FROM ITEM DEVELOPMENT TO ANALYSIS: THE DEVELOPMENT OF A COMPUTERISED ENTRANCE PLACEMENT TEST USING A VLE

doi.org/10.61425/wplp.2020.15sp.43.69

Zsuzsanna Soproni
International Business School, Budapest
zssoproni@ibs-b.hu

Abstract: This paper reports using a virtual learning environment (VLE) for English language reading and writing competence testing as part of a higher education entrance placement test for a business college. The suitability of the VLE for computer-based testing, as well as the effect of its mode of delivery, including the perceptions of test-takers are discussed through the findings of quantitative and qualitative action research investigations. More specifically, within suitability, the transferability of specific tasks, security issues, user-friendliness and data processing are examined. It was found that prior computer-based test-taking experience is not an unfair advantage for applicants, but sample tests need to be made accessible to cater for the time needs of applicants with no prior computer-based experience. The VLE has proved to be a suitable platform for the purpose, in terms of test design, security, transferability and data processing, but only with the help of database engineers was some of the data retrievable. Thus, the paper seeks to emphasise the need for collaboration between information technology personnel and language testing professionals.

Keywords: computers, higher education, language testing, technology, Virtual Learning Environment

1 Introduction

The use of technological devices in teaching and the global English language teaching industry had become a norm in many parts of the world even before the Covid-19 pandemic hit the Globe. New educational platforms were born, tested and promoted in modern educational institutions. Over recent decades, the 21st century student has become accustomed to classrooms with computers and projectors, colleges and libraries with Wi-Fi access, teachers using mp3 files, presentations software and perhaps some online quiz platforms such as Kahoot, Slido or Mentimeter. The goal set by Bax (2003, p. 11) for technology to become "invisible, embedded in everyday practice and hence 'normalised'" in language education has been achieved, at least in the developed world. In 2007, 74% of teachers in 27 European schools reported that they had used information and communication technology (ICT) in class in the previous year (Korte & Hüsing, 2006) and the proportion has probably increased since then. Beyond the school perimeter the Internet has greatly enhanced the learning process with an abundance of information, with opportunities to build vocabulary or to practise grammar. Massive Open Online Courses are available to those wishing to study on their own. With regard to English language coursebooks, they are published with a view to allowing access to practice websites and virtual classroom forums in lieu of workbooks, which could be used either in or outside the classroom, and not only for doing traditional exercises, but also for posting comments or submitting and marking written homework. The introduction of Web 2.0 tools, interactive whiteboards and social networks have transformed computer-assisted language learning (CALL) (Dudeney & Hockly, 2012). Some of the applications have become available on

smartphones, which are used to motivate students and even for modernising testing in higher education (e.g., Solihati & Mulyono, 2018).

The pandemic, however, made technology-enhanced-learning opportunities the one and only viable alternative for face-to-face teaching almost overnight in the spring of 2020. The only way students and teachers were able to continue teaching and learning was through online means. Internet-based classrooms and other virtual solutions, video conferencing applications such as Microsoft Teams, Skype and Zoom, suddenly became the only way students and teachers could interact in a contactless manner. Soon, in Hungary, at-home (online or internet-based) testing, with previously unprecedented speed in legislation, was also introduced on the basis of Article 8(3) of the Government decree No. 101/2020. of 10 April (101/2020. (IV. 10.) Kormányrendelet, 2020). As a result of safety concerns and measures, test centres worldwide had to close, so international exam providers too developed their own online tests, for example, TOEFL introduced their special home test which was "available around the clock every Tuesday, Wednesday, Thursday and Friday" (TOEFL iBT® Special Home Edition, n.d.).

The use of 21st century educational technology has been and still is promoted in the International Business School in Budapest (IBS), both before and amid the coronavirus pandemic. On its website, IBS promotes itself as a modern, innovative and dynamic business school (The IBS Story, n.d.). In addition, the student body is highly capable of using digital devices not only because of their age and socio-economic background but because more than half of the students are foreigners, who invariably rely on technology to maintain family ties and, sometimes, to run a remotely located business. As part of the innovative teaching and learning methods and good practices, it is thus not a surprise that IBS sought to computerise its English language entrance placement test.

This paper aims to present the computerisation project and summarise the lessons learnt from the process. The action-oriented research approach was followed and the experience of using a virtual learning environment will be summarised with reference to quantitative and qualitative data collected in the process (Cohen et al., 2007).

2 The pros and cons of computer-based testing

2.1 Institutional advantages

In this section, the benefits of using computers for language testing purposes will be discussed. First, innovations to improve efficiency in language testing will be briefly discussed, followed by a discussion of advantages for analysts. The section will end with the discussion of the appealing nature of computerised tests to young test-takers.

2.1.1. Efficiency

It is without doubt that by the end of the 20th century, computers have occupied "a major role in test construction, item banking, test administration, scoring, data analysis, report generating, research, and the dissemination of research" (Fulcher, 2000, p. 93). However, efficiency has been an issue in measurement for a hundred years now whenever a high number of candidates were tested. As a milestone, the intelligence of large numbers of potential soldiers was tested with the help of perforated cardboard or transparent celluloid masks in the US to speed up the marking (Yoakum & Yerkes, 1920, p. 159). Scorability appeared to be a key

advantage (Fulcher, 2000) but since then many different methods have been devised to assist test-developers, markers and analysts. Moreover, test-takers themselves have come into closer contact with computers. Thus, the term "computer-assisted language testing" emerged to refer to "tests that are administered […] on personal computers" (Brown, 1997, p. 37). A further step in the development was the introduction of web-based tests, which were no longer administered on a closed network. Within web-based tests, Roever differentiates between low-tech and high tech web-based solutions depending on the availability of funds, expertise and the complexity of the systems and algorithms behind (2001, p. 85).

As usual, global language exam providers have been leading the innovation. The Educational Testing Service (ETS) introduced the computer-based TOEFL test and its internet-based version in 1998 and in 2005 respectively (Educational Testing Service, 2007). The Cambridge Preliminary English Test was launched in a computerised format in 2005 (Jordan et al., 2011). In 2016, International English Language Testing System (IELTS) introduced its computer-based English language exam (Computer-delivered IELTS, n.d.). Likewise, in 2016, the Business Language Testing Service (BULATS) began offering an exclusively online business language exam (BULATS: About the test, n.d.). The technological development in education has infiltrated assessment (Hill & Barber, 2014; Suvorov & Hegelheimer, 2014).

In addition to computer-based tests, computer-adaptive testing was also made possible by the availability of large item banks, IRT-based analysis and the monitoring of test-taker responses during the test. As a result, tests may become shorter and more efficient and can be administered with additional security since no two tests are identical in terms of individual items (Fulcher, 2000). The first computer-adaptive tests were launched by Cambridge ESOL and were available on CD ROMs in 1990s (Jordan et al., 2011). TOEFL introduced computer-adaptive Listening and Structure tests in their computer-based testing in 1998 (Educational testing Service, 2007).

Thus, computerisation brings with itself lower costs, higher efficiency and easy scoring. The general public is prone to believe that all is needed for a computer-based test "is a computer with a web browser and an internet connection" (Roever, 2001, p. 88). The efforts of professionals are often taken for granted and the ensuing costs are often underestimated (Fulcher, 2000; Pathan, 2012). Little is known by the test-taking population or the users of test results, about test-specifications, statistical analyses or pre-calibrated items. To do justice to all these beliefs, using an item-bank can save money and time but establishing an item bank requires substantial funds and expertise.

2.1.2. Data collection and analysis

When paper-based and computer-based tests are compared, a salient advantage of computer-based tests for measurement specialists is that data are readily available for analysis immediately after an exam is taken once the test-takers have entered the answers themselves. Computer-based tests can be automatically and time-efficiently scored (Mohammadi & Barzgaran, 2010), especially the sections that include unambiguous short answers, that is, discrete-point items, although recent developments indicate that even constructed responses can be reliably rated by computers (Gomaa & Fahmy, 2011). Analysis is enhanced by the data being available instantly or can even be done automatically. Computerisation also allows researchers to access data they were previously unable to collect, for example they can study underlying cognitive processes by analysing responses that test-takers decide to modify or the way they edit their writing. Test-taker behaviour, "their route through the test" (Pathan, 2012, p. 35),

more specifically, their eye movements may be studied as well (e.g., Brunfaut & McCray, 2015). Another advantage that markers often cite is the elimination of the onerous task to decipher students' handwriting.

2.1.3 Customer expectations

Encountering state-of-the-art technology during the application process to a university might be more attractive to young applicants for whom one of the most important objects in their lives is typically their mobile phone or their laptop. Even in 2001, a study found that more young people preferred taking an experimental computer-based test than a paper-based one (44% vs. 35%) and "the majority of the students (59%) believed taking the test on the computer was less tiring than taking an equivalent paper-and-pencil test" in an American high school research context (Bridgeman et al., 2001, pp. 14-15).

It is most likely that it is young people who apply to university although it is also possible that older generations apply for the different programmes. Thus, young people are the primary target age group of higher education institutions. Young people, members of generation Z, born after 1995, are reliant on technology both in Hungary (Bernschütz et al., 2016) and elsewhere (Schwieger & Ladwig, 2018). The older generation often has the impression that the computer literacy of those born after 2000 does not even need to be developed, as young people today seem to be born with it. As Barlow (1996) puts it, this generation comprises "natives" in the cyberspace where a teacher can only be "an immigrant". According to the World Bank (n.d.) in 2016, for example, 76% of Hungarian households owned personal computers (World Bank, n.d.), so test administrators can expect test-takers to have sufficiently good computer skills and can assume that their performance will not be negatively affected. Furthermore, institutions may believe that computer-based entrance tests represent additional marketing value matching their customers' preferences.

