucial tool in the modernization and effectiveness of education. The information is gained by the research part of which is run through statistical data processing. Not only teachers but also students benefit from this research based on recording, monitoring, gathering and interpreting data. Our research is based on online testing.

Online testing

Our department decided to replace written paper-based tests with more modern online testing. There are many websites, applications and different types of software offering the e-learning possibility and allowing for the creation of author tests or quizzes. The university has predefined Moodle but, with students, we think it is not easily used because of its extensive option range, continued updates and changing environment. We tried to find a better solution by working with Hot Potatoes, Kahoot and Jeopardy, but the last choice looked like the most suitable even for our students.

We started using an online free (or paid platform), a website called Socrative. Recently, it topped teacher rankings as their favorite website. According to us, this tool is more user-friendly and easier to use. You can create, modify or delete the test any time or share it with your colleagues by signing in with your email and password. It offers quizzes to check whether students have learned vocabulary or what they think or know about the topic, offering them the space to present their ideas. It all happens in a virtual room to which students must first be allowed access by a teacher, who also launches test for them where necessary. The only limitation is that just three types of tasks are on offer – multiple choice, true/false and short answer, which other teachers may consider insufficient for various types of tasks offered in a paper-based test. Reading and listening tasks, pictures and videos, abbreviations and matching: all of these can be created and embedded in this online test platform. One only needs to adapt them to the three types of tasks mentioned.

The crucial advantage of Socrative is that it supports formative assessment and boasts two strands: immediate feedback for students and our research and further motivation to learning.

“Formative assessment refers to frequent, interactive assessments of student progress and understanding to identify learning needs and adjust teaching appropriately. Teachers using formative assessment approaches and techniques are better prepared to meet diverse students’ needs – through differentiation and adaptation of teaching to raise levels of student achievement and to achieve a greater equity of student outcomes” (CERI, 2008: 1).

Using Socrative online testing, we can prove that we are implementing formative assessment features to our language lessons, which is a sign of modern trends in recent years. This program helps us to assess student progress immediately, develop cooperation to improve student results and point out the needs of students and establish the strong and weak points of our education process, which we can then adjust.

Creating an appropriate test is a long process. If you draft a test well, it will help the education process because students will understand their tasks and identify what is needed to achieve these. On the other hand, if a test is drafted in a wrong way, it will affect students’ results and you will still have to adapt tests to students’ needs. You can consider everything closely by checking the item analysis.

Item analysis

Item analysis is based on examination and analysis of individual tasks in a test, the relevant success rate and its influence on overall results. It entails assessment of how a test is done. By item we understand each separate task within the test in accordance with the test specifications of our department. We collected the data for item analysis from all students of a specific field of study (in our case, informatics) attending foreign language lessons and assessed them according to their achievement in two online tests – *a mid-term test* in the middle and *a final test* at the end of the winter term of 2018/2019 (both tests have a predefined number of forty items).

This kind of analysis is dually-oriented. It has a qualitative (reliability of the test, comprehensiveness and clarity of given tasks and grammatical and stylistic mistakes) and a quantitative aspect (validity of the test and difficulty of tasks). Thanks to item analysis, we can compare the results and monitor student progress in two language skills such as reading with

comprehension and listening with comprehension. The test has parts focusing on lexis and grammar, too. Speaking is involved in almost 70 % of during lessons, so teachers can evaluate oral performance continuously and writing (as homework) is included in the final evaluation, too.

Items which have the form of multiple choice questions, matching or pictures are the easiest ones and even students with low language levels are able to do them. It is proved by almost all results of 100 % shown in Fig. 1. Contrary to the easy tasks mentioned, there are the hard ones such as definitions and grammar tasks, which is also clear from the success rate, ranging between 20 and 50 %.

Item type	Def	Def	Gram	Gram	Multi	Multi	Match	Match	Pict	Pict
Group 1	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19
Stud 1	1	1	1	1	1	1	1	1	1	1
Stud 2	0	1	1	1	1	1	1	1	1	1
Stud 3	0	0	0	0	1	1	1	1	1	1
Stud 4	0	0	0	0	1	1	1	1	1	1
Stud 5	0	1	0	0	1	0	1	1	1	1
Stud 6	0	0	1	0	1	1	1	1	1	1
Stud 7	1	0	0	0	1	1	1	1	1	1
Stud 8	0	0	0	0	1	1	1	1	1	1
Stud 9	0	1	0	0	1	1	1	1	0	1
Stud 10	0	1	0	0	1	1	1	1	1	1
Class scoring	20%	50%	30%	20%	100%	90%	100%	100%	90%	100%
Difficulty index of items	0,2	0,5	0,3	0,2	1	0,9	1	1	0,9	1

Fig. 1 Comparison of different types of tasks

Reading and listening are in most cases divided between the two tests, e.g. *the mid-term test* includes reading and *the final test* includes listening or vice versa. In very few cases does choice depend on which activity is practiced more during the lessons because only some topics are more speaking-oriented and some listening-oriented. For our teaching response, education process effectiveness and further student motivation, the most important point is to define the difficulty of a test and its items.

