COMPUTER-BASED TESTING: A NECESSARY EVIL OR A SENSIBLE CHOICE?

Wojciech Malec
ORCID: 0000-0002-6944-8044
Assistant Professor,
Institute of Linguistics,
John Paul II Catholic University of Lublin
Aleje Racławickie 14, 20-950 Lublin
wojciech.malec@kul.pl

https://doi.org/10.28925/2518-7635.2020.5.10

ABSTRACT

For many students and teachers working in online environments during the current pandemic crisis, the use of computers for educational testing is often an unavoidable predicament. This may be due to the fact that computer-based materials are not merely a useful addition to the learning and teaching resources, but rather the only option available. However, although in some contexts computers may indeed be a significant hindrance to test developers and test takers alike, they actually offer a number of substantial benefits. It is also worth pointing out that, by and large, educational tests delivered through online platforms with the aim of measuring progress and achievement in learning have a lot in common with traditional paper-based tests. This article is thus an attempt at balancing the advantages and disadvantages of computerized testing with a view to finding out whether this mode of testing can be recommended as the preferred choice. Based mainly on a literature review of research and practice in the area of computerized and online educational testing, the paper provides a synthesis of key issues relevant to using electronic devices for the purpose of constructing, administering, and analyzing tests and assessments. In particular, the discussion focuses on the models of test administration, the merits and demerits of computer-assisted testing, the comparability of paper-based and computer-based test scores, as well as selected features of web-based testing systems, such as text-to-items converters, test generators, full-screen delivery mode, automated scoring (and human verification thereof), score reporting, feedback, as well as quantitative analysis of test scores. The article also puts forward some arguments in favour of developing one's own testing application.

Keywords: *computerized testing, web-based technology, merits and demerits of electronic educational testing*

INTRODUCTION

As anticipated over two decades ago by Chalhoub-Deville (1999, p. xv), computers are fast becoming the primary mode of delivering educational tests and assessments. Computer-based testing (CBT) is particularly prevalent in technologically advanced countries such as the United States (Way, Davis, Keng, & Strain-Seymour, 2016; Winke & Isbell, 2017). Nevertheless, in many schools and universities around the world, paper-based testing (PBT) still remains the status quo, which often coincides with the continuing predominance of traditional materials for learning and teaching provided in printed form. The reasons behind the persistent dominance of PBT may include an objective lack of access to the necessary technologies, or insufficient resources for implementing CBT, as well as personal preferences for traditional methods of testing. However, in the last several months, many educators have had to turn to web-based testing as part of a general move to online instruction, as dictated by pandemic restrictions imposed in many regions. It is thus important to review the key issues involved in computerized testing because in switching to this mode of test delivery and development, both teachers and students should be able to make informed decisions and take the best possible advantage of the benefits afforded by technology.

CBT types

One of the key points to bear in mind with regard to CBT is that this approach to testing is not always radically different from traditional PBT. The degree of similarity between CBT and PBT partially depends on the model of CBT administration. These models include the following types: linear fixed-form, linear on-the-fly, multistage, and adaptive (Mills & Breithaupt, 2016).

To all intents and purposes, a linear fixed-form computer-based test (CBT) is simply a computerized version of a paper-based test (PBT). In this model, every test taker receives the same test items, which is typically the case with traditional PBTs. Linear on-the-fly tests also closely resemble traditional PBTs, although they are not exactly the same for every test taker. Rather, when the test is administered, a different test form is created for each examinee by importing items from an item bank. What all linear tests, whether fixed-form or on-the-fly, have in common is that they are equal in length for every test taker.

On the other hand, multistage and adaptive tests represent a major departure from traditional PBTs. Multistage tests, sometimes called semi-adaptive tests (Winke & Fei, 2008; Ockey, 2009) or branching tests (Fulcher, 2010), consist of several testlets (sets of items), and the selection of every next testlet is adapted to each examinee's level of ability, as estimated on the basis of his or her responses to the items in the previous testlet. Multistage tests are thus

adaptive at the testlet level, whereas tests which are known as computer-adaptive tests (CATs) are adaptive at the item level, which means that the selection of every next item is tailored to the examinee's level of ability. Adaptive tests may well be different in length for each individual because tests of this type present the examinees with questions that are gradually more difficult or easier for as long as an accurate assessment of their ability is obtained (e.g. Chapelle & Douglas, 2006).