It must be added though that the fact that the majority of households and young people labelled digital natives nowadays have computers and smartphones does not necessarily mean that all test-takers are capable of efficiently using these devices for serious purposes. Some of them may only use the computer for entertainment, and may not know for example how to format Word documents. As it was pointed out by Bennet et al. (2008), "it may be that there is as much variation within the digital native generation as between the generations" (p. 779). Therefore, the expectations of universities and language examination centres might be unrealistic and institutions have to be prepared to provide assistance to their test-takers.

2.2 Test-takers' advantages

This section will examine what benefits test-takers may encounter when taking a computerised test. After discussing flexible scheduling, time-efficient scoring and reporting will be discussed. Last, functions not available in paper-based tests will be reviewed.

2.2.1 Flexibility and speed

Cambridge Assessment offers nearly "700 test sessions a year" worldwide (Jordan et al., 2011, p. 2), which suggests that there is demand for the high number of test-taking opportunities allowed by the use of advanced technological solutions. Another examination provider allows applicants to take their exams every twelve days (About the TOEFL iBT® Test,

n.d.), so, evidently, the flexibility with which exam providers offer their computer-based tests is a much appreciated benefit. With computer-based tests, there are fewer logistical constraints than with mass paper-based ones, as Roever highlighted (2001).

In addition to flexible scheduling, Mohammadi and Barzgaran (2010), Pathan (2012), and Taylor et al. (1999) emphasise that immediate scoring and fast reporting are also advantages for the test-taker. For example, test-takers will receive their IELTS results 13 calendar days after the test date (Results – When and how to get your score, n.d.). Linguaskill, the online test by Cambridge Assessment which replaced the BULATS test at the end of 2019, promises to release results in as little as 48 hours (Linguaskill, n.d.). Cambridge Assessment releases its A2 Key to C1 Advanced level computer-based results earlier (two to three weeks after the exam date) than the paper-based results (four to six weeks after exam date) (Results, n.d.). TOEFL scores can be seen online six days after the test day (Getting your TOEFL iBT® Scores, n.d.). Out of the four language examination centres in Hungary that began to offer at-home testing in Hungary in May 2020, two promised to release results seven days after the test date the latest (iTOLC website information, n.d.; iXam website information, n.d.). Fierce competition between examination centres puts pressure on them and promising to report results early is clearly a must these days since the pressure is equally large on examinees to be able to present language certificates.

2.2.2 Navigability and editability

During test-taking, another advantage is that candidates can review and correct their own responses with ease, just as well as revise their written work neatly (Russel, 1999) unlike on paper. Yet another benefit is that candidates can work at their own pace (Pathan, 2012; Taylor et al., 1999), and may finish earlier. Most computer-based test-taking applications indicate it clearly where candidates are within the test, which items have not been answered yet and how much time they have left, the last of which is actually a requirement specified in the Hungarian Accreditation Manual on language examinations (Akkreditációs Kézikönyv, 2020, p. 25). The software might warn candidates if they have left an item or task out and might even have an in-built spellcheck or dictionary in one or some part of the test. This latter appears to be a major advantage since generation Z tends to use online dictionaries instead of printed ones. A Japanese study found that high school students "have a lesser sense of closeness to a printed dictionary than college students" (Koyama & Takeuchi, 2003, p. 73), so one can assume that the younger the candidates are, the more they will appreciate the availability of an in-built dictionary.

2.3 Disadvantages of computer-assisted testing

2.3.1 Theoretical considerations

One disadvantage some professionals thought might exist is the effect of the medium, namely, the computerised format, on test results. The construct validity of computer-assisted tests was examined by many (e.g., Fulcher, 2000; Mohammadi & Barzgaran, 2010) but the findings are inconclusive. Bangert-Drowns (1993), for example, found that the quality of texts produced on computer improved (as cited in Russel, 1999). These findings refer to scripts produced in class under untimed conditions (Russel, 1999), but under timed conditions in a crossed design study on language, arts, science and maths Russel obtained mixed results. Test-takers with high word per minute (WPM) keyboarding speed were found to have scored

significantly higher on computer tests (p. 24). On the other hand, Russel remarked that "it is likely that more students will develop solid keyboarding skills and, thus, will be adversely affected by taking open-ended tests on paper" (p. 42). Mohammadi and Barzgaran (2010) also pointed out that familiarity with computers was on the rise. After factoring out the sequence effect, based on correlations and paired samples T-tests, Mohammadi and Barzgaran (2010) concluded that the computer-based and paper-based language tests measure the same construct.

In order to examine whether familiarity with computers affects test scores, researchers' attention turned to the issue of computer literacy, the degree of familiarity with computer use. For example, in a 2010 study Mohammadi and Barzgaran found no statistically significant difference between the scores of groups with high computer familiarity and low computer familiarity in terms of a language test. Taylor et al. (1999) had analogous findings in a large-scale study in the context of the TOEFL test. First, a computer familiarity research tool was devised based on the responses of 90,000 test-takers. The final 11-item questionnaire gave a composite score for computer familiarity and test-takers were grouped into high, low, and moderate computer familiarity groups. After taking a tutorial, the 1,200 test-takers did a 60-item computer-based test and no meaningful relationship was found between the test-takers' level of computer familiarity and level of performance on the 60-item test.

Exam providers soon realised that the more familiar candidates are with the exam format, the less effect computerised delivery will have on their scores. Therefore, the role of providing information to candidates not only on the different components of the test, but on the mode of delivery became much more important. For example, ETS introduced compulsory tutorials and gave free sampler CDs in the early days (Fulcher, 2000). For the same reason, there are numerous tutorials available on the Internet for various examinations at present, for instance, ETS even has its own TOEFL TV channel on YouTube and candidates may try their hand at the mode of delivery Oxford University Press tests have on their website (The Oxford test of English demo, n.d.).

Although computer-based tests are considered to be more suitable than paper-based ones for technology-savvy applicants, especially in the developed world (Casey, 2013), it is highly probable that computer anxiety still exists. Some applicants, even members of generation Z, and some of those in charge of administering computer-based tests may still have what is referred to as "unfamiliarity, anxiety, or hostility" towards the computer as a testing medium (Henning, 1987, p. 137). As a result of the age difference, the staff responsible for administering or invigilating computer-based tests are less familiar with computer technology than the candidates who take the tests. In modern institutions, though, as Pathan (2012) pointed out, an "understanding of the nature of computer-assisted language testing" is expected from colleagues (p. 38). Stress may occur because a new system needs to be learnt, or because of the fear that that system could crash (Pathan, 2012, p. 40).

2.3.2 Technicalities

A further disadvantage of computer-assisted testing is related to limitations of the medium. Some tasks may very easily lend themselves to computerisation, while others do not. Open-ended answers may need to be manually graded one by one, slowing down the otherwise time-efficient scoring process. In testing reading, for example, items on vocabulary that can be found in certain lines have to be indicated in other ways than by referring to the given line. In terms of scoring, the computer will only accept solutions that it has been instructed to, therefore, a detailed and carefully pretested and coded key needs to be carefully worked out (Suvorov &

Hegelheimer, 2014). Test-takers may achieve a lower score because of misspelt words, using capitals instead of lowercase letters, accidentally hitting the keyboard, or moving the mouse wheel, problems that can again only be dealt with manually. Although computer screens have been growing recently, a further limitation is the size of the screen, longer passages may have to be excluded (Pathan, 2012) or replaced by several shorter ones for easier navigation.

A further issue that arises is the difference between the capacity or configurations of different computers or devices. Language exam providers eliminated the problem by standardising equipment in their test centres but with the number of test centres with computers increasing and the unstoppable development of the computer industry, this is becoming a challenging task. Internet-based tests may be running on very different devices as more and more computers are used across and within different institutions which operate as test locations. With the introduction of at-home testing internationally (TOEFL) and in Hungary applicants' own computers are used in the homes of test-takers, which makes it impossible to standardise technology. One solution is to define minimum configuration requirements, but there are doubts whether that is sufficient to ensure equality of opportunity.

Text length and its visibility on the screen may present difficulties to some test-takers as well. One study (Haas & Hayes, 1986, as cited in Russell, 1999) found that when passages covered more than one page, computer-based reading tests yielded lower scores than paper-based tests. In a 2001 study, Bridgeman et al. examined the effect of screen size, screen resolution, and display rate in the context of a maths and reading test with 357 high-school juniors and found that screen size did not, but resolution did affect verbal performance scores (but not maths scores). The participants who used screens with higher resolutions tended to score significantly higher (p. 11). Although the study used screens of 15 and 17 inches and resolutions of 1024 x 768 and 640 x 480, which are clearly out-dated today, the findings are worth mentioning since with a lower resolution participants were able to see less of the text and needed to do more scrolling. In an accompanying questionnaire, 66% of the participants reported that scrolling interfered with taking the computer-based test to some degree while font size, use of the mouse, screen clarity, screen size were reported not to have interfered (p. 15). Computer literacy again might interfere with test-taking since more computer-literate candidates are more likely to be able to adjust letter size and the proportion of the text they can see at the same time. Evidently, the use of technology in itself always carries some risks, paper-based tests are doable if the candidate is equipped with a pen and the tasksheets, whereas the lack of electricity, an Internet service outage or merely a slowdown, an inadequate setting or simply forgetting a password might cause unnecessary loss of time, anxiety or even failure to administer a test.

Although the project described here reports work related to an English language entrance test in an international school, the question of what keyboard layouts and settings are available arises naturally. Test-takers may be used to a certain keyboard and using a different one may slow them down or make them anxious, thus putting them in a disadvantaged position. Similarly, special characters may be necessary for testing the competence of speakers of languages with accented characters, such as Hungarian (ö, ü, ó, etc.) or French (é, ê, ç, à, ï, etc.), which are mostly accessible on Qwerty or Qwertz keyboards by using ASCII codes or other shortcuts, but their use is certainly not very convenient. By, for example, typing ALT 0244, one will get ô. Non-western languages may present additional difficulties. In Hungary, the keyboard layout characteristic of the given foreign language is a minimum requirement for accredited language examination centres (Akkreditációs Kézikönyv [Accreditation Manual], 2020, p. 25).

