2019-04-01

A Comparison of Mobile and Computer Receptive Language ESL Tests

Aislin Pickett Davis
Brigham Young University

A Comparison of Mobile and Computer Receptive Language ESL Tests

Aislin Pickett Davis

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Troy L. Cox, Chair
Benjamin L. McMurry
Robert Joshua Reynolds

Department of Linguistics

Brigham Young University

ABSTRACT

A Comparison of Mobile and Computer Receptive Language ESL Tests

Aislin Pickett Davis
Department of Linguistics, BYU
Master of Arts

The option to *bring-your-own-device* (BYOD) to educational settings is becoming more prevalent as mobile technologies are more accessible than ever, yet little research has been done to examine the effect of those devices on language assessment. In this study, participants (n=175) were divided by stratified random sampling into four groups. Using a Latin square design to control for ordering, two forms of a multiple-choice reading and listening exam were administered over two days. On each day, participants took one test on a BYOD mobile device and one on a computer. A repeated measures ANOVA was used to determine the effect that device type had on score. During the administration of the test, the BYOD condition revealed a number of difficulties that would caution against full-scale adoption for high stakes testing, but the test scores on the computer and BYOD mobile version of the exam were not significantly different in either skill area.

Keywords: computer-based language assessment, mobile-based language assessment, ESL

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

The pervasiveness of technology has resulted in a world where nearly everything can be put online, and mobile phones with data plans, essentially functioning as pocket-sized computers, have changed the way people access technology. The language field is no exception. Over the last few decades, computer assisted language learning shifted from being a novelty to a common practice (Garrett, 2009) with computer-based assessments following suit necessitating the need to have different use cases validated empirically. When predicting future directions of assessment technology, some researchers have proposed mobile devices as potential options (e.g. Al-Emran et al., 2018; Chou et al., 2017). This is logical due to the rise in both quality and quantity of such devices. While it is certainly possible to port computerized assessments to mobile platforms, certain factors may impact the adoption of these assessments. Even if a test can be administered on a mobile device, how, when, and why should we do so?

Generally speaking, it seems that new and emerging technologies are more readily adopted in pedagogical situations. Use of new technologies in assessment, especially those that are high stakes, often lags behind. One example of this pattern is the history of the TOEFL, a widespread exam that is often taken as an entry requirement for English university programs. Originally, Educational Testing Services (ETS) administered the TOEFL on paper (PBT) and a bubble sheet was automatically graded (Saadian & Bagheri, 2014). In 1998, ETS began using a computer-based test (CBT); this version of the TOEFL included an essay section in which students could choose to either write by hand or use a word processor (Breland et al., 2004). The introduction of word processors on this high stakes test came several decades after they became popular in other contexts (Haigh, 2006). The internet was not used to administer the TOEFL until 2005, when the internet-based test (iBT) replaced the CBT (Alderson, 2009). As access to technology throughout the world has increased, the PBT has declined in use, but it is still offered when the technological infrastructure is poor (Alderson, 2009). The TOEFL is not the only

example of this phenomenon of technology adoption by society first, followed by pedagogical uses subsequent prior to its adoption in high-stakes assessment. For any high-stakes assessment, it is easier to identify potential reasons for the later adoption time frame the expense of securing and validating the assessments cannot be minimized. Stakeholders were likely concerned about the validity, reliability, and practicality of moving away from the paper medium. Test security, proctoring, and data analysis were also likely among other concerns.

Literature Review

Considering the increasing use of mobile devices in the classroom, the next logical step would be to weigh the benefits and drawbacks of using these devices as assessment delivery tools. Certain tools will lend themselves better to testing certain skills or to testing in different environments, such as the case where TOEFL iBT replaced the CBT but has not completely overtaken the PBT (Alderson, 2009).

Testing Mediums

Sometimes a new technology does a better job than the previous version and sometimes changing assessment methods is unnecessary. When considering mobile technology as a vehicle for assessment, it is important to examine other various potential testing mediums, effects of the medium on the user, test constructs and contexts, and the strengths and weaknesses for testing different skill areas.

Construct and Target Language Use. For any assessment, one of the first steps in the test development process is to define the constructs that the assessment will measure and to decide what administration medium is most appropriate for that construct. According to Bachman (2002), a construct is a hypothesized latent trait of a specific ability, such as listening comprehension, that is not directly measurable but can be inferred through the responses to different items. When defining the construct, it is essential to consider the Target Language Use (TLU). TLU refers to the situations in real life where the examinee will use the specific language ability (Bachman, 2002). An unclear construct can make the rest of the assessment creation process vague and difficult, and the implications of the scores are not

useful. Chapelle (1999) adds that "tests must be evaluated in view of the contexts for which they are intended" (p. 266). This principle includes choosing a testing medium (e.g. paper, computer, mobile) that is appropriate for both the construct TLU and the test taking environment. Certain tools are more suited to measure some constructs than others, so understanding the characteristics of each type of instrument can help make an informed decision about which is chosen.