Difficulty index

Difficulty is one of the main variables in items which allow us to prove that the test itself was too easy or vice versa. Each item is shown as a percentage result and the difficulty is assessed by means of a difficulty index. The difficulty index is a ratio of the total number of correct answers and the total number of test participants (Markechová, 2011: 235). The difficulty index differentiates allows for a distinction between the difficulty of individual test items and the overall difficulty of the test. To arrive at a final difficulty value range, it is necessary to define intervals for these. The value range belongs to the closed interval $<0;1>$ representing 0 % to 100 % result. Since university belongs to the tertiary level of education and the percentage for passing both optional and compulsory lessons stands at 61%, we adopted the range of items as easy, optimal and hard and their difficulty indexes vary in intervals as follows:

- easy item – difficulty index has the interval $(0,9; 1>$
- optimal item – difficulty index has the interval $<0,7; 0,9>$
- hard item – characterized by difficulty index from interval $<0; 0,7>$.

Figure 2. shows the final value range from the test done by only ten first-year students of Informatics. The forty-item test was concentrated on technical terms and collocations, grammar and reading comprehension. In accordance with test specifications, we selected the following types of tasks:

- for technical terms and collocations – definition, definition with first letters given and definition with translation
- for grammar – word formation and
- for reading comprehension – summary and filling in the correct terms.

Our difficulty was compounded by the need to prepare the test precisely because technical terms had not to occur in test twice and we had to avoid repeating them. If it was the answer for one item it was not allowed to be used again as a word in another item, description or reading.

Group 1	Total score	Correct answers	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Stud 1	95	38	1	1	1	1	1	1	0	1	1
Stud 2	85	34	1	1	1	1	1	1	0	0	1
Stud 3	60	24	1	1	1	0	0	0	0	0	0
Stud 4	63	25	1	1	1	1	0	0	0	1	1
Stud 5	80	32	1	1	1	0	1	1	0	1	1
Stud 6	60	24	1	1	1	1	0	0	0	1	0
Stud 7	68	27	1	1	1	0	0	0	0	1	0

Stud 8	50	20	1	0	1	0	0	0	0	1	1
Stud 9	85	34	1	1	1	1	1	1	0	0	1
Stud 10	80	32	1	1	1	1	1	1	0	1	1
Class scoring	73	29	100%	90%	100%	60%	50%	50%	0%	70%	70%
Difficulty index of items			1	0,9	1	0,6	0,5	0,5	0	0,7	0,7

Fig. 2 Difficulty index of items 1-9 with the impact on technical terms

The first 10 tasks were centered on knowing technical terms, which were also the tasks in whose results we were the most interested because we wanted to find out if students really knew the terminology, especially to define it. The difficulty index values of the first three items were in the interval of values (0,9; 1), meaning that those items were too easy. Their aim was to motivate students with a lower language level. The subsequent ones had various values, which indicates divergence in preparing for the test and the knowledge of every student. Student 1 had the highest achievement but for Student 3 we had to analyze weaknesses and what said student had to study further. The tasks chosen were definition tasks, which also reflected that we will have to practice these more with students in this study group. Items 7-9 have the index 0 and 0,7, automatically showing they are hard and students had problems with them. Task 7 had the worst result, which led us to examine where the problem was – if students just did not know the answer because they had not studied or if it was the result of a some teaching mistake. At the end, after the final comparison of results in all groups, we came to the conclusion that the problem was in the item itself (we had chosen a definition which did not explain the term exactly). On the other hand, we can say, this part of the test was balanced with similar number of easy, optimal and hard items.

In the last part of the test (items 30-40) shown in Fig.3 it is obvious that items 30 and 31 were very hard for students because they had an index of 0,2 and 0,3.

Group 1	Item 30	Item 31	Item 32	Item 33	Item 34	Item 35	Item 36	Item 37	Item 38	Item 39	Item 40
Stud 1	1	1	1	0	1	1	1	1	1	1	1
Stud 2	0	1	1	0	1	1	1	1	1	1	1
Stud 3	0	0	1	0	0	1	1	1	1	1	0
Stud 4	0	0	1	0	0	1	1	1	1	1	1
Stud 5	0	0	1	0	1	1	1	1	1	1	1
Stud 6	0	0	1	0	1	1	1	1	1	1	1
Stud 7	0	0	1	0	1	1	1	1	1	1	1
Stud 8	0	0	1	0	0	0	1	0	1	0	1
Stud 9	1	1	1	0	1	1	1	1	1	1	1
Stud 10	0	0	1	0	0	1	1	1	1	1	1

Class scoring	20%	30%	100%	0%	60%	90%	100%	90%	100%	90%	90%
Difficulty index of items	0,2	0,3	1	0	0,6	0,9	1	0,9	1	0,9	0,9

Fig. 3 Difficulty index of items 30-40

After reviewing the results of all study groups we did a comprehensive item analysis using a difficulty index and compared the results of all study groups and teachers. We found out that only the group discussed above had such a low success rate in those two items. It led us to examine individual teacher methods and the result was that the teacher of this study group had not completed one topic. That was why students had not chosen the correct answers. Items 35-40 were reading exercises with an approximate difficulty index of 0,9, which indicated reading is a very easy task and we should change distractors before the next use. Remarkable results were found in items 32 and 33, which had opposite values of their respective percentages and indexes.