Despite their many unquestionable advantages (such as greater efficiency and accuracy, or the potential to increase motivation thanks to the fact that the test takers are not presented with inordinately easy or difficult questions), CATs are not really a viable option for classroom-based testing. This is mainly because adaptive testing entails developing extremely large item banks to guarantee item security, and it requires sophisticated IRT calculations (Ockey, 2009). Moreover, from the perspective of the examinees, it may not be quite clear how the final estimate of ability has been obtained on the basis of a test which was different for each individual. Taking into consideration all the technical and practical issues with implementing CATs, Fulcher (2010) pointed out that linear tests actually have many advantages over CATs, particularly in small-scale testing. It is therefore safe to conclude that the type of CBTs which are most likely to be useful in school and university contexts as measures of progress and achievement have a lot in common with traditional PBTs.

One more type of computerized testing should be mentioned, namely web-based testing (WBT). According to the classic definition given by Roever (2001), web-based tests (WBTs) are essentially CBTs written in the HyperText Markup Language, or HTML. Besides the language in which the test content is coded, WBTs differ from CBTs in how the tests are delivered (Suvorov & Hegelheimer, 2014). In the case of WBTs, the mode of delivery is the internet. Although WBTs can be constructed by the test developer as static web pages and uploaded to a hosting server with the aid of an FTP client (Marczak, Krajka, & Malec, 2016), it is now common practice to use an authoring tool for this purpose, either embedded in an online LMS (learning management system) or installed as an add-on extension (Douglas & Hegelheimer, 2007).

Merits and demerits of CBT

A variety of advantages and disadvantages of CBT have been widely discussed in recent years. For example, Brown (2016) provides a review of the drawbacks and benefits of computerized language testing, grouped into physical and performance drawbacks on the one hand, and test designer, test administrator, and examinee benefits on the other. Some of the points addressed by Brown (2016) relate specifically to CAT, some predominantly to WBT, while others to CBT in general. The following

discussion is particularly pertinent to linear WBTs because tests of this type are currently most commonly used in schools and universities delivering online instruction.

Potential problems with CBT

Without wishing to state the obvious, many of the merits and demerits of CBT do not apply universally to all testing situations, but rather have to be evaluated in each given context. A case in point may be the first two physical drawbacks listed in Brown (2016). These drawbacks relate to the availability and quality of equipment. Although access to high-quality hardware and software is undoubtedly essential for CBT, this physical drawback is increasingly waning in importance as various electronic devices are becoming more easily available. Moreover, thanks to the internet, no specially designed software or delivery platforms are needed for computerized testing. In fact, WBTs may be said to be platform independent (Fulcher, 2000), requiring only a standard web browser. On the other hand, equipment-related drawbacks cannot be entirely disregarded. For example, we still cannot take it for granted that no technical problems will occur, such as server failure or browser incompatibility (Roever, 2001). It also bears pointing out that certain types of equipment, for example small mobile devices, may not be the best choice for all testing situations.

The significance of another physical drawback, namely limitations in item types, should also be considered in the light of the particular circumstances. First, testing systems vary widely in the types of items that they allow to construct (and score automatically). Second, testers who need, for example, multiple-choice items only will not be bothered by the fact that a specific testing system does not allow them to create constructed-response items, such as gap-filling. Therefore, limitations in item types will represent a major drawback only as long as the requirements set out in the test and item specifications cannot be satisfied by the testing system being used. Interestingly, CBT may actually allow testers to develop new and innovative item types (as pointed out by Brown, 2016).