In sum, computer-based testing presents some theoretical and practical challenges. On the other hand, paper-based testing has always had its own risks and demerits as well. It must be borne in mind that it is very likely that the two platforms are appealing to two distinct groups of test-takers and the availability of both platforms would be the ideal solution. The two platforms, however, both have their shortcomings. The challenges concerning the administration of computer-based tests deserve researchers' attention because of their relative novelty.

3 The project

3.1 Research context

The project was undertaken in a business and management college in Budapest. IBS is a private higher education institution in Hungary, the first of its kind, established in 1991. IBS offers two kinds of bachelor programmes: 1) Programmes accredited by the Hungarian state and 2) programmes validated by the University of Buckingham. In addition, a one-year intensive language programme, master's and PhD programmes in business and management related fields are also taught to altogether approximately 1,000 students a year. In most programmes, teaching is in English, therefore, the language competence of applicants is measured when they apply to start the degree programmes. IBS accepts international certificates as well as Hungarian state-accredited ones, for instance, B2 level state-accredited certificates, a score of 6.0 on the IELTS Test for the bachelor programmes, or a C1 Certificate for its master programmes. Applicants who do not have certificates at the level required by the college take the internal entrance placement exam, the Single English Test of IBS (SETI). In order for the entrance requirements to be a coherent system, a study was conducted to link the SETI Test and its scores to IELTS Scores.

SETI Components	Points	Total
Use of English	15	
Reading 1 Academic	15	
Reading 2 Newspaper article	20	
Writing 1 Letter or email	20	
Writing 2 Mini-essay	30	
Reading and Writing		100
Listening 1 Everyday conversations	20	
Listening 2 Lecture or talk	30	
Speaking	50	
Listening and Speaking		100

Table 1. Components of the SETI Exam in 2020

The SETI measures the four skills: reading, writing, listening and speaking (For details, see Table 1). The focus of this study is on the reading and writing components, which were computerised well before the 2020 pandemic forced the school to conduct all exams, including English Listening and Speaking exams online. The SETI exam has multiple aims: 1) to filter and sort into groups incoming applicants according to their language skills, 2) to exempt students from English during their studies. The exam is seen as a high-stakes test because passing it allows an applicant to start their studies at the bachelor's or master's level. If applicants score high enough in all the four skills, they are exempted from a number of English

language development classes in the first year. English language development classes are taught in two modules, English Reading and Writing (ERW) and English Listening and Speaking (ELS), in eight contact hours weekly over two semesters in the first year. Students may be exempted from either module or both (e.g., ERW), depending on their test results.

Using Suvorov and Hegelheimer's framework (2014), the Reading and Writing components of the SETI are linear tests. Test-takers are allowed go back to previous items and change their answers. The Written Section of the SETI is a single medium test with no multimedia content. All tasks target one skill, there are no truly integrated tasks although reading and writing inevitably get integrated. The electronic version of the SETI is web-based and is evaluated by both human raters and the computers and will be referred to as the E-SETI. Applicants are to give both selected and constructed responses during the test.

The school has adopted Moodle as a virtual learning environment (VLE) to offer college students a blended learning experience. In the VLE, to which newcomers are automatically added since first-year modules are centrally allocated, students have access to the intended learning outcomes, the syllabus, some course content, requirements, the assessment criteria and some supplementary sources as well. For example, some tutors upload their slides, handouts, links to videos and practice exercises. Students have to upload their assignments to Moodle and this is where their marks and absences are registered. Moodle is a VLE, the first version of which was originally developed in Australia in 2002 (History, n.d.) with the intention to improve distance learning. In line with the school policy, English language testing in IBS was transferred to a Moodle-based website in 2017, a site which is not integrated into the Moodle platform that is used for teaching. Moodle 3.5.6 today is used to measure incoming applicants' English language competence with the E-SETI.

Applicants to IBS are young people, half of whom are Hungarians while the other half are non-Hungarian speaking applicants, from various countries in Asia (Mainland China, Azerbaijan, Kazakhstan, Turkey, Iran, etc.) Eastern and Northern Europe (Ukraine, Serbia, etc.) or Africa (Nigeria, Cameroon, Libya, etc.). The typical Hungarian applicant has lived or studied abroad either because of mixed parentage or parents working and living abroad or because the parents find it important to send their children abroad. Foreign applicants, too, may be considered to be cosmopolitan in many respects.

3.2 Aims and research questions

During the computerisation project, action research was conducted in order to continually review and carry out the project as well as possible (Cohen et al., 2007). Data were collected in different ways in order to be able to answer the following questions:

1 How suitable is Moodle for administering the SETI exam?
2 Does the mode of delivery affect the performance of applicants?
3 How do applicants perceive the electronic language exam?

3.3 Participants

To supplement the lessons learnt during the computerisation project, data were collected by asking 225 applicants to answer three questions about their test-taking experience after their

entrance test. One third were Hungarian and two thirds were from elsewhere, 40 of whom were Erasmus students. Over 60% of the applicants were male, almost 40% female.

3.4 Design and instruments

Many of the answers to the research questions are rooted in experience and were worked out in collaboration with the information technology experts of IBS and the available Moodle documents (Moodledocs, n.d.) on the massive support site that accompanies the VLE. Some of the lessons were learnt form comparing Moodle statistics with those obtained from analysis carried out on the same data extracted from Moodle with Excel spreadheets. Quantitative and qualitative data were collected to answer questions 2 and 3 in the second half of 2018 in the form of responses to three further questions added to the E-SETI:

1. Is this the first time you have taken a test on the computer?
2. Please rate today's test-taking experience on a scale of 1-5, 1 meaning very bad, 3 meaning neither bad nor good, 5 meaning excellent.
3. Please tell us why you gave the score above or make any comments on your experience.

Since the whole test is delivered using the English language, all instructions are in English, the above questions, too, were in English. Applicants were free to skip the questions if they wished. Answers to the questions were not anonymous in the sense that they were attached to the test-takers' responses to the E-SETI for the purposes of the analysis.

Applicants' language competence was measured with the computerised version of the internal language exam of IBS, the E-SETI. As a result of computerisation, E-SETI is delivered in a computer room using a Moodle testing site specifically set up for this purpose. Unlike in Table 1, at the time of the research, the SETI Reading and Writing exam consisted of a Writing Component (35 points), a Grammar and Vocabulary component (35 points) and a Reading Component (30 points) (as summarised in Table 2), but the Specifications were modified in 2019 to allow more room for the testing of the four skills and to have a less grammar focused entrance test. The tasks that were excluded as a result of the modification are italicised in Table 2.

E-SETI Components	Points	Total
Writing 1 Text-based open gap-fill	15	
Writing 2 Paragraph	20	
Writing Total		35
Grammar and Vocabulary: Grammar multiple choice	*10*	
Grammar and Vocabulary: Vocabulary multiple choice	*15*	
Grammar and Vocabulary: Word formation	*10*	
Grammar and Vocabulary Total		35
Reading: Academic reading	15	
Reading Newspaper article summary gap-fill	15	
Reading Total		30
Reading and Writing SETI Total		100

Table 2. Components of the SETI Exam in 2018 excluding Listening and Speaking

3.5 Procedures and analysis

Before the introduction of the computer-based test, the format was validated by using verbal protocols and focus group interviews. First, a student with C1 language competence and a relevant certificate was asked to do the E-SETI and verbalise his thoughts at the same time. Then, post-test focus group interviews were conducted with two students, who were asked to comment on the test-taking experience. These validation steps greatly enhanced the computerisation project by identifying areas that need to be improved. As a result of the validation procedures, the layout and formatting of the tasks was modified to better separate the instructions from the text of the tasks and some regulations were changed, for instance, applicants were allowed to use paper for their notes.

In the quantitative phase of the data collection, applicants were asked to answer the three research-related questions listed in 3.4 above. They were also invited to give textual feedback on their experience. Responses varied in length and depth with the lowest answer rate for the last, open-ended question. Descriptive statistics, independent samples T-tests and Mann Whitney U-tests were calculated using Excel and the SPSS software package.

4 Findings

4.1 Suitability

In the following section the answer to the first research question in 3.2 above will be discussed, that is., how suitable the platform proved to be. The findings are based on the experience of using the Quiz function of Moodle and a comparison of Moodle analytics and the results of statistical analysis based on classical test theory.

4.1.1 Transferability

Both the tasks specified in the earlier specifications of the test (Table 2) and the latest specifications (Table 1) were easily transferable to the Moodle platform. The task types available on the Moodle platform or as downloadable plug-ins included sentence-based multiple choice and gap-fill exercises, for example, which were painlessly imported in large numbers from an Excel file. The variety of tasks includes multiple choice, true or false, matching, short answer (gap-fill) exercises, drag and drop exercises and text-based gap-fill (embedded answers in Moodle terminology). Multiple choice items constitute one question of Moodle and text-based tasks as well, which could present a challenge to test-taker and the analyst as well. Test-takers might underestimate the importance and time requirements of a longer text-based task if it looks exactly the same on the navigation panel as a single sentence with a multiple-choice gap in it. An example of the coded background text of a text-based task (shown later in Figure 3) can be seen in Figure 1, which shows that the gap-fill task (Quiz question in Moodle terminology) of Moodle allows the developer to enter multiple correct answers. Analysts, on the other hand, will need database specialists to extract the item-level data from the platform, as the in-built analysis tools calculate discrimination and facility values (discrimination index and facility index in Moodle terminology) for Moodle questions, and this information is not available for each and every item in the case of text-based gap-fills (gapfill or embedded answers type of question in Moodle terminology). This might distort the

automatically calculated reliability (Coefficient of internal consistency, CIC, in Moodle terminology) figures as well.