Confidence and familiarity. Confidence and familiarity with the testing mode (e.g. paper, computer, or mobile) play a role in how successful the experience is for test takers (Davis, 1989). For example, in a 1990 study comparing scores on paper and computer tests, the paper version had higher scores (Bugbee & Bernt, 1990). In contrast, a 2003 study comparing the same things found that the scores were "highly comparable" (Choi et. al, 2003, p. 316). This change in results is likely due to the participants' increased confidence with computers in the 2003. In 2014, another study examined the impact of familiarity with mobile devices on reading comprehension scores and found that participants with higher familiarity did perform significantly better (Chen et al., 2014).

Many of the difficulties with using technology can be eased if users feel comfortable with the administration medium and if there are proctors near who are trained in troubleshooting (Dearnley et al., 2008). When students bring their own device they can increase their confidence and familiarity levels, however, since there is no set model for the proctors to prepare for (Chou et al., 2017) the test experience might not be as standard. On the other hand, when students are using their own devices, they are more likely to know basic troubleshooting on their own.

Low versus high stakes testing. When the results of a test are used in a low-stakes setting, such as formative feedback in a classroom setting, there is much more flexibility in the adoption of different testing modes. For instance, replying to a text message is an authentic TLU and examinees may feel very comfortable and familiar with that task, so creating a classroom-based texting assessment could meet both of those criteria. High-stakes testing, however, relies on the precept that the tests are

standardized (Cox, 2018) and fair to all the participants. That is, anyone taking the test will have the same test administration environment, and this can impede early adoption of new technologies. For instance, if some students had phones with autocorrect and others did not, then using a bring your own device model for TLU assessment of texting would disadvantage on set of students based on technology instead of language ability. When the testing is high stakes, the risks to validity and lack of standardized experiences are amplified (Kenyon & Malabonga, 2001).

Examining Testing Mediums

In order to determine the effectiveness of test instruments, developers and researchers need to examine the similarities and differences between each medium. The various methods and instruments have varying benefits and drawbacks. In addition, it is essential to consider the effect that changing the test instrument has on student scores.

Paper versus Technology-assisted Language Tests (TALT). Paper is one of the most established tools used to administer tests, but how do paper tests compare with TALTs? Each have strengths and weaknesses throughout the process including setting up a test administration, taking the test, and grading and reporting.

For one, paper tests generally do not require a lot of external equipment or electronic tools to administer and the cost is relatively low, however, a physical copy of the test must be created for each new person who takes a paper test which puts more work on the front end of the process to ensure the test is ready, with enough copies made prior to any administration. While TALTs have more technology requirements and increased cost, they can be distributed with greater ease for instance, by hosting the test on a website. In fact, a *bring-your-own-device (BYOD)*or *even use-your-own-device* with remote proctoring allows examinees a familiarity with the testing medium that would have positive affect on their performance.

TALTs can utilize security and accessibility features that are not available on paper tests. It is possible to protect tests by requiring a password to access them or to randomize the order in which the items appear to examinees. Information about the time spent on each item is also usually available, which can alert graders to potential problems with cheating if irregular patterns are noted. Technology can also help make tests more accessible with features such as voice dictation, reading text out loud, and more.

When examinees take paper-based tests, responses range from filling in a bubble on a MC test to writing legibly. With listening tests, speakers must be set up in the room as well and structured in a way that all students can hear equally well. Other than a pencil or pen breaking, however, there is very little troubleshooting involved with paper-based tests. One major drawback, however, is that in our evolving world, the exclusive use of paper-only in language contexts does not reflect the TLU as well. TALTs, on the other hand, require familiarity with the technology that affect users differently (Sawaki, 2001) than paper-based methods. The risk of something going wrong with a website or application can potentially affect users' scores and confidence on a test, and the burden of fixing the problem often falls on the proctor. Furthermore, if BYOD were to be employed, it would be difficult to ensure that unauthorized materials such as dictionaries and translation tools were not available and that no copying of the exams occurred via screen grabs or other means.

Paper-based tests are difficult to enforce time constraints. Even when examinees are told to stop writing, they might continue. And since the whole test is available to the examinee, items lack local independence when time constraints are imposed. When time is introduced, tests become **speeded** and total scores can be confounded (Sawaki, 2001). TALTs, however, can use timing at the item level and users can be prompted to move on so they do not spend too much time on any one question. Even if time limits are not imposed, TALTs allow for prompting examinees to use good strategies by recommending when they should reach different sections of the test.

Grading and reporting scores on TALTs can be faster than paper tests. Many items can be scored automatically without a human rater. This is especially useful in situations such as placement tests and self-assessments where quick results are necessary.