The remaining physical drawbacks apply particularly to CATs. As mentioned earlier, the need for extremely large item banks (as a precaution against overexposure of test items) is a serious limitation, which may preclude the possibility of implementing adaptive tests in certain contexts. CATs may additionally require programming expertise and specialist knowledge of IRT. The same limitations may also affect linear computerized tests, but to a much lesser extent.

While most of the physical drawbacks of CBT may well be deemed insignificant in many situations, or they may be relatively easy to overcome, performance drawbacks are more likely to have serious consequences. Physical drawbacks can make it difficult to construct and administer tests in accordance

with the design specifications, whereas performance drawbacks can actually threaten the appropriateness of score interpretations and uses.

The first two performance drawbacks examined by Brown (2016) pertain to the differences between CBT and PBT scores as well as to computer familiarity and anxiety. It has long been recognized that the comparability of CBT and PBT scores may be affected by computer familiarity (see, e.g., Dooey, 2008). This problem can be resolved by allowing examinees to choose their preferred mode of testing, one which is consistent with their regular routine (as pointed out by Maycock & Green, 2005). Given that we cannot assume that all learners have easy access to digital technologies, this may be the best way to guarantee fairness (Papp & Walczak, 2016). On the other hand, this solution may simply not be a viable option in many situations. Moreover, if the test is not administered in the same conditions for all examines, questions arise about the relevance of test standardization (Jones & Maycock, 2007).

An alternative solution to the potential lack of CBT and PBT score equivalence is to give the test takers a tutorial before they take a computerized test. For example, after administering a specially designed CBT tutorial, Taylor, Kirsch, Eignor, and Jamieson (1999) found that inadequate previous computer familiarity did not actually affect CBT scores. In all likelihood, tutorials of this kind will inevitably become redundant for younger learners who are increasingly familiar with computer technology (Chapelle & Voss, 2017).

An interesting observation regarding fears related to computer anxiety can be found in Maycock and Green (2005). Specifically, the researchers found that expectations of lower test performance resulting from lack of computer familiarity actually proved to be false. In other words, the examinees believed that those who were more experienced in using computers would do better on a CBT, but in reality no such effect was detected.

The problem of score equivalence has been investigated by, for example, Wang, Jiao, Young, Brooks, and Olson (2008), who performed a meta-analysis of 36 studies devoted to CBT and PBT administration effects. Their study found no statistically significant differences between reading achievement scores obtained from the two modes of administration. On the other hand, a study by Clariana and Wallace (2002) showed computer-based test delivery to have a positive impact on scores relative to PBT, whereas the reverse was found to be the case in Pomplun and Custer (2005). A research review by Blazer (2010) suggests that score comparability may be conditional on examinee demographic characteristics, computer skills, test and item characteristics, as well as content areas tested. In another study by Keng, McClarty, and Davis (2008), mode effects were more significant when examinees were required to scroll lengthy reading passages on computer screens or performed graphing and geometric manipulations on the computer. In short, the conclusion that may be drawn from a review of the literature is that differences between CBT and PBT performance continue to be a potential drawback of CBT, and test developers must be aware of it.

The next performance drawback pertains to security and cheating. This is indeed a significant disadvantage of most CBTs, even when these are administered in supervised conditions (see Bartram, 2006, for more on levels of supervision in CBT). Supervision during a test delivered online may be conducted through the use of webcams. However, while taking a WBT, students may be able to access the answer key by viewing the source code of the web page (Randolph, Swanson, Owen, & Griffin, 2002). One solution to this problem is to encode the keyed responses as strings of unicode characters (as in *Hot Potatoes*). Nevertheless, an online unicode converter (e.g. www.branah.com/unicode-converter) can be used to easily decode the answers (Marczak et al., 2016). A better solution is to entirely prevent the key from being delivered to the browser during test administration. This can be done with the aid of a MySQL database and server-side scripting (see also Malec, 2018).