When we think **(e.g.)…*of*…** the North Sea, we think of ice, icebergs and animals [*like*] penguins and seals. However, it may all change in [*the*] next few years or decades, because the number of dolphins, whales and other marine creatures [*is*] rising in the North Sea. They are usually found in warmer waters, [*but|although|though*] recently up to six different types of dolphins [*have*] been spotted along the coast of Great Britain, stretching [*from*] Northumberland to North Yorkshire. During one sighting, a school of about 250 white-beaked dolphins were seen 40 km out from Cullercoats. Since 2003, Newcastle University's Dove Marine Laboratory has been looking at dolphins and whales in the region. The University recorded sightings of 614 individual creatures [*in|within|during|over*] a 12-month period.¶

Figure 1. Example of a coded background text

4.1.2 Security

Naturally, test security issues were a major concern. Applicants take the tests in computer rooms where strict invigilation is ensured, for instance, electronic devices need to be switched off. The invigilator ensures that applicants do not communicate in any way, and do not open other sites or a word processor. Functions of Moodle that enhance security are the resettable password or the random selection function. If administrators use previously calibrated tasks and items, they can programme Moodle to randomly select items and tasks from categories (Moodle terminology) for certain item or question positions. Question positions in Moodle refer to the place of an item (specifically selected or randomly inserted from a pool) in a series of items. Random selection means enhanced security since test-takers who sit next to each other will probably be given a different item or task of the same difficulty. Additional security is achieved when applicants choose to do the items and tasks in the order they prefer, which further complicates copying from a neighbour.

At the beginning of the project, a Moodle testing site accessible only from the IBS campus was set up to prevent outsiders from accessing it, which meant enhanced security since only school computers were used to access the site. After about a year, the site became accessible openly from the Internet because foreign applicants were supposed to be able to access it. The site is not yet used for distance testing but is an important virtual venue to register for entrance placement tests. Each applicant sets up their own account and password. To ensure a safe testing environment, before the exam session on campus, all applicants received their password for the test (Quiz in Moodle terminology) in addition to their own personal user name and password to access the testing site. With online registration, test-takers establish their accounts themselves, so not even the invigilator knows their confidential data. With the help of the Quiz password, only applicants who are present in the examination room can access the test. Password use, however, proved to be a somewhat risky and time-consuming process at the beginning of the test sometimes requiring help from invigilators and system administrators on the spot since applicants do not remember their passwords or get confused with having to use several.

Moodle allows administrators to compile different variants of the same test. For example: if two reading comprehension tasks of approximately the same difficulty are moved

into a category and are given to the test-taker randomly, test-takers have a 50% chance of having the same task in their test as the person sitting next to them. If, however, the test administrator decides to classify Task 3 into a separate category, that will be given to all applicants. The items or tasks given may be shuffled and the applicants' chances of having the same item or task as their peers at the same time is reduced. Since the categories are established by the test administrator, using the same item or test on different administrations can be avoided. Moreover, the multiple-choice items, as well as their distractors, can be re-shuffled to make cheating even more difficult. To illustrate the category structure, the above example is presented in Figure 2.

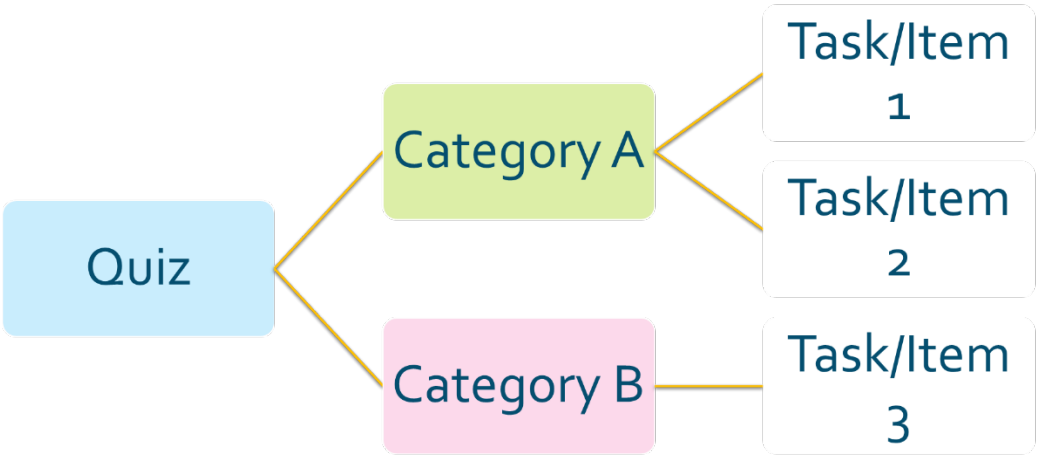


Figure 2. An example of the Category structure

In a test in which there are three question positions, for each of which questions are randomly selected from a category in which there are three questions, the result totals 27 tests that to some extent differ in content but, if the items are classified adequately, do not differ greatly in difficulty. Thus, the applicants sitting next to each other will almost certainly not be given the exact same set of questions (P = 0.03). An inherent risk is if items are not or not well enough calibrated, because in this case applicants will be given test variants with varying difficulty. As the Moodle support site says, "You don't want one variant to be much harder than the others" (Quiz statistics report, n.d.) and advises them to study facility values and discrimination indices (FVs and DIs).

Increasing test security by using random items or tasks from a previously selected pool, therefore, assumes the careful monitoring of item and task difficulty. This requires that records of previous administrations or pre-tests are kept to make informed decisions concerning items and tasks. Although the underlying Question bank (Moodle terminology) is called a bank, it is not an item bank in the sense that no indices are stored alongside the items or tasks. The FVs and DIs can be accessed by using the in-built features, but the item bank needs to be established by collecting the information on calibrated items and tasks.

4.1.3 User-friendliness

Another concern was the user-friendliness of the platform, that is, its navigability and appearance for test-takers. Until the platform was only accessible from within the institution, applicants saw the application for the first time when they actually had to perform well on the

language competence test, which was far from ideal. Accidentally, after being given detailed instructions they easily learnt to use it. Today, with the site open to all applicants, a Sample Test is available to them so they can familiarise themselves with the features of the application.

The layout and some of the features of the Quiz function of Moodle can be seen in Figure 3. (The coded background text of the task was shown in Figure 1 earlier.) The navigation panel showing applicants that they have to do five tasks is on the left and the passage to fill in can be seen on the right. Applicants may choose to hide the navigation panel if they wish. There were no specific questions concerning the layout and applicants did not comment on the organisation of the different elements on the page. The task (question in Moodle terminology) can be flagged by applicants if they wish to return to it later. Administrators may edit the task by clicking on the cogwheel icon that can be seen in the middle (Edit question). The screenshot below was taken in preview mode, note the *Start a new preview* button. Previewers or applicants can see the amount of time still available for the preview or the actual test on the left hand side of the screen, one hour 39 minutes in this screenshot.

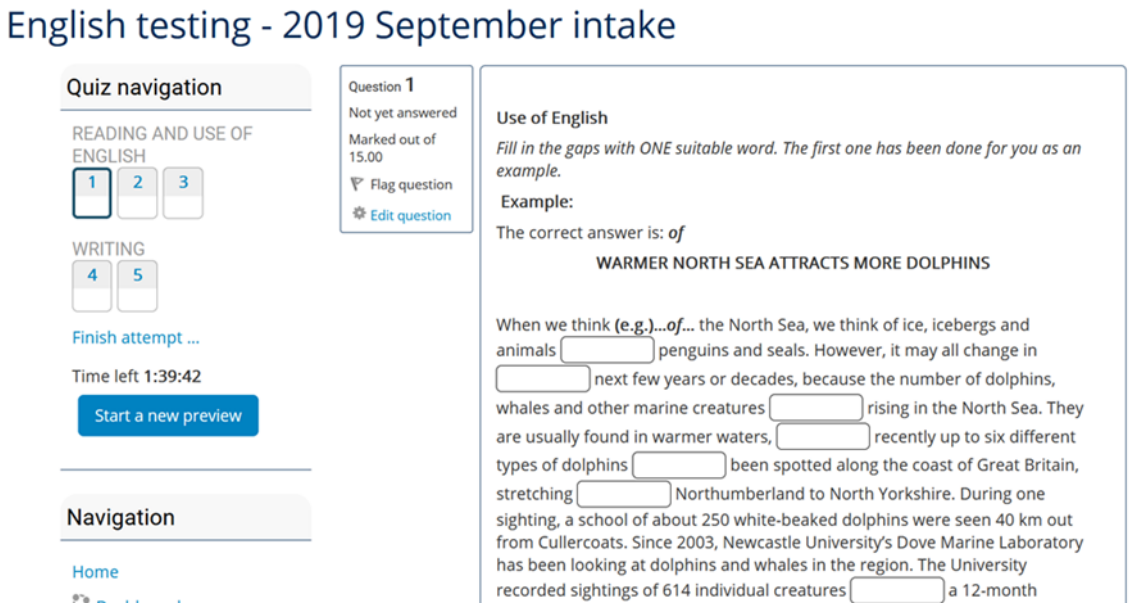


Figure 3. Appearance of a Use of English task in the E-SETI Test in Moodle

By clicking on *Previous page* or *Next page* visible in Figure 4, the test-taker can easily navigate between different tasks. Before the submission of the attempt, applicants are automatically reminded to review their answers. They can also see which questions they have left unanswered or flagged and make any final necessary changes. If the administrator wishes to insert other reminders or special instructions into the test, they can do so by adding a description (Moodle terminology).