Computer tests versus mobile tests. While mobile devices and computers have a lot of similarities, they are not the same. Mobile devices have a much smaller screen and no external keyboard, but they have a built-in microphone and speaker. Also, they can be carried from room to room easily while desktop computers are often in a more permanent location. Many students own their own mobile devices, thus a BYOD model of high stakes assessment is much more suited for mobile than computer tests. It is not likely that schools will purchase sets of mobile phones to use for tests, while computer labs are frequently included in school facilities. In addition, mobile devices can access the internet through a data plan or a wireless internet connection, while computers do not typically have the option to use a data plan. As of now, there are very few studies that address mobile devices as a medium for language assessment (e.g. Garcia Laborda, et al., 2016; Gordon, 2015) and further research is needed to compare them with other testing devices.

Benefits of Mobile-Based Assessment

When looking broadly at the use of mobile devices as part of a testing experience, there are some clear benefits for users.

Mobility, portability, and availability. First, the mobility, portability, and availability of the devices can result in greater convenience for users and administrators. As already mentioned, test takers could use their personal devices for the assessment (Chou et al., 2017) which would provide institutions with a less expensive alternative to elaborate computer labs that often go unused except for program-wide assessments (Castillo-Manzano et al., 2017). Besides being devices with which examinees might be more comfortable and familiar with using, the constraint of limited computer availability is negated when everyone can simply use what they are comfortable with. Furthermore, the shift from expensive

computers to more cost-effective devices such as tablets, could have a positive financial impact in that the space required for test administration could be more multi-purposed.

Drawbacks of Mobile-Based Assessment

Although mobile devices are useful and accessible in many cases, there are certain limitations to using this technology.

Cheating. In a typical test environment, mobile devices are not allowed to be used during the assessment. When these devices become the mode of testing itself, the lines of appropriate behavior can become blurred. While cheating is also possible on a paper or computer test, the facility of sending messages and searching for information on the internet makes it much easier to be dishonest on a mobile exam. Smaller screens are often held at an angle that makes it difficult for proctors to see, resulting in even more opportunities for cheating than on computers. Proctors for mobile tests need to be aware of the difficulties and differences involved with these kinds of exams. Test developers should consider this issue while designing the format of a technology-based exam and proctors should be aware of potential cheating during the test.

Inconsistency of devices. When mobile devices are used as a testing medium, there is a chance that test participants use different device types. This can introduce problems if they have unmonitored applications and tools on their phone. Also, the speed and quality of devices can differ, which introduces an element of inequality.

Testing Different Skill Areas

Item types and test formats vary depending on the language skill being assessed regardless of the testing medium. Some skills are easier to test and score with technology than others are (Choi et al., 2003). The productive skills are more difficult to record and grade than receptive skills with selected response items, especially when testing a large group of students. Furthermore, there are potential reliability problems introduced when multiple human raters score spoken and written responses. Due to

this difficulty and the need for precise results, this study will focus only on testing listening comprehension and reading comprehension using multiple-choice questions. It is important to recognize the reasons behind assessing these receptive skills on mobile devices and to address and minimize potential negative effects while capitalizing on the benefits.

Reading comprehension. Reading comprehension tests traditionally involved reading a text on paper and producing an answer to an accompanying question. In the 1910s and 1920s, these items were usually short answer, and in the 1930s multiple choice questions started being used as well (Sarroub & Pearson, 1998). Occasionally, essays and oral responses are used to test reading comprehension, but the multiple skills required to complete those kinds of task cause issues with construct validity. In the 1980s, the definition of reading comprehension assessment grew more broadly, and expanded to include tasks such as retelling stories and think-alouds (Sarroud & Pearson, 1998).

With the rise of the internet and electronic devices, some reading contexts have changed. Newspapers, ebooks, instructions, and a plethora of other written genres can now be accessed on a screen. Fonts, colors, and pagination can be altered. The flow of the text may emulate a page turn or readers might simply scroll to the end of what they are reading. The glow from these devices differs from the reflected light coming from paper. In 2001, the International Reading Association predicted that "traditional definitions of reading…[and] traditional definitions of best practice instruction derived from a long tradition of book and other print media will be insufficient" (cited in Coiro, 2003). This quote has been proven to be true, as many forms of reading are now primarily digital and reading using technology is a part of functioning in a language (Singer & Alexander, 2017). Thus, it is natural that many reading comprehension assessments use computers, and even mobile devices, as a presentation tool. In fact, successful assessment on mobile-based reading assessments may provide a better indicator of real-world reading comprehension.