The last two performance drawbacks involve reliability and validity. Brown (2016) includes reliability in the list of drawbacks due to the fact that the reliability of CBT scores is often unknown (in truth, this is also the case with many teacher-made PBTs). Apart from that, CBT has the potential to actually contribute positively to reliability, thanks to the consistency of automated scoring (see below). Referring to validity, CBT can result in construct-irrelevant variance, defined by Haladyna and Rodriguez (2013) as systematic error in test scores, i.e. variance which is unrelated to the construct being measured. This happens when, for example, computer literacy is unintentionally measured by a CBT (Davidson & Coombe, 2012). Additionally, validity may be threatened when the testing system imposes methods of assessment which are completely distinct from the learning tasks. On the other hand, CBTs may be able to contribute to task authenticity in language testing (Douglas, 2010) and in this way increase the validity of score interpretations. For example, test tasks may be constructed to include sound and video and thus closely resemble real-life situations, such as telephone conversations, university lectures, or job interviews.

Advantages of CBT

As regards the advantages of CBT, these have been grouped by Brown (2016) into three categories, depending on who they are mostly relevant to, i.e. test designers, test administrators, and examinees. The rest of this section will concentrate on selected advantages of CBT, i.e. those which are particularly applicable to classroom-based testing.

One of the oft-repeated test designer benefits is that CBT scoring is more accurate and reliable. This is indeed true because the sources of unreliability in PBT include various kinds of errors made by humans while scoring the test, even when it is a multiple-choice test, and while transferring the scores to computer records (Fulcher, 2010). In the case of CBT, such errors are practically eliminated because computers are fully consistent in how they

process test data. Thanks to automated scoring, CBT guarantees stability of scoring across different test forms and administrations. Moreover, although human judgment may sometimes be required for scoring constructed-response items, even those where the expected response is relatively short (as in gap-filling), CBT systems are generally capable of ensuring that identical responses from different examinees are never given different scores. In PBT, by contrast, human rating may negatively affect both inter-rater and intra-rater reliability. As reported in Malec (2018), raters are liable to disagree in their evaluation of even short answers consisting of only several words. It is worth adding that automated scoring of extended responses, though still not widely available, is already technically feasible (for a study devoted to automated assessment of essays, see, e.g., McNamara, Crossley, Roscoe, Allen, & Dai, 2015).

CBT advantages also include design and administration flexibility. As mentioned above, the use of computers allows testers to develop innovative task types (such as those involving the use of drag-and-drop), which may additionally incorporate multimedia, glosses (displayed on mouse hover) with explanations of difficult words, as well as various forms of support for examinees with disabilities (Stone, Laitusis, & Cook, 2016). Design flexibility also means that, depending on the system being used, test forms can be compiled by running a test generator programmed to import items from an item bank. Moreover, the items themselves may be created automatically (Gierl & Haladyna, 2013) or by parsing a piece of text, so that it is not necessary to write individual items from scratch (Malec, 2016). Referring to administration flexibility, CBT makes it possible to administer tests either to an entire group of examinees or to selected individuals only, with or without a time limit. Moreover, the same test form can be presented to the test takers either on a single (scrollable) web page or one item per page with a navigation menu (without the necessity of making any changes to the test form itself). Finally, students can be easily allowed to retry a CBT, even an unlimited number of times. By contrast, traditional PBTs would have to be printed out and distributed again.

Another advantage of CBT is related to feedback (an examinee benefit). Among other things, the effectiveness of feedback depends on its timing (Alderson, 2000). Thanks to automated scoring, feedback on examinees' responses can be provided immediately after completing a test or individual items, rather than being substantially delayed, as is the case with PBT. In addition to that, computers can easily deliver answer-specific feedback, which means that it will be presented only to the examinees who submitted a given response, thus enhancing the relevance of feedback. Answer-specific feedback is illustrated in Figure 1, which is part of a WBT administered at www.webclass.co. This figure also shows examples of automated marking with incorrect words being crossed out and the missing ones provided in the form of speech bubbles.