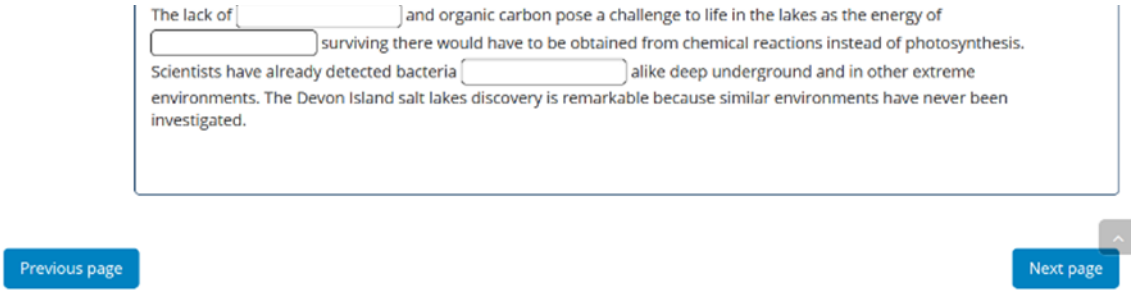


Figure 4. An example of the navigation buttons

4.1.4 Technicalities

When the first E-SETIs were administered, system administrators gave the computers used for testing priority over all others on campus to prevent applicants from being held up by their computers. Over the years, not more than a dozen instances occurred when applicants were warned by Moodle that their answers were not being saved and these typically lasted less than one minute.

Productive writing tasks can be completed in additional dialogue boxes (Essay question in Moodle terminology). There is no word count function, which means candidates have to guesstimate the length of their letters and essays. Although one would expect a computer-based platform to be equipped with this feature, this was not considered to be a major drawback since applicants taking a paper-based test do not have an automatic word count function either.

Automatic scoring, a major advantage of the application, speeds up scoring: multiple alternative solutions or spelling variations (e.g., both British and American, case-sensitive or non-case-sensitive) can be given in the key, the system can be programmed to award an equal number of points, or if necessary, a different number of points for each alternative solution. The automated scoring can easily be double-checked by an administrator who can remedy unexpected problems (e.g., a misspelt word in the key can be corrected). Very often, problems are easier to identify than to solve, for example, the key may not work because of a question mark and it may take time even for information technology experts to find the solution to use the Disable regex function by trial and error. The paragraph writing task of the E-SETI exam, later renamed as mini-essay and supplemented by a letter writing task, made it necessary to include a manually graded essay question. The section which included automatically scorable items was scored by Moodle while the essay questions were marked by the raters.

In the essay questions, however, a minor problem was identified, that of checking the spelling. Originally, in their handwritten form, in these productive tasks student were not allowed to check their spelling in any way and by default, there is no spellcheck within Moodle but browsers automatically underline incorrectly spelt words. In our experience, this proved to be an advantage for the applicants who notice this in the stressful conditions of a test, and a disadvantage for those who do not. Even if applicants are informed of the availability of the spellcheck, those less familiar or simply slower with computers will be at a disadvantage.

4.1.5 Data processing

One might even learn from the robust Moodle support site where, for instance, advice is given on the use of statistical calculations available in Moodle (e.g., acceptable figures of

standard deviation). About the average grade achieved by students, for example, the site says users should aim for "between 50% and 75%. Values outside these limits need thinking about" (Moodle quiz report statistics, 2010). Another page calls teachers' attention to "broken questions", where the discrimination index might show if the correct solution was marked incorrectly in the key (Quiz statistics report, n.d.). In terms of the CIC, the advice given is "Anything above 75% is satisfactory. If the value is below 64%, the test as a whole is unsatisfactory and remedial measures should be considered". This tallies with the requirement set out in the Accreditation Manual for accredited language exam centres in Hungary (Akkreditációs Kézikönyv [Accreditation Manual], 2020, p. 38) and the assessment literature (e.g., Alderson et al., 1995; Hughes, 2003).

What follows is an overview of the functions Moodle offers for data analysis. One of these is the visually helpful quick analysis tool, the distribution table based on the performance of applicants attempting the same test, which informs administrators whether there is any skew. To illustrate this Moodle function, an example of an automatic Moodle chart is provided in Figure 5 below showing the distribution of the scores of the reading and use of English components of the E-SETI test that first year students took in May 2018. By clicking on *Show chart data* at the bottom, one can have access to the actual number of test-takers achieving these scores.

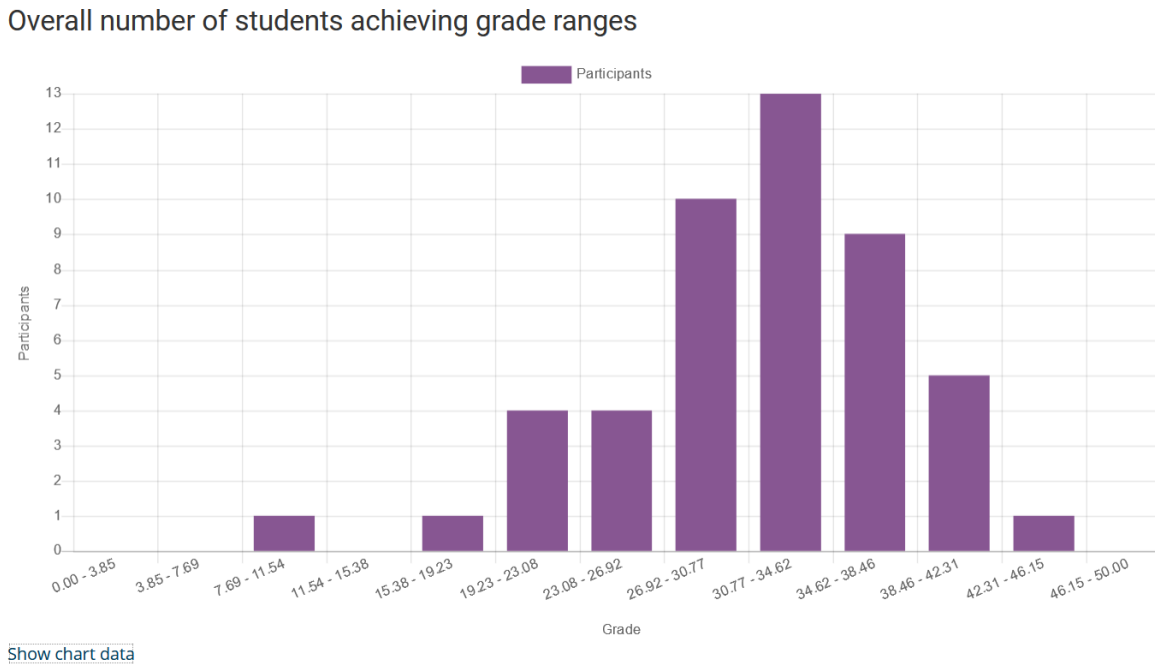
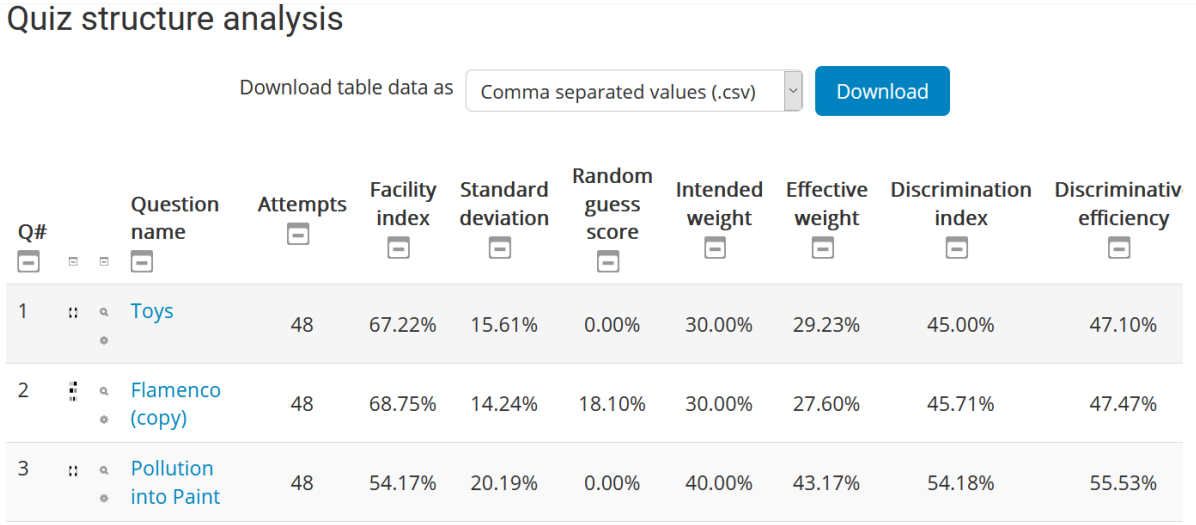


Figure 5. An example of the distribution of test-takers' scores provided by Moodle

In addition to distribution tables, Moodle also provides facility values (FV, facility index in Moodle terminology) and discrimination indices (DI). In Figure 6, examples of these statistics are given for three different task types: the Use of English (Gap-fill), Reading 1 (embedded answer) and Reading 2 tasks (gap-fill question types in Moodle) (See Table 1 for Components) of the SETI test whose distribution was depicted in Figure 5. The original Moodle title (Quiz structure analysis) and format has been retained to demonstrate the way data are offered in Moodle, but as it can be seen, the data are downloadable in .csv or .xls format as well.



Q#	Question name	Attempts	Facility index	Standard deviation	Random guess score	Intended weight	Effective weight	Discrimination index	Discriminativ efficiency
1	Toys	48	67.22%	15.61%	0.00%	30.00%	29.23%	45.00%	47.10%
2	Flamenco (copy)	48	68.75%	14.24%	18.10%	30.00%	27.60%	45.71%	47.47%
3	Pollution into Paint	48	54.17%	20.19%	0.00%	40.00%	43.17%	54.18%	55.53%

Figure 6. Examples of FVs and DIs in Moodle

Based on the information provided in Quiz structure analysis tables like the one in Figure 6 above, it is possible to monitor the way items and tasks work, necessary adjustments can be made for a later reuse of the items and tasks or the foundations of an item/task bank can be established. Although Fulcher says that "the general rule of thumb is that items within a range of .3 to .7 should be included" (2010, pp. 182-183), the support site does not have any recommendations in this respect.

