Journal of Research in Curriculum & Instruction
2018, *Vol. 22, No. 3, pp.* 219-225
https://doi.org/10.24231/rici.2018.22.3.219
Printed in the Republic of Korea

Computer-based Language Testing: Potential and Pitfalls

Shin, Sang-Keun[*,†]

** Professor, Dept. of English Education, Ewha Womans University (†Corresponding author, E-mail: sangshin@ewha.ac.kr)*
(Received June 15, 2018.; Reviewed June 20, 2018.; Accepted June 23, 2018.)

ABSTRACT. The computer is positioning itself as the primary base for performing language assessment. Unlike earlier studies that focused on the comparability of paper-and-pencil and computer-based testing, more recent studies on computer-based assessment are actively examining how the testing method may affect performance. This paper briefly highlights some seminal studies on computer-based language assessment, and then looks at some major recent research trends. This paper goes on to introduce computer-based tests that make the most of the advantages of computer technology: a listening test that allows test takers to control the number of times they listen to a recorded passage; a computer-based reading test that enables test takers to make markings on the screen; and a writing test that allow internet search, and a computer-based reading test that lets test takers adjust the setting for typeface, font size, and line spacing. This paper concludes with introducing the latest research trends and with calling for more research on the construct of and access to computer-based tests.

Key words: Computer-based testing, Construct, Authenticity, Test methods, Comparability

I. Introduction

Many assessments these days are Internet-based or in other computer-based formats. Some tests have been developed specifically for CBT or IBT formats, while others have been adapted into those formats based on existing paper-and-pencil assessment tools. TOEFL is a good example. Many graduate schools require students to submit GRE scores and it is now also administered in CBT format.

Although there is not enough time to go into a detailed history of computers in language assessment, the computer-based TOEFL exam represents the most critical turning point. This was the moment when language assessors and educators started to give serious thought to computer-based language assessment. Language teachers could divide computer-based language testing research according to the pre- and post- CBT-TOEFL era. Of course, prior to that moment, too, there were many cases of computer-based tests being used by certain educational institutions at an individual level. Harold Madsen (1986) of Brigham Young University is considered a pioneering researcher in this area. His attempts at computer adaptive testing in the early 1980s were a remarkable research achievement on the vanguard of testing.

Another essential reading for researchers in the field of computer-based language assessment is a 1997 paper by J. D. Brown that summarized the advantages and disadvantages of computer-based testing. As discussed above, the trigger that sent research in this area into full swing was, of course, the introduction of the computer-based TOEFL. There are three key reasons behind the significance of this test. First, it occurred at a global level. The conversion to a computer-based test for no apparent reason forced students all over the world to take a computer-based test if they wanted to study in the U.S. This was true regardless of whether they owned a computer or had any familiarity with computers. The test fees also went up. How much are test fees these days? It's interesting to note that the test fees can vary slightly, depending on the country. In Korea, it's around 190 dollars. Since people generally take the test at least once, the fees are something of a burden.

The TOEFL could only be administered at testing centers equipped with computers. So, a great many people could not sign up for the test because the demand was higher than the availability of spots. This became a serious social issue. One test-taker in New York Times article stated that it would have been easier for a camel to pass through the eye of a needle than to sit for the TOEFL in Korea.

Another major change was the addition of a writing requirement. With the paper-based TOEFL, the Test of Written English was optional, and was administered during even-numbered months only to those who wanted to take it. But when the TOEFL-CBT was introduced, the writing component became required. The writing score was reflected in the grammar score. Initially, since many test-takers were not familiar with typing, they were

allowed to choose between writing by hand and using a word processer, whichever they felt more comfortable with. Finally, language testers can point to another major change, which was the introduction of images in the listening component. Visual images started to be provided to illustrate dialogues or mini-talks.

The reasons for the introduction of the computerized test are unclear. This is because, at least in terms of assessment, practically no new types of items were introduced that capitalized on the advantages of computer-based testing. Basically, it went no further than merely transferring the paper-and-pencil assessment to a computer screen. During that period, the chief focal points of research in this area were the effect of computer familiarity and the comparability of different test methods.