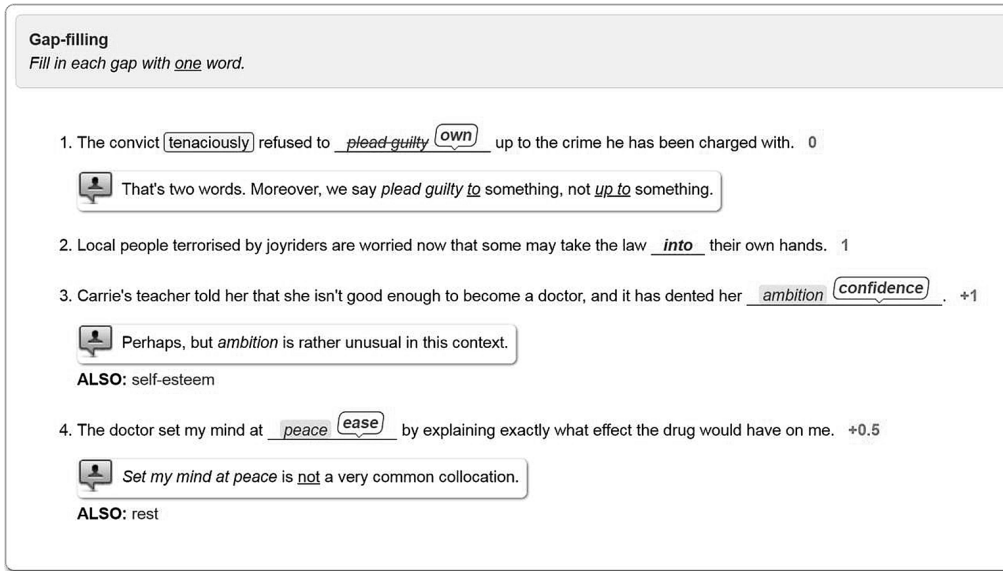

Figure 1. Feedback on responses to computer-based gap-filling items

Finally, it is probably true that computers are most beneficial and convenient for those interested in statistical evaluation of tests and items (as pointed out by Kubiszyn & Borich, 2013). Indeed, for test developers who need to carry out quantitative analyses of test scores on a regular basis, CBT may actually prove to be the only viable option. This is because PBT necessitates entering large amounts of numeric data into, for example, spreadsheets before this data can be used for analysis. CBT systems, by contrast, are often capable of conducting test and item analysis automatically (see also Malec, 2018, for more on this).

Selected features of CBT Tools

While many of the merits and demerits of CBT discussed above generally apply to all computerized tests (such as the anytime/anyplace availability of WBTs), the tools and platforms that are used for test development and delivery also have their own specific strengths and limitations. For example, as noted by Green (2014), the implementation of innovative task types is heavily dependent on the functionality of the software being used. In fact, if none of the available test authoring tools offers a feature that is critically important to the test developer, the only option may be to develop one's own in-house testing system (Shin, 2012).

With respect to test construction, the above-mentioned availability of specific task formats, as defined in the design specifications, may be a decisive factor in determining the appropriateness of a given system. In addition to that, however, there are certain practical issues worth considering. One of them is the amount of time required to create test items and compile them into a test form. This time is often much longer than in the case of PBTs, which may be simply edited in a

word processor and then printed out. Indeed, some testing systems may require dozens of mouse clicks to create a single test question. Generally speaking, it is useful to have some kind of tool that converts user-formatted text into test items. Another issue related to practicality is the possibility of reusing items from previous tests. If items are stored in an item pool, it may even be possible to fully automate the process of compiling test forms by using a test generator.

As far as test administration is concerned, among the issues worthy of note is the option of delivering tests in full-screen mode. When this mode is activated at the start of the test, switching to a different program or changing the browser tab may be additionally disallowed. In case examinees actually do so, their test may be automatically suspended until the test administrator allows them to continue. Another feature that is potentially very useful from the point of view of test administration is the possibility of changing the time limit during the test. In some situations, such as the occurrence of technical problems, it may be necessary to lengthen the time allotted, and if the system disallows this, examines will be disadvantaged.