Figure 6 shows that the statistics are based on 48 attempts, all test-takers completed the same three tasks and, as intended, the third task is the most difficult. Once data such as the above have been collected on the tasks, items and tasks of approximately the same difficulty can be identified assuming the number of test-takers is sufficiently high and they are of roughly the same ability in each administration. Ideally, some pretesting should be done to examine the quality of the items and tasks and to link data sets. Figure 7 illustrates an example after the calibration has taken place.

Q#	Question name	Attempts	Facility index	Standard deviation	Random guess score	Intended weight	Effective weight	Discrimination index
1	Random (UoE1 CYE Sample)	39	38.46%	29.16%		30.00%	29.83%	65.42%
1.1	Dolphins in the North Sea (copy)	14	38.57%	31.07%	0.00%	30.00%		74.65%
1.2	Low Fertility (copy)	25	38.40%	28.69%	0.00%	30.00%		86.44%

Figure 7. Example of the statistics for two tasks in the same question position

Here, the first column presents the identity number of each task, the second presents what type of Quiz question was used, the fourth the title of the task. The fifth column informs the user of the number of attempts made, the sixth column about the facility index, and the last one shows the discrimination index. Following previous administrations, here, two tasks of the

same type (text-based gap-fills: 1.1 Dolphins in the North Sea and 1.2 Low fertility) were selected as tasks of the same level of difficulty and were classified into the same category. Moodle then randomly gave them to test-takers as a position 1 question in the 39 attempts. The data comes from a sample test, where students were allowed to do the test any number of times. The same difficulty is reflected by the FVs given in Moodle (38,57% and 38,4%) although, naturally, the two groups (those attempting task 1.1 and those attempting 1.2) are not the same with probably some overlap. The DIs proved to be slightly different, but the two tasks have approximately similarly high discriminating power. The Moodle support site says, if the DI is above 50, the item/task discriminates very well between able and less able students (Moodle quiz report statistics, 2010). This recommendation is in line with the recommendations in the relevant literature (e.g., Alderson et al., 1995; Bachman, 2004; Bárdos, 2002).

In the following example, a sharp drop in CIC will be more closely examined. When the specifications of the SETI exam were changed and grammar and vocabulary-focused sentence-based items were dropped in favour of a skills focus and text-based tasks, the CIC, which had previously shown a high reliability, usually above 0.7, had decreased to below 0.5. As the Moodle support site and the literature recommends, such a low figure indicates that test-takers might achieve a different score on a second administration of the test. Similarly, experts agree that the reliability figure needed to be higher. To quote Fulcher and Davidson (2007, p. 107), "tests that do not achieve reliabilities of 0.7 are normally considered to be too unreliable for use". Since the test included 40 items and detailed data were only available for 3 Moodle Questions, it was assumed that the reliability of the 40 items is probably higher. Therefore, with the help of data engineers the test-takers responses were extracted from Moodle and the reliability (KR-20) was calculated with an Excel spreadsheet. The results of the comparison can be seen in Table 3. In Tables 3 and 4 all figures are presented as percentage figures as in Moodle.

Statistic	Recommendation on Moodle support website	Moodle-calculated figures	Excel-calculated figures
Average	50-75%	62.46%	61.66%
CIC/Reliability	≥75%	45.55%	72.8%
Standard deviation	12-18%	13.45%	6.48%
Standard error	<8%	9.93%	3.38%

Table 3. Comparison of the E-SETI Exam statistics

Table 3 shows that the reliability and standard error calculated by Moodle were worrying while the actual reliability and standard error figures based on item-level responses were reassuring. It was not possible to identify what caused the difference in the average score, but it is a negligible difference. The comparison showed that further work was necessary to ascertain the reliability of the test. Without the help of information technology colleagues, this would have been impossible and even with their help, it was not a rapid procedure. The responses of gap-fills (Tasks 1 and 3 in Figure 6) were easy to extract while the responses to the embedded short answer question (Task 2 in Figure 6) were more time-consuming and needed to be manually modified since alternative correct solutions were not extractable in a data file. The facility values presented in Figure 6 and revisited in Table 4 below were double-checked and insignificant differences were found.

E-SETI Task	Moodle-calculated FVs	Excel-calculated FVs
Toys	67.22%	65.69%
Flamenco	68.75%	68.75%
Pollution into paint	54.17%	53.33%

Table 4. Comparison of the FVs of three E-SETI tasks

In conclusion, Moodle offers useful statistics concerning test-taker and test performance that might call the user's attention to errors in the key, a slant in the distribution or differences of difficulty in term of items and tasks provided the user reads the information presented on the support website or is familiar with professional standards. More care needs to be taken with text-based tasks (gap-fills and embedded answers) that include multiple items while the reports are easy to use with sentence-based gap-fills. With regard to text-based tasks, item-level data are also available, but the user will need the help of a database specialist to extract the data. Thus, obtaining statistical information concerning several items within one Quiz question (i.e., a text-based task as embedded answers question type) requires more time, expertise and the close cooperation of database engineers and measurement specialists. Although the setting up of an item or task bank is possible, it requires further efforts that go beyond a mere use of Moodle.

4.2 Mode of delivery

The question whether test-takers' prior experience with computer-based tests affected their language competence scores was examined by collecting data from 225 test-takers. The distribution of their total written E-SETI scores (Reading, Use of English and Writing) can be seen in Figure 8 below.

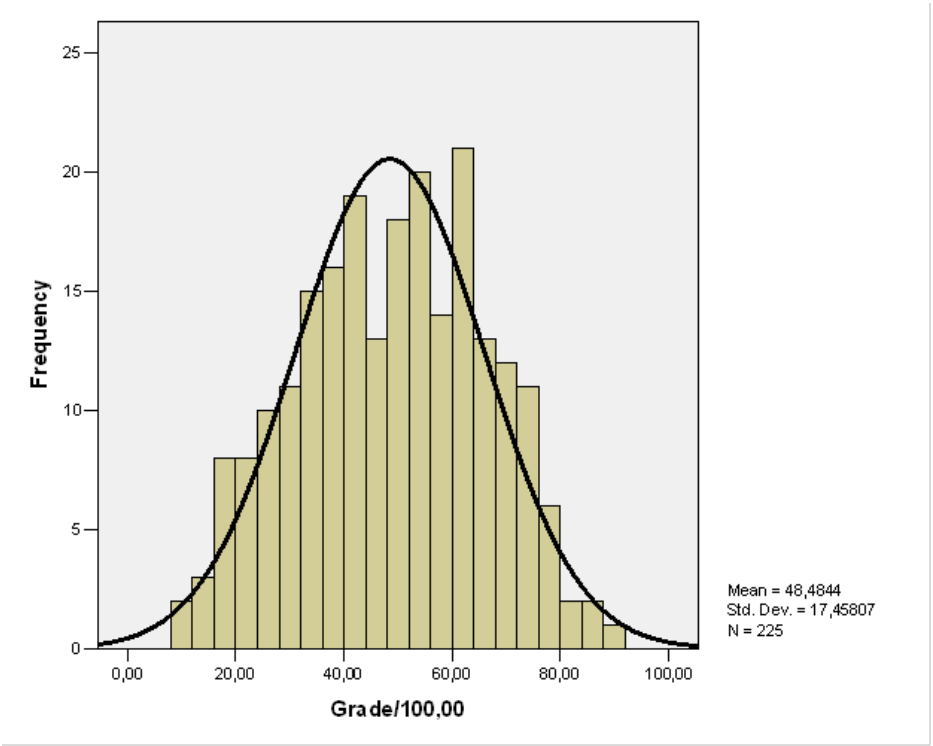


Figure 8. The distribution of the scores of the sample produced by SPSS

On the basis of whether this was the first time that test-takers encountered a computerised test they were classified into two groups. Forty-two applicants did not answer this question. For 99 test-takers this was the first computerised test in their lives while for 84 it was not. Table 5 shows that the mean scores of the two groups did not differ critically, and the independent sample t-test showed that the difference was not significant ($t_{(183)} = -.102$, $p = .919$; see Table 5 for further details).

Is this your first computerised test?	N	Mean	Standard Deviation	Std Error Mean
No	84	48.5774	18.70	2.04
Yes	99	48.8535	17.91	1.08

Table 5. Comparison of mean scores of test-takers with and without prior computer-based experience

However, test-takers without prior computer-based test-taking experience needed significantly more time for completing the test. For those who had already taken computer-based tests, it took 82 minutes on average to complete the test, while for those without computerised test-taking experience, it took 89 minutes. The independent samples t-test showed that the 7-minute difference in the means was significant ($t_{(183)} = .201$, $p = .001$).

Test-takers' scores were normally distributed as can be seen above in Figure 8, however, the data showing how much time applicants needed for the test were not completely normally distributed, thus the two data sets had different shapes. Therefore, to validate the above findings, the non-parametric Mann-Whitney U-test was also calculated. The Mann-Whitney U-test produced similar results: the group with no prior computer-based test-taking experience needed more time, as shown in Table 6 below:

Is this your first computerised test?		N	Mean rank	Sum of ranks
Time taken in minutes	No	84	72.20	6484.50
	Yes	99	104.56	10351.50
	Total	183		

Table 6. Mann-Whitney time mean ranks

To answer the second research question whether the mode of delivery affects the performance of applicants, the following conclusions were drawn. The applicants who did not have any computer-based test-taking experience and those who had some did not achieve significantly different scores, however, the group with no computer-based test-taking experience needed statistically significantly more time than the other group ($U = 2914.5$, $p = .000$).

4.3 Test-takers' perceptions

The third research question aimed to explore how prospective and present IBS students perceive the E-SETI. In the following sections, quantitative and qualitative data is presented concerning their views. Overall, applicants and current students felt positive about having

taking exams on the computer, the sample gave the test-taking experience an average rating of 2.79 on a 5-point Likert scale (see Figure 6).