II. Comparability Studies

Before looking at the research from that period, it is necessary to touch on two important concepts. The first concept is construct, which refers to the specific language skill that tests intend to assess in test takers. The construct could be the ability to comprehend an academic lecture, the ability to write an experiment report, or the ability to grasp the main ideas when reading a college textbook, for instance. The other important concept is testing method. The language skills of test takers are assessed through a particular testing method. The issue is that test-takers' performance is affected not only by their language ability, but also by the testing method. Of course, the effect of the testing method is not what the assessment tool is supposed to measure, so the impact of test method effect reduces the accuracy of the assessors' inferences about language skills.

For example, suppose you give a reading comprehension test to two groups using the same reading passage. One group takes a multiple-choice test, and the other group takes a free-response short-answer test. Which group do you think will show higher scores? The group taking the multiple-choice test is probably going to score higher. This shows that, even if you use the exact same reading passage, the results may vary, depending on the testing method used (In'mani & Koizumi, 2009; Shohamy, 1984; Wolf, 1991). When looking at research on computer-based language assessment, researchers always have to consider what skill is supposed to be assessed and how the assessment results are affected by the testing method. And another factor testers have to think about is the quality of the inferences they can make about the test takers' language skills, based on those results.

This section looks at some of the major studies that were carried out in the early days following the introduction of computer-based assessment. First, studies were conducted on how test performance was affected by computer familiarity, or lack thereof. Some may wonder if there could possibly be any test takers unfamiliar with how to use a computer. But the world is full of students less privileged than students in advanced countries. In addition, it should be noted that this study was conducted in 1998, before computer use had become as widespread as it is now.

One of the representative studies examining the effect of computer familiarity is by Taylor et al. (1998). The main question was whether test takers' familiarity with computers affects their performance on computer-based tests, and no effect was identified in the study. More than eleven hundred students at 12 international sites were divided into two groups: one with low-computer familiarity and one with high-computer-familiarity. All received a computer tutorial and a set of 60 computer-based TOEFL items. Before examining the effect of computer familiarity, the researchers adjusted for language ability through a series of ANCOVAs, using TOEFL paper-and-pencil test scores as the covariate. After controlling for language ability, the researchers found no significant relationship between level of computer familiarity and level of performance on computerized language tasks among TOEFL examinees who had completed the computer tutorial.

There's something interesting about the background of the students who participated in this study. The critical question is whether the students who participated in this study are really a representative sample of the whole test taker group. In this study, there is one whole continent missing: Africa. There are also many countries in Asia that are missing. Apparently the sample is lacking test takers from countries where economic development is comparatively behind. This is a very important issue, but practically no research on this question had been carried out since then. It also needs to be noted that many studies have reported that computer familiarity actually does have an effect (Goldberg & Pedulla, 2002; Kirsch et al., 1997).

The second topic of research in that period was the comparability of paper-and-pencil assessment and computer-based assessment. Basically, researchers administered identical tests with identical items in paper-and-pencil and computer-based formats, and investigated the differences. For example, would there be any difference in performance? If a test-taker had to take a very important

reading test, and he or she was allowed to choose between paper-and-pencil and computer-based testing, which format would the test-taker choose?

There are a multitude of research articles on the comparability of reading tests in different formats. The findings are divided about fifty/fifty. Half say that different test methods produce differences in reading speed and level of comprehension, while the other half says there is no difference. The most cited study is Choi et al. (2003). Another highly cited paper is one by Sawaki (2001) that reviewed a number of factors that affect the comparability of paper-and-pencil and computer-based assessment.

A lot of research on comparability has also been conducted in the area of writing assessment. For example, when essay tests are written by hand, are the scores different from when they are written on a computer? Would the results of a writing exam be different in the two settings? The findings are more complicated than you might think. Some studies have found that students with low English proficiency or less familiarity with computers demonstrate higher essay scores when they write by hand (Wolfe & Manalo, 2004).

How about scoring? Do you suppose that when raters are shown the same essay, alternately written by hand and typed using a word processer, they might score the essay differently? Classroom teachers often encounter handwritten student essays that are not exactly legible. The legibility of test-takers' handwriting may lead to a difference in score. It's also possible that the rater's eyesight could be an issue. Try being my age—the older you get, the harder it is to read those answer sheets, with all those tiny characters packed together. The research findings vary. Some find no significant differences (Harrinngton, 2000), and some find higher scores for handwritten essays (Arnold et al., 1990; Bridgeman & Cooper, 1998; Powers et al., 1994; Sweedler-Brown, 1991), and conversely, some find higher scores for word-processed essays (Lee, 2002).