With regard to test scoring, one of the biggest benefits of CBT compared to PBT is that this scoring may be entirely automated. Nevertheless, machine scoring sometimes introduces problems of its own. For example, some scoring algorithms are not capable of accepting responses containing additional spaces or punctuation marks, while errors of this kind may well be regarded as irrelevant by a human scorer. In addition to that, it is very important that the system should allow the tester to modify the scoring key after the test has been submitted by the examinees. Otherwise, the system would be simply assuming that the test is a final product. In reality, this is often not the case. For example, humans can make inadvertent errors when defining the correct options of multiple-choice items. Moreover, it may be impossible to predict all of the acceptable responses to certain gap-filling items. It is thus essential that the testing system should offer some kind of human verification of the automated scoring. In addition to the above, the scoring procedure may require awarding partial credit to answers which are not completely wrong. Again, the availability of this feature depends on the testing system being used.

Differences between testing systems also apply to score reporting and feedback. For example, teachers who have administered several tests to a group of students in a period of study may or may not be able to access tabulated summary reports containing scores obtained by all of the students in the group, with average final results calculated for each student. As for feedback, it may be of several different types, such as immediate or delayed, simple (i.e. information about the score and correct response) or elaborated (with some additional comments), general (irrespective of the response given) or answer-specific. The marking of the examinees' responses can also be styled in various ways. If these are different from the keyed responses, they may additionally be auto-corrected (as in Figure 1 above). Testing systems may vary substantially in how they implement such features.

Finally, probably some of the most obvious differences between testing systems lie in the way they handle quantitative analysis of test scores. As a matter of fact, the possibilities for test and item analysis offered by educational platforms are often either very rudimentary or simply non-existent. Quite naturally, many teachers may believe that they do not need this functionality. However, those who do may be limited to very basic item and test statistics, not necessarily the right ones. A case in point is the test reliability estimate, which is often restricted to Cronbach's alpha, or the item discrimination index, which is often expressed as the Pearson correlation coefficient. However, in principle, these statistics should only be used with norm-referenced tests, while most of the teacher-administered tests are actually criterion-referenced. The need to conduct the appropriate statistical analyses of test scores was one of the main reasons behind the decision to develop the testing system that is part of WebClass (as discussed in Malec, 2018).

CONCLUSION

Leaving aside the current pandemic situation, the increasing importance of CBT in education is a natural consequence of the popularity and significance of computers in learning. Given that one of the principles of assessment is that test tasks should mirror learning tasks, the switch from PBT to CBT is a welcome change. As pointed out by Winke and Isbell (2017), the present dominance of CBT is a sign of normalization, which has been possible by virtue of the wide availability of low-cost testing software and the fact that test takers are increasingly more familiar with computer technologies. Thanks to this, CBT is no longer perceived as a peculiarity.

Referring to the question posed in the title of the paper, it does not seem to be possible to give a definitive answer because, as has been shown, the choice between PBT and CBT is dependent on a wide array of factors. This paper has reviewed a number of strengths and limitations of CBT, tentatively implying that the strengths compensate for the limitations. Indeed, CBT has a lot to recommend it as a better option than PBT. One the other hand, certain disadvantages of CBT might be somewhat difficult to overcome (such as cheating in online testing). It is therefore up to the individual to decide whether the merits of CBT actually outweigh the demerits.

REFERENCES

Alderson, J. C. (2000). Technology in testing: The present and the future. *System, 28*, 593–603. DOI: https://doi.org/10.1016/S0346-251X(00)00040-3

Bartram, D. (2006). Testing on the Internet: Issues, challenges and opportunities in the field of occupational assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-Based Testing and the Internet: Issues and Advances* (pp. 13–37). Chichester, UK: Wiley. DOI: 10.1002/9780470712993.CH1

Blazer, C. (2010). Computer-based assessments (Information Capsule Vol. 0918). Miami, FL: Miami-Dade County Public Schools.