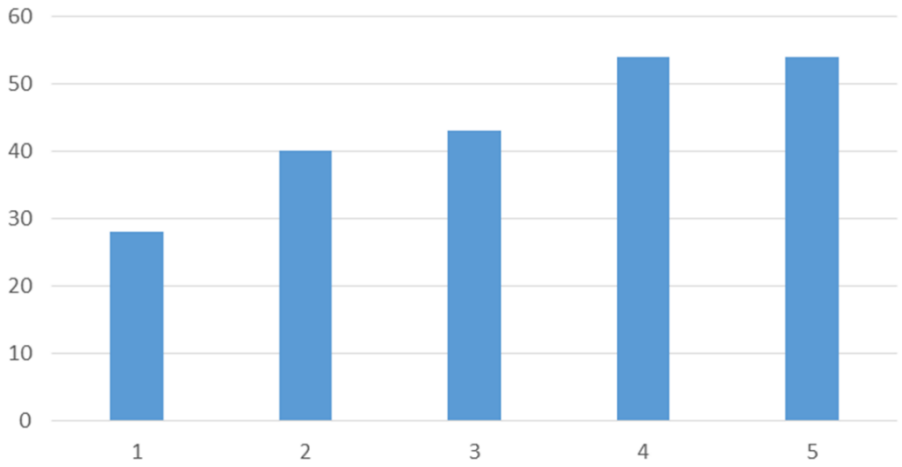


Figure 9. Test-takers' ratings of their computer-based test-taking experience

There seemed to be a significant but very weak correlation between the ratings and the scores ($r = .189$, $p < 0.01$) and a significant but weak inverse correlation ($r = -.260$, $p < 0.01$) between the ratings and how much time applicants needed. Since the ratings given by applicants were not normally distributed, the Mann-Whitney U-test was also calculated. The Mann-Whitney U-test, however, indicated no significant difference either between the ratings of applicants with computer-based experience and applicants without such experience ($U = 3986$, $p = .617$). It was therefore concluded that the performance of an applicant did not significantly affect the rating that they gave for the test-taking experience; meaning it was not only the high-scoring test-takers who rated the test-taking experience as a good one.

Following the live administration of the E-SETI test, 38% of the applicants gave textual comments on the test-taking experience (the text and spelling of the comments was unchanged). Test-takers had quite different views of the experience ranging from positive to negative.

Below, some student comments are presented. From neutral comments, it seems that the mode of delivery is not of critical importance in young test-takers' eyes. As one applicant put it: "I had already done tests like this so it was nothing new, not a good or bad experience". Or, as another student commented: "It is either good or bad. We will see…". For some test-takers, taking the E-SETI on the computer was a positive experience, for instance, one wrote "Did not run into any saving or refreshing page problems". For others it was not particularly challenging, as one of them wrote "because in the computer it is much more easier [*sic*] to write". Another applicant wrote "this test was clear and the system is easy to get adapted to". Overall, having to take the E-SETI exam did not prove to be stressful because "It was comfortable and understandable", to quote another applicant. Navigability was also commented on: "It's faster and you can easily go back to questions you haven't answered yet, or you're not sure about". For others, the traditional paper-and-pencil test is the usual accepted way of taking a test: "I still prefer to have tests on paper because I am used to it". As another applicant added, "I prefer old fashioned paper test, alhought [*sic*] this test was good". For some applicants, computer-based test-taking was a new experience, but they did not mind the challenge. As one applicant remarked, "this was the first time doing an exam on computer. I have to get used to it, but I like it".

Some comments have already been dealt with. For example, one applicant remarked that they did not know "if the system sometimes accepts multiple answers". As a consequence, the instructions have been modified to warn applicants that several solutions may be appropriate and will be accepted by the computer.

A concern that emerged from the textual comments was related to keyboards and typing skills, as it can be seen from the following comment: "There is one another issue that is my typing speed…" Applicants needed time to familiarise themselves with the special Hungarian layout of the keyboard, as one of them commented: "I loved the test but i [*sic*] was not familiar with the keyboard so it took me a while to get used to it. Other than that it was very good" . This appears to be a major problem for some students at IBS, where on one occasion one of the students brought his own keyboard with him from home. Most IBS keyboards work with the Hungarian layout of characters, which is a modified QWERTZ layout with some keys for special accented characters such as ö or í. The non-Hungarian speaking students such as the Chinese student who carried his own keyboard to the campus above find it more tiring to work on IBS keyboards than their own. Typing speed is also a research area discussed in the literature. Russel (1999) found that students with high word per minute keyboarding speed performed significantly higher on computer-based tests. He also pointed out that typing could be a challenge but transferring solutions from paper to computer might be equally challenging.

Another topic that emerged from the comments was the difficulty of reading passages. Test-takers found that reading tasks were different on the computer due to text length. As one applicant commented: "In the reading test [it is] hard to go back always and check the text." The applicant may be referring to the tiresome scrolling that Bridgeman et al. identified in their study (2001). Russel (1999) also states that if the text is longer than one page, performance gets poorer. Another applicant pointed out that note-taking was not allowed: "it is difficult to do reading exercises on the computer because it is not possible to make notes" [*sic*]. One may conclude that the underlying cognitive process in reading on the computer may actually be different from paper-based reading. Russel also mentions that computerised reading tests are more tiring for applicants (1999, p. 8), although that might be changing as generations read more and more online, young and old.

Another issue that applicants often referred to was the time available for taking the test. One applicant, for example, made the following comment: "Maybe, i [*sic*] wanted a little bit more time to fell [*sic*] more free and confident" [*sic*]. Frequent short comments, for instance "No time", emphasise that sufficient time needs to be given for adapting to both the keyboard and the platform.

In summary, the comments suggest that test-takers can easily adapt to having to take the entrance placement test on the computer and have a positive attitude to the computer-based test-taking experience, but they have also pointed out some issues that need to be resolved, for example, by giving more and clearer instructions. Equally important could be allocating them some additional time to familiarise themselves with the keyboard and the functions to help them feel more relaxed at the testing venue, for example, by adjusting the size of letters to avoid scrolling. Applicants' typing and keyboarding skills, however, cannot be improved overnight. That is a vital skill they may have to acquire earlier in their education, perhaps as early as in their primary school (Layton, 2013; Porter, 2015).

5 Conclusion

5.1 Findings

The Moodle platform proved to be a relatively easy-to-use, secure and user-friendly platform for managing the in-house entrance placement test in IBS. All task types were transferable to the platform. With text-based tasks, the concession had to be made that item level statistical information (FVs and DIs) will not be readily available and further data analysis would be necessary. In addition, it was found that sufficient time needs to be given to applicants and administrators before the test opens for logging in and for system managers to help those who have forgotten or mixed up their passwords. Ensuring test security thus has its time requirements. The platform also proved to be versatile. When the school decides to administer the same version, test-takers are seated with one seat empty between them so that no cheating is possible. When items and tasks have already been calibrated, items and tasks of the same difficulty may be randomly assigned to candidates; consequently, they may sit next to each other, with chances of cheating reduced.

The platform proved to be applicant-friendly, although the availability of the Sample Test and detailed instructions for applicants are necessary to ensure smooth operation. Allocating sufficient time for adaptation to the platform appears to be crucial. The platform is definitely administrator-friendly but ideally needs to be used by someone who already is literate in assessment-related issues. Some technical glitches were easily overcome, while others needed time and collaboration with information technology experts. As Pathan (2012) underlined, "cross-disciplinary knowledge" is very important (p. 32). Data processing and analysis is largely assisted by Moodle with the automatic distribution tables and the availability of FVs and DIs. If a test-administrator can competently evaluate the quality of the tasks and items with the help of data available on the platform, these functions can facilitate the compilation and calibration of tests, or with some additional efforts, the building of an item bank.

The support site is quite informative in relation to the construction of quizzes and the interpretation of the statistical information. Without an understanding of classical test theory, however, the tables, charts and figures could mislead users, especially in the case of text-based tasks with multiple items. For example, without an understanding of the Cronbach alpha, or the CIC (coefficient of internal consistency), facility values or the discrimination index, it is hard to imagine everyone can use the Moodle platform for serious testing purposes. Today, there are downloadable sets of questions accompanying textbooks, for instance, as part of course-packs, theoretically facilitating test development. Such packs are ideal for practice quiz construction, nevertheless, it is only with the necessary savoir-faire, that the operation of high-stake exams can be adequately monitored and without appropriate assessment literacy "bad educational decisions" may be made (Green, 2014, p. 6).

Since Moodle stores items or tasks in a convenient format and allows the user to compile tests with relative ease and makes other information concerning item or task content and psychometric characteristics accessible, those with measurement expertise and experience may build an item bank and may compile equivalent test variants with relative ease. Attention must be paid, however, to sample size since only based on the responses of a large enough sample (Henning, 1987) can suitable items or task be selected for item banking. The statistical information attached to the items or tasks needs to be continuously monitored to ensure that

new items and tasks added to the bank contribute to the compilation of fair tests of similar difficulty.

In the business school environment and VLE and with young applicants, their prior computer-based test-taking experience did not affect the language competence grades. This is a locally important finding since IBS applicants come from very different backgrounds in terms of infrastructure. Earlier computer-based test-taking experience did not affect applicants' performance, but it needs to be emphasised that applicants without computer-based test-taking experience needed more time to complete their work. It may well be that the time applicants need depends on their familiarity with the given keyboard layout and their keyboarding speed. Therefore, to reduce this disadvantage at least to some extent, opportunities to practise using the platform need to be given to test-takers for whom this is the first computerised test.

As a result of the action research carried out in the project, the collection of applicant views via the additional questions significantly contributed to the development of the Moodle testing site, the relevant school regulation and the layout of the E-SETI. The clarity and information content of the instructions were improved in order to make the E-SETI more test-taker-friendly. The investigations into the statistical information available on Moodle also shed light on the suitability of the platform. For institutions willing to use this VLE platform, the first step should be to establish a task force that includes both assessment literate teachers and database experts.