In summary, the research findings on comparability of paper-and-pencil and computer-based assessment are more complicated than we might expect. One thing that is certain is that new assessment methods may also affect assessment results. Although utilization of computers has the potential to create a more authentic assessment environment, it also affects the assessment results in unintended ways. Therefore, Bachman (2000) suggests that there is a need for more research on new assessment methods. He noted that the new task formats and modes of presentation that multimedia computer-based test administration makes possible raise all of the familiar validity issues, and may require testers to redefine the very constructs they believe they are assessing.

III. Studies on the Effects of Test Method Facets

Entering the mid-2000s, researchers moved away from their focus on comparability. This was because computer assessment had caught on as the primary medium for assessment, despite suggestions that much more research was needed. Research now is divided chiefly into three areas. Studies examining the effect of various facets of computer-based assessment on assessment results make up the majority. The test method facets proposed by Bachman (1990) are principally composed of 1) the testing environment; 2) test rubrics; 3) the nature of the input; 4) the nature of the expected response; and 5) the interaction between the input and the response. Among these, quite a lot of research is being done on the element of input.

In particular, several major studies have been done on listening assessment. As presented earlier, visual images had been introduced in the listening section of the computer-based TOEFL. Looking at a blank screen while you're taking a test is also awkward, but in order to increase the authenticity of the listening assessment, photographic material is being presented in the TOEFL CBT. Most test-takers would probably be in favor of showing visual images in listening assessment. Nobody is listening to an academic lecture with their eyes closed. It goes without saying that visual information plays a critical role in daily listening situations.

In any case, if visual material is provided in listening assessment, the key question is: what will happen to the exam results? First, items like the one below asking about location would probably not be presented in the test.

5. Where does the conversation most likely take place?
① hospital
② restaurant
③ police station
④ train station
⑤ lost & found center

***Figure* 1.** Role of Visual Images in Listening Comprehension Test

In truth, this item, which is extracted from 2018 nationwide listening test for the 8th graders, can be considered the kind of question that has no authenticity whatsoever.

In an authentic situation, a person hearing the kind of dialogue spoken in a library is already in a library, so asking them where the dialogue is taking place makes no sense.

Let's look at a more natural question Suppose there's a dialogue about someone purchasing one blanket and two pillows at a store, and the item asks test-takers to choose the correct amount that a character has to pay. In fact, this item is extracted from English section of 2018 College scholastic ability test. Would it make any difference in the exam results whether an image of the store is provided or not?

9. Listen to the conversation, and choose the amount the man has to pay.
① $36 ② $45 ③ $54 ④ $60 ⑤ $63

***Figure* 2.** Role of Context Visuals in Listening Comprehension Tests

What if the visual images that are provided include details related to the correct response? Visual images should be divided into contextual visuals that reveal the situation only and content visuals that involve details related to the correct response.

Suppose a test taker does not comprehend the listening material, yet determines the correct answer by using the information in the visual material. How should the results be interpreted? For example, in the case of an item about a weather forecast for a certain region, if a test taker answers the problem using a map displayed on the screen that shows weather symbols and the temperature of that area, how should the results be interpreted? Now, in this case, how about still images and video-based materials? In the same situation described above involving purchasing goods at a store, what kind of difference would it make in the results if the test provides a still image or a video?

Prior research shows that many students initially look at the visual materials. At some point, though, when they realize that the visual materials don't provide any crucial hints in deciding on the answer, they close their eyes and concentrate only on the audio, because the video is actually a distraction. So, a paper has been published with an interesting title (Wagner, 2007). It asserts that even when video clips are provided, many students don't look at them. In fact, some students show lower scores on a video exam.

At this point, there is a question that must be asked. In a case where students don't look at the video very much, even if it is available, should the video still be provided? Some test-developers may argue that testers just save the expense and not provide it. Others may believe that video should still be provided to enhance the authenticity of the test, even if it does not produce any difference in scores. Suppose that some students will get lower scores if video is available. Should it still be provided?

IV. Studies on Tests Utilizing the Advantages of Computer Technology

The existing test was nothing more than moving the paper test to the computer screen. In recent years, attempts have been made to develop test items that take advantage of computer tests that were not available in paper tests. Let's look at prior studies capitalized on computer technology to assess language skills.