Brown, J. D. (2016). Language testing and technology. In F. Farr & L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology* (pp. 141–159). London: Routledge. DOI: 10.4324/9781315657899

Chalhoub-Deville, M. (1999). Preface. In M. Chalhoub-Deville (Ed.), *Issues in Computer-Adaptive Testing of Reading Proficiency* (pp. ix–xvi). Cambridge: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1177/0265532209026001060 2

Chapelle, C. A., & Voss, E. (2017). Utilizing technology in language assessment. In S. L. Thorne & S. May (Eds.), *Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment* (3rd ed., pp. 149–161). Cham, CH: Springer International Publishing.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593–602. DOI: https://doi.org/10.1111/1467-8535.00294

Davidson, P., & Coombe, C. (2012). Computerized language assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 267–273). Cambridge: Cambridge University Press.

Dooey, P. (2008). Language testing and technology: Problems of transition to a new era. *ReCALL, 20*(1), 21–34. DOI: https://doi.org/10.1017/S0958344008000311

Douglas, D. (2010). *Understanding Language Testing*. London: Hodder Education. DOI: https://doi.org/10.1177/0265532210373604

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics, 27*, 115–132. DOI: 10.1017/S0267190508070062

Fulcher, G. (2000). Computers in language testing. In P. Brett & G. Motteram (Eds.), *A Special Interest in Computers* (pp. 93–107). Manchester: IATEFL Publications.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education. DOI: https://doi.org/10.1177/0265532210394641

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic Item Generation: Theory and Practice*. New York and London: Routledge. DOI: https://doi.org/10.4324/9780203803912

Green, A. (2014). *Exploring Language Assessment and Testing: Language in Action*. New York, NY: Routledge. DOI: 0.1111/modl.12233

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York, NY: Routledge. DOI: https://doi.org/10.4324/9780203850381

Jones, N., & Maycock, L. (2007). The comparability of computer-based and paper-based tests: Goals, approaches, and a review of research. *Cambridge ESOL: Research Notes, 27*, 11–14.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas Assessment of Knowledge and Skills. *Applied Measurement in Education, 21*(3), 207–226. DOI:10.1080/08957340802161774

Kubiszyn, T., & Borich, G. D. (2013). *Educational Testing and Measurement: Classroom Application and Practice* (10th ed.). Hoboken, NJ: John Wiley & Sons.

Malec, W. (2016). Automating the construction of selected-response items with a text-to-items converter. *CBU International Conference Proceedings, 4*, 864–872. DOI:10.12955/cbup.v4.866

Malec, W. (2018). *Developing Web-Based Language Tests*. Lublin, Pubisher: KUL. DOI: https://doi.org/10.31261/tapsla.7701

Marczak, M., Krajka, J., & Malec, W. (2016). Web-based assessment and language teachers — from Moodle to WebClass. *International Journal of Continuing Engineering Education and Life-Long Learning, 26*(1), 44–59. DOI: 10.1504/IJCEELL.2016.075048

Maycock, L., & Green, T. (2005). The effects on performance of computer familiarity and attitudes towards CB IELTS. *Cambridge ESOL: Research Notes, 20*, 3–8. DOI: https://doi.org/10.1017/S0261444808005399

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59. DOI:10.1016/j.asw.2014.09.002

Mills, C. N., & Breithaupt, K. J. (2016). Current issues in computer-based testing. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational Measurement: From Foundations to Future* (pp. 208–220). New York: The Guilford Press.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *Modern Language Journal, 93*, 836–847. DOI:10.1111/j.1540-4781.2009.00976.x

Papp, S., & Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (Ed.), *Assessing Young Learners of English: Global and Local Perspectives* (3rd ed., pp. 139–190). Cham, CH: Springer International Publishing.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 reading tests. *Journal of Educational Computing Research, 32*(2), 153–166. DOI:10.2190/D2HU-PVAW-BR9Y-J1CL

Randolph, G. B., Swanson, D. A., Owen, D. O., & Griffin, J. A. (2002). Online student practice quizzes and a database application to generate them. In M. Khosrow-Pour (Ed.), *Web-Based Instructional Learning* (pp. 142–149). Hershey, PA: IRM Press.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology, 5*(2), 84–94. DOI: 10125/25129

Shin, S.-Y. (2012). Web-based language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 274–279). Cambridge: Cambridge University Press.