The overall difficulty level of test is the next part of our statistics affecting the effectiveness of the education process. With the purpose of measuring the overall difficulty of the whole test, it was necessary to calculate the ratio of the arithmetic mean quotient of all points of the test-takers and the total number of points which students could gain (Fig. 4). As an example, let n be the number of correct answers of Student 1, m – the number for Student 2 and 40 items for the test, then the overall difficulty level of test (**ODT**) is calculated as follows:

$$ODT = \frac{n+m}{40} .$$

Group 1	Total score	Correct answers	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Stud 1	95	38	1	1	1	1	1	1	0	1	1
Stud 2	85	34	1	1	1	1	1	1	0	0	1
Stud 3	60	24	1	1	1	0	0	0	0	0	0
Stud 4	63	25	1	1	1	1	0	0	0	1	1
Stud 5	80	32	1	1	1	0	1	1	0	1	1
Stud 6	60	24	1	1	1	1	0	0	0	1	0
Stud 7	68	27	1	1	1	0	0	0	0	1	0
Stud 8	50	20	1	0	1	0	0	0	0	1	1
Stud 9	85	34	1	1	1	1	1	1	0	0	1
Stud 10	80	32	1	1	1	1	1	1	0	1	1
Class scoring	73	29	100%	90%	100%	60%	50%	50%	0%	70%	70%
Difficulty index of items			1	0,9	1	0,6	0,5	0,5	0	0,7	0,7

ODT index	0,7	$(290/10)/40=$	
		0,7	
		290	a total of correct answers by S1-S10
		10	a number of respondents
		40	a number of items in test

Fig. 4 Calculating the ODT index

The final figures concern the index of test difficulty, which also belongs to the closed interval $<0;1>$. Overall test difficulty has the same intervals as those employed in the item analysis:

- easy test – ODT index has the interval $(0,9; 1>$
- optimal test – ODT index has the interval $<0,7; 0,9>$
- hard test – characterized by ODT index from interval $<0; 0,7)$.

The results show that the ODT index is definitely on the threshold of optimal difficulty but it is worth pointing out that ODT is also based on individual items. After the finishing the final detailed percentage analysis of both tests, we can univocally state that the results of students were significantly better at their final test and their overall results improved during the semester.

Likewise, in order to gain the most relevant and accurate result, it is necessary to test as many from the analyzed study group respondents as possible and these should have different language levels because, for every study group, the test and its items may vary in difficulty. Data that are correctly collected and evaluated can inform us if we have heterogeneous or homogeneous groups of students from the point of their language level. In the future, we hope to consider results in detail in order to distinguish between various types of groups. Our second aim is to understand and study in detail “student gain” which is the individual gain or achievement of every student during the semester. We hope to do this by implementing *pre-tests* at the start of every language course. It will show us what students know when they start attending a course and how much they have achieved.

Conclusion

Based on proven facts of item analysis and difficulty index data, we can affirm that this kind of statistical data collection and processing helps teachers greatly but that it is also useful for students. Although the preparation of test items, tests themselves and their analysis with statistics often constitute difficult and time-consuming processes, they result in giving students the opportunity to see the dynamic of their results in real time. The modernization of education

processes lies primarily in using IT during lessons as an assistive tool towards gaining and practicing new knowledge but our team has evidence that its substantial role is to function as a performance enhancer to improve student results and motivate students. Another sign that online testing really makes education modern and effective is the instant feedback for both students and teachers the faculty at large, too. We have also stepped up formative assessment tools by offering feedback and self-assessment options. Feedback relates firstly to the level of individual students' knowledge, secondly to the results of the group tested, thirdly an item analysis of a test and lastly to the difficulty of said test itself. Students have a positive stance towards online testing, especially towards instant feedback and the much faster results than when they sit paper-based tests and have to consider explanations which can be included in the test. Even though the drafting of tests takes longer because of the standards set for test items, which must meet certain criteria, this time is then spared by the immediate response of the system. Item analysis and the difficulty index are also two crucial tools for our team allowing us to enhance the quality and effectiveness of education. The aim of our research and this paper was to outline an initiative relevant to the assessment of students, language teaching and education.

Bibliography

ERI 2008: OECD/CERI International Conference „Learning in the 21st Century: Research, Innovation and Policy“ Assessment for Learning. Formative Assessment. Paris, 2008. available online: <<https://www.oecd.org/site/educeri21st/40600533.pdf>>

Markechová 2011: Markechová, G., Tirpáková, A., Stehlíková, B. Základy štatistiky pre pedagógov. Nitra: Univerzita Konštantína Filozofa v Nitre, 2011.