The first study (Shin, 2011) involves listening assessment, specifically the number of times test takers play back the listening material. How many times do you think test takers are allowed to listen to the material in a listening test? Most tests only play a passage once, but some tests play it twice. Some test-takers may think that it should be played just once. Others disagree with them and argue that it should be played twice. Still others may think that test takers should be able to listen to it as many times as they want.

Those who believe that a passage should be played only once argue that, since there is only one listening in actual communicative situations, the material should be heard only once on a test, too. Those who argue that it should be played twice claim that in real-life listening situations, interlocutors know what the situation is and can predict the topic and details of what the other person will say. In a test setting, on the other hand, this information is nonexistent, so the problem should first be presented by initially playing the passage to convey the listening purpose and to set up the situation, and then played again to test comprehension. This, they say, would enable us to better assess real-life listening ability.

In any case, how should testers assess listening tasks based on real-life communication situations that allow multiple listening? For instance, real-life listening tasks include examples such as voicemail messages that can be heard multiple times. To infer that comprehension is poor when the material is played only once could be considered problematic. So, in the test that this researcher created, the material presented in the exam was the kind that, multiple times, could be listened to multiple times, if

needed. Furthermore, the test takers could opt to listen as many times as they wanted. After test takers were divided into high, intermediate, and low proficiency groups, plus control groups at the same levels, the study (Shin, 2011) investigated how many times they listened to a passage and which parts they repeated, that is, whether they listened to the whole passage from the beginning or only to the part that relevant to the correct answer. This paper also looked at the difference in run time for the test, in addition to performance. The results showed that, after controlling for the effect according to proficiency level, there were no differences in the high and low groups. But at the intermediate level, the group that was able to listen multiple times showed much higher scores than the group that could only listen once or twice. Of course, it is not clear what exactly this signifies, in terms of the validity of inferences surrounding the construct being measured, in other words, making inferences about listening ability.

This study clearly demonstrates the advantages of computer-based assessment. This methodology was possible because it was a computer-based assessment. In fact, a wider range of data could be collected, beyond just the scores. The researcher was able to collect various data such as exactly how many times each individual listened to a passage, which part they listened to, and how long it took to respond to the item. Can this kind of information be used to make inferences about a test taker's listening ability? Is it possible to consider the speed of the response, for example, as an indicator of listening proficiency?

An interesting attempt has also been made in the reading evaluation. Lee, 2012 involved a computer-based reading test that allowed test takers to use marking strategies. One of the biggest differences between paper-and-pencil and computer-based assessment is that test takers cannot underline on computer screens or make simple annotations. For test takers who are accustomed to paper-and-pencil tests, not having the option to use marking is reported as an inconvenience. How much do test takers really use marking while taking an exam? Does marking affect the test results? Is it feasible to allow them to use a marking feature in a computer-based exam? With these questions in mind, our team produced a computer-based reading test that had a marking function added. Using so-called "digital ink," test takers could perform underlining and make simple memos. In the study, the researcher compared the quality and quantity of the test group's marking behavior with a control group taking a paper-and-pencil test. This shows an actual example of underlining and various annotations made by students on a computer screen.

Language testers can point to two reasons why this study is significant. First, it attends to the complaints of test takers. It seems that if many test takers complain about a particular inconvenience, the testing agency is obliged to find out whether the complaint is valid. The bottom line is this: Say a person's test score is lower on a computer-based test, and the reason is that they were unable to make use of a vital reading strategy. Suppose they could have taken the same test in a paper-and-pencil format and gotten a higher score by using a strategy like marking. In this case, our inferences about that person's reading ability would inevitably be less accurate, simply because the test had been converted to a computerized format. The distinction between construct and testing method really needs to be carefully considered.

An interesting attempt has also been made in the writing assessment, where the internet search was allowed in the writing test (Jun, 2014). This test format attempts to overcome the limit of timed essay exams which do not allow test-takers to search the Internet and to look up the internet dictionary which they normally do in non-testing situations.

Lastly, an interesting study conducted most recently was related to interface design. As Sawaki (2001) shows, in computer-based reading, speed and level of comprehension are influenced by factors like typeface, font size, line spacing, screen size, and screen resolution. Then, let's consider the configuration of the reading section of the Internet-based TOEFL. What would happen if, before taking the test, test takers could configure those factors according to their own preferences? Lim (2013) divided students into three groups. One group took a test using the default screen settings; another group was able to adjust the settings before the test according to their preferences; and the last group was able to adjust the settings any time they wanted to. The researcher studied the effect of these different conditions on the test results.