Stone, E., Laitusis, C. C., & Cook, L. L. (2016). Increasing the accessibility of assessments through technology. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 217–234). New York and London: Routledge.

Suvorov, R., & Hegelheimer, V. (2014). Computer-assisted language testing. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (Vol. 2, Chap. 36, pp. 594–613). London: John Wiley & Sons. DOI: https://doi.org/10.1002/9781118411360.wbcla083

Taylor, C., Kirsch, I., Eignor, D., & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219–274. DOI: https://doi.org/10.1111/0023-8333.00088

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24. DOI: 10.1177/0013164407305592

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 260–284). New York and London: Routledge. DOI: 10.31094/2020/1

Winke, P. M., & Fei, F. (2008). Computer-assisted language assessment. In N. Van Deusen-Scholl & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education, Volume 4: Second and Foreign Language Education* (2nd ed., pp. 353–364). New York, NY: Springer.

Winke, P. M., & Isbell, D. R. (2017). Computer-assisted language assessment. In S. L. Thorne & S. May (Eds.), *Encyclopedia of Language and Education, Volume 9: Language, Education and Technology* (3rd ed., pp. 313–325). Cham, CH: Springer International Publishing.

КОМП'ЮТЕРНІ ТА ОНЛАЙН ТЕХНОЛОГІЇ ТЕСТУВАННЯ В ОСВІТІ: ОГЛЯД ОСНОВНИХ АСПЕКТІВ

Войцех Малец, лектор, Інститут лінгвістики, Люблінський католицький університет Яна Павла II, Рацлавицький проспект, 14, 20-950 Люблін, Польща, wojciech.malec@kul.pl

Для багатьох студентів та викладачів, які працюють в Інтернет-середовищі під час пандемічної кризи, використання комп'ютерів для навчального тестування часто є невідворотним безвихідним становищем.

Це пов'язано з тим, що матеріали он-лайн є не просто корисним доповненням до навчальних та викладацьких ресурсів, а, скоріше, єдиним доступним рішенням. Попри те, що комп'ютери ускладнюють як процес розробки тестів так і їх проходження — вони мають ряд суттєвих переваг. Варто також зазначити, що освітні тестування як ефективний спосіб перевірки прогресу та досягнень у навчанні відбувається у електронному варіанті, що має багато спільного з паперовим варіантом. У статті представлені переваги та недоліки комп'ютеризованого тестування з метою з'ясування чи можна рекомендувати такий режим тестування як альтернативу паперовому варіанту. Стаття узагальнює ключові питання, що стосуються використання електронних пристроїв з метою побудови, адміністрування та аналізу тестування та оцінок на основі огляду літератури досліджень та практики в галузі комп'ютеризованих та онлайн-навчальних тестувань. Зокрема, у дослідженні зосереджено увагу на моделях адміністрування тестів, висвітлено переваги та недоліки комп'ютеризованого тестування, виокремлено порівняння результатів тестування у електронному та паперовому варіантах, а також виділено особливості веб-систем тестування, таких як перетворювачі текстових повідомлень, генератори тестів, повноекранний режим, автоматичне оцінювання (та їх перевірка), звітування про бали, зворотний зв'язок, а також кількісний аналіз тестових балів. У статті також запропоновані ідеї щодо розробки власної програми тестування.

Ключові слова: *комп'ютеризоване тестування, веб-технологія, переваги та недоліки електронного навчального тестування*

Received: 02.08.2020
Accepted: 26.11.2020

Passed for printing 14.12.202?
Format 70x108/16. Minion Pro.
Offset paper / Conventional printed sheets 10,5.
Number of copies 300. Order 1-02.