5.2 Limitations and further research

As this paper reports action research, it may be criticised for lacking generalisability and replicability (Burns, 2015). The conclusions are deeply rooted in the experience and data collected in the given educational context, a relatively small business school in Central Europe; therefore, they are specific to this international higher education context and cannot be generalized or indeed replicated in another context. However, this report may be useful to similar higher education institutions and colleagues who need to test English language competence with the assistance of a similar, perhaps, VLE-based, more specifically, Moodle-based platform.

Since the findings are based on three single questions and the accompanying test results, they may not be well-substantiated. A multi-item questionnaire would have provided a more reliable research instrument. Findings may be biased because they are based on the voluntary responses of those who understood the questions and had the time and motivation to provide their answers. Eliciting answers from applicants in their mother tongues would probably have meant a wider sample and more sophisticated responses. Also, the questions were closely tied to an important language competence test, so some of the respondents might have opted out for this reason and the responses collected may be not be completely sincere.

The answers to the questions, which partly provide the foundation for the conclusions, were elicited with the help of a computer, therefore, it is quite likely that responses came from participants who did not mind studying, working or even doing tests on computers and were already familiar with computer technology. Accordingly, there may be some positive predisposition among participants towards the use of modern technology in terms of the comments presented in the qualitative section.

The classification of participants into two distinct groups with prior test-taking experience and without it did not allow refined inquiries and conclusions. In the second phase, further questions were added to compute a composite computer-familiarity score and to allow more detailed investigations.

Proofread for the use of English by: Julian Salánki, Department of English Language Pedagogy, Eötvös Loránd University, Budapest.

References

About the TOEFL iBT® test (n.d.). https://www.ets.org/toefl/ibt/about.

Akkreditácós Kézikönyv [Accreditation Manual]. (2020). OH-NYAK. https://nyak.oh.gov.hu/nyat/doc/ak2020/word/Akkreditacios_Kezikonyv_2020.pdf

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge University Press.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bárdos, J. (2002). *Az idegen nyelvi mérés és értékelés elmélete és gyakorlata*. Nemzeti Tankönyvkiadó.

Bax, S. (2003). CALL—past, present and future. *System*, *31*(1), 13–28. https://doi.org/10.1016/s0346-251x(02)00071-4

Barlow, J. P. (1996). Declaration of the independence of cyberspace. *Electronic Frontier Foundation.* https://www.eff.org/cyberspace-independence

Bennett, S., Kervin, L., & Maton, M. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, *39*(5), 775–786. https://doi.org/10.1111/j.1467-8535.2007.00793.x

Bernschütz, M., Dörnyei, K., & Nováky, E. (2016). A Z-generáció a jövőről – empirikus vizsgálat eredményei [Generation Z about the future – findings of an empirical investigation]. In A. Tóth & A. S. Gubik (Eds.), *Magyarország 2025-ben és kitekintés 2050-re: Tanulmánykötet Nováky Erzsébet 70. Születésnapjára [Hungary in 2025 and a look into 2050: Research volume to commemorate the 70th birthday of Erzsébet Nováky]* (pp.63–89). Arisztotelész Kiadó.

Bridgeman, B., Lennon. M. L., & Jackenthal, A. (2001). *Effects of Screen Size, Screen Resolution, and Display Rate on Computer-Based Test Performance. Research Report (RR-01-23)*. Educational Testing Service.

Brown, J. D. (1997). Computers in language testing: present research and some future directions. *Language Learning and Technology, 1*(1), 36–54.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processing whilst completing reading tasks: A mixed-methods eye-tracking and stimulated recall study.* British Council.

BULATS: About the test (n.d.). https://www.cambridgeenglish.org/exams-and-tests/bulats/test-format/

Burns, A. (2015). Action research. In J. D. Brown & C. Coombe (Eds.), *The Cambridge guide to research in language teaching and research* (pp. 99–104). Cambridge University Press.

Casey, M. (2013, July 28). Has technology ruined handwriting? *CNN.* https://edition.cnn.com/2013/07/26/tech/web/impact-technology-handwriting/index.html

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. Routledge Falmer.

Computer-delivered IELTS (n.d.). https://www.ielts.org/about-the-test/computer-delivered-ielts

Dudeney, G., & Hockly, N. (2012). ICT in ELT: how did we get here and where are we going? *ELT Journal*, *66*(4), https://doi.org/10.1093/elt/ccs050

Educational Testing Service (2007). *Test and score data summary for TOEFL® computer-based and paper-based tests: July 2005—June 2006 Test DataTest of English as a Foreign Language™*. Educational Testing Service.

Fulcher, G. (2000). Computers in language testing. In P. Brett & G. Motteram (Eds.), *A special interest in computers: Learning and teaching with information and communications technologies* (pp.93–107). IATEFL publications.

Fulcher, G. (2010). *Practical language testing.* Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An advanced resource book*. Routledge.

Getting your TOEFL iBT® Scores (n.d.). https://www.ets.org/toefl/test-takers/ibt/scores/getting

Gomaa, W., & Fahmy, A. (2011). *Tapping into the power of automatic scoring*. The 11th International Conference on Language Engineering, Egyptian Society of Language Engineering (ESOLEC '2011), Cairo. https://www.researchgate.net/publication/259182018_Tapping_Into_The_Power_of_Automatic_Scoring

Green, A. (2014). *Exploring language assessment and testing – Language in action.* Routledge.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Newbury House.

Hill, P., & Barber, M. (2014). *Preparing for a renaissance in assessment*. Pearson.

History (n.d.). https://docs.moodle.org/37/en/History

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

iTolc website information. (n.d.). https://itolc.hu/otthoni-vizsgazas-folyamata/

iXam website information. (n.d.). https://ixam.hu/documents/10186/13506395/MINDEN+ONLINE+VIZSG%C3%81VAL+KAPCSOLATOS+INFORM%C3%81CI%C3%93_1/87f00aff-c220-40c4-82a3-1fa0d9c3ca57

Jordan, S., Hughes, G., & Bets, C. (2011). Technology in assessment. *Cambridge ESOL: Research Notes, 43*, 2–6.

Korte, W. B., & Hüsing, T. (2006). Benchmarking access and use of ICT in European schools 2006: Results from Head Teacher and A Classroom Teacher Surveys in 27 European Countries. *eLearning Papers*, *2*(1), http://www.ictliteracy.info/rf.pdf/Use%20of%20ICT%20in%20Europe.pdf

Koyama, T., & Takeuchi, O. (2003). Printed dictionaries vs. electronic dictionaries: A pilot study on how Japanese EFL learners differ in using dictionaries. *Language Education and Technology, 40*, 61–80. https://doi.org/10.24539/let.40.0_61.

Layton, L. (2013, October 13). Elementary students learn keyboard typing ahead of new Common Core tests. *The Washington Post*. https://www.washingtonpost.com/local/education/elementary-students-learn-

keyboard-typing-ahead-of-new-common-core-tests/2013/10/13/d329ba66-3289-11e3-9c68-1cf643210300_story.html
Linguaskill (n.d.). https://www.cambridgeenglish.org/exams-and-tests/linguaskill/
Mohammadi, M., & Barzgaran, M. (2010). Comparability of computer-based and paper-based version of writing section of PET in Iranian EFL context. *The Journal of Applied Linguistics, 3*(2), 144–167.
Moodledocs (n.d.). https://docs.moodle.org/37/en/Main_page
Moodle quiz report statistics. (2010). https://docs.moodle.org/dev/Quiz_report_statistics
101/2020. (IV. 10.) Korm. Rendelet 8. § (3). [Article 8(3) of the Government decree No. 101/2020. of 10 April]. Quiz statistics report. (n.d.). https://docs.moodle.org/38/en/Quiz_statistics_report
Pathan, M. M. (2012). Computer assisted language testing [CALT]: Advantages, implications and limitations. *Research Vistas*, *1*(4), 30–45.
Porter, C. (2015, February 4). Common core-linked tests spur schools to teach typing. *The Wall Street Journal.* http://www.wsj.com/articles/common-core-linked-tests-spur-schools-to-teach-typing-1423073700
Results. C1 Advanced (n.d.). https://www.cambridgeenglish.org/exams-and-tests/advanced/results/
Results – When and how to get your score (n.d.). https://www.britishcouncil.hu/en/exam/ielts/results
Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5*(2), 84–94. http://dx.doi.org/10125/25129
Russel, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20), 1–47.
Schwieger, D., & Ladwig, C. (2018). Reaching and retaining the next generation: adapting to the expectations of gen Z in the classroom. *Information Systems Education Journal (ISEDJ), 16*(3), 45–54. https://files.eric.ed.gov/fulltext/EJ1179303.pdf
Solihati, N., & Mulyono, H. (2018). Designing and evaluating the use of smartphones to facilitate online testing in second-language teacher education (SLTE): An auto-ethnographic study. *International Journal of Emerging Technologies in Learning, 13*(1), 124–137. https://doi.org/10.3991/ijet.v13i01.7683
Suvorov, R., & Hegelheimer, V. (2014). Computer assisted language testing. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp.593–613). John Wiley and Sons. https://doi.org/10.1002/9781118411360
Taylor, C., Kirsch, I., Jamieson, J., & Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–274. https://doi.org/10.1111/0023-8333.00088
The IBS Story (n.d.). https://www.IBS-b.hu/about-IBS/v/the-IBS-story/.
The Oxford test of English demo. (n.d.). https://fdslive.oup.com/www.oup.com/elt/general_content/global/ote/demo-v3/#/
TOEFL iBT® Special Home Edition. (n.d.). https://www.ets.org/s/cv/toefl/at-home/
World Bank (n.d.). TCData360. https://tcdata360.worldbank.org/indicators/entrp.household.computer?country=BRA&indicator=3427&viz=line_chart&years=2012,2016#table-link
Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests.* Henry Holt and Company.