These studies are noteworthy in that, by developing test tasks that are similar to target language use tasks, they attempted to identify assessment methods that enable test takers to best demonstrate their language ability. In the words of Swain (1984), testers should "bias for best" in test design.

V. Conclusion

This paper have looked at some recent studies on the effect of computer-based assessment on assessment results. Language testers witness two current research

trends. First, the field is seeing attempts to assess language skills that have not been previously assessed. For example, a group of language testers attempted to measure interlanguage pragmatics through the use of technology (Roever, 2006). Roever (2006) suggests that test tasks themselves can be more innovative. The field can look forward to more research in this area.

Another research topic that is gaining in significance recently involves attempts to utilize computers for scoring. For example, speech recognition technology is being utilized for speaking assessment, and computerized essay scoring is actually being done. In fact, for the writing section of the ibt TOEFL, scoring is done both by humans and computers. Can scoring be adequately performed by computers? Can a computer truly assess ideas and creativity? Indeed, the level of consistency between human raters and computers is very high (Coniam, 2009; Lee, Gentile, & Kantor, 2010). What standards do computers use to assess essays written by us humans? Is it acceptable to continue utilizing computer scoring just because the consistency ratio is high, when language testers and teachers don't understand the method of assessment? In the case of speaking assessment, the issue of computerized scoring becomes even more serious. What is starting to happen is that the capabilities of computerized scoring are actually used to determine the assessment tasks. The Versant Spoken Language Test is a good example.

This paper have looked at various issues related to computer-based assessment. It needs to be noted that technology has limitless potential for the development of new assessment methods (Chapelle & Douglas, 2006). Since new assessment methods inevitably affect the assessment results, language testers and teachers have to consider a number of questions regarding construct validity (Ockey, 2007, 2009). Computers will surely play a greater role in language assessment. It is language testers' job to make sure that the technology, that is, the testing method, does not drive the construct. Finally, language testers and teachers should always bear in mind that not all test-takers have an easy access to computers remember those who are less advantaged than most of us. We have to realize, there may be a student somewhere who cannot take the test because he or she does not have 190 dollars.

References

Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers? A study in scoring hand-written vs. word-processed papers.* Unpublished paper, Rio Hondo College, Whittier, CA.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing, 17*(1), 1-42.

Bridgeman, B., & Cooper, P. (1988). *Comparability of scores on word-processed and handwritten essays on the graduate management admissions test.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44-59.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge University Press.

Choi, I. C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. L*anguage Testing, 20*(3), 295-320.

Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCall, 21*(2), 259-279.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133-167.

Goldberg, A. L., & Pedulla, J. J. (2011). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement, 62*(6), 1053-1067.

Harrington, S. (2000). The influence of word processing on English placement test. *Computers and Compositions, 17,* 197-210.

In'mani, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244.

Jun, H. S. (2014). *A validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test*. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1997). Computer familiarity among TOEFL examinees. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*(1), 4-26.

Lee, S. H. (2012). *The effects of marking on computer-based reading test performance*. Unpublished master's thesis, Ewha Womans University, Seoul.

Lee, Y. W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied*

Linguistics, 31(3), 391-417.

Lim, J. (2013). *The effects of options in typeface, size, and line spacing on computer-based reading test performance*. Unpublisehd master's thesis, Ewha Womans University, Seoul.

Madsen, H. (1986). Evaluating a comptuer-adaptive ESL placement test. *CALICO Journal, 4*(2), 41-50.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24*(4), 517-537.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal, 93*(s1), 836-847.

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*(3), 220-233.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing, 23*(2), 229-256.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. Language Testing, 23(2), 229-256.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*(2), 38-59.

Shin, S. K. (2011). The effects of input repetition and proficiency Level in internet-based listening tests. *Multimedia Assisted Language Learning, 14*(1), 211-224.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1,* 147-170.

Swain, M. (1984). Teaching and testing communicatively. *TESL Talk, 15*(1 & 2), 7-18.

Sweedler-Brown, C. O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic score of essays. *Research and Teaching in Development Education, 8,* 5-14.

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. TOEFL Research Report 61; ETS Research Report 98-08.

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology, 11*(1), 67-86.

Wolf, D. F. (1991). The effects of task, language of assessment, and target language experience on foreign language learners performance on reading comprehension tests. (UMI No. 9124507)

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning and Technology, 8*(1), 53-65.