Test development. Mobile devices have some similar and different constraints when compared with traditional tests. Part of creating any reading test is selecting the text students will read, writing comprehension questions, and determining the best way to present them to students. Also, there needs to be an accessible way to present the items to the test taker. On paper and mobile tests, there are choices to make about the presentation. Developers need to consider font size and placement on the page, as well as whitespace and numbering. Computers and mobile devices need to consider the screen size, while paper tests usually use a standard size sheet. Another consideration is the placement of the questions in relation to the reading passage. On paper, the answer sheet is sometimes separate from the reading, but this is difficult to achieve on a computer or mobile phone. Because two distinct pages is not an option, some tests display the questions side by side with the reading and some simply have the questions after the passage. In order to create a successful reading test, it is important to understand the type of device it is administered on and present the items in an intuitive way to the test taker.

Testing experience. Reading on a screen can have physical and cognitive consequences for test takers. One of the primary concerns with reading on a digital screen is eye strain (e.g. Boo, 1997; Choi et. al, 2003; Larson, 1999). Many students who have taken reading tests on computers or mobile devices complain about fatigue and discomfort in their vision after staring at a screen for a long time. In some cases, the students performed worse on digital reading comprehension assessments than they did on paper-based tests (Mangen et al., 2013). One cause of cognitive problems is the scrolling required on longer reading tests given on screens. While a paper test simply requires turning a page, computer and mobile tests often ask users to scroll down a screen. This can cause spatial instability, which has a negative effect on comprehension (Mangen et al., 2013). Therefore, when we create a mobile-based assessment, we should minimize the amount of time students will be looking at a screen and focus on making the text as readable as possible.

Listening comprehension. The presentation of audio on listening comprehension exams changes based on the technology available. The technology used in listening comprehension practice and assessment started with phonograph recordings, then continued on to tapes, and then into more digital formats (Jones, 2008). With access to internet-based listening such as the news, podcasts, and audiobooks on the rise, mobile-assisted listening tools are increasing in popularity. Part of the TLU for listening comprehension involves interpreting audio that is delivered through a mobile device. With these developments in technology in mind, it makes sense to include mobile listening assessment in some listening comprehension exams.

Test development. As previously mentioned, the initial part of the testing process is to identify the construct being assessed. Once there are clear goals for the tests, decisions need to be made about visual appearance and item type. Listening test development addresses many of the same visual concerns as reading test development, such as presentation of images, fonts, and whitespace. However, listening test developers have the unique challenge of finding or developing audio tracks that are appropriate for the target use of the language (Coniam, 2006). In addition, the placement of the audio player is an important decision.

Several types of items are used to assess listening comprehension, including multiple-choice, cloze, and open-ended questions (Cheng, 2004). A key difference between paper and technology is that on screens, students may need to scroll to see all the options. One advantage of a TALT listening test is that there is an option to allow students to play the audio at their own time and multiple times. In order to create a successful listening test, it is necessary to use the chosen medium effectively to present the items in an intuitive way to the test taker.

Testing experience. Listening comprehension assessments are different experiences based on the mode used to deliver them. Paper-based listening comprehension tests require some kind of audio player (or a human reader) that is separate from the test sheet itself. TALT versions of these tests are often

more pleasant for test takers because they can listen to the test on their own time. There can also be accompanying images, like the visuals added to the listening section of the TOEFL CBT that are not present in the PBT. One downside to using computers or mobile devices where students can listen to the audio individually is that it requires headphones. Testing institutions either need to provide headphones to students or expect them to bring their own, which places a new burden on the students. This same problem occurs when students are required to provide their own mobile device. If students forget, or have an inferior instrument, they might be disadvantaged or excluded (Chou et. al, 2017).

Research Questions

The existing research in the field of mobile language assessment is not abundant, and what has been done has left more unanswered questions. It is clear that using a mobile device to take a test is a different experience than a paper or computer test, but does that mean they are less effective? The administration tool can also affect student attitude. Gordon (2015) conducted a comparison study with mobile-based and paper-based multiple choice language assessments. , and found some information about EFL student attitudes towards the different tests forms. He learned that students have positive feelings about mobile assessments, at least in in-class low stakes assessments. In addition to the attitude findings, this study discovered that there were no significant score differences between the mobile and paper tests (p. 30). While this study had 150 participants, all of them were from the same language background (Korean), proficiency level (low), and age range (19-22). Most of the students were highly literate in the use of mobile technology and had expressed positive feelings towards implementing new tools in the classroom. Furthermore, the content of the tests was focused on a single classroom unit rather than overall language proficiency. The more class-specific and casual testing setting could have made using a mobile device feel like a review game rather than a formal assessment. A potential positive effect on student attitude without changing test scores is just one of the benefits offered by mobile devices, specifically.

It is possible that the differences in user experience and proctoring impact test results. Reading and listening comprehension tests with selected response items are some of the most straightforward tests given to ESL students and how the scores are affected based on the type of device used to take the tests.

RQ1: Does device type (computer or mobile phone) have an effect on English listening comprehension scores?

RQ2: Does device type (computer or mobile phone) have an effect on English reading comprehension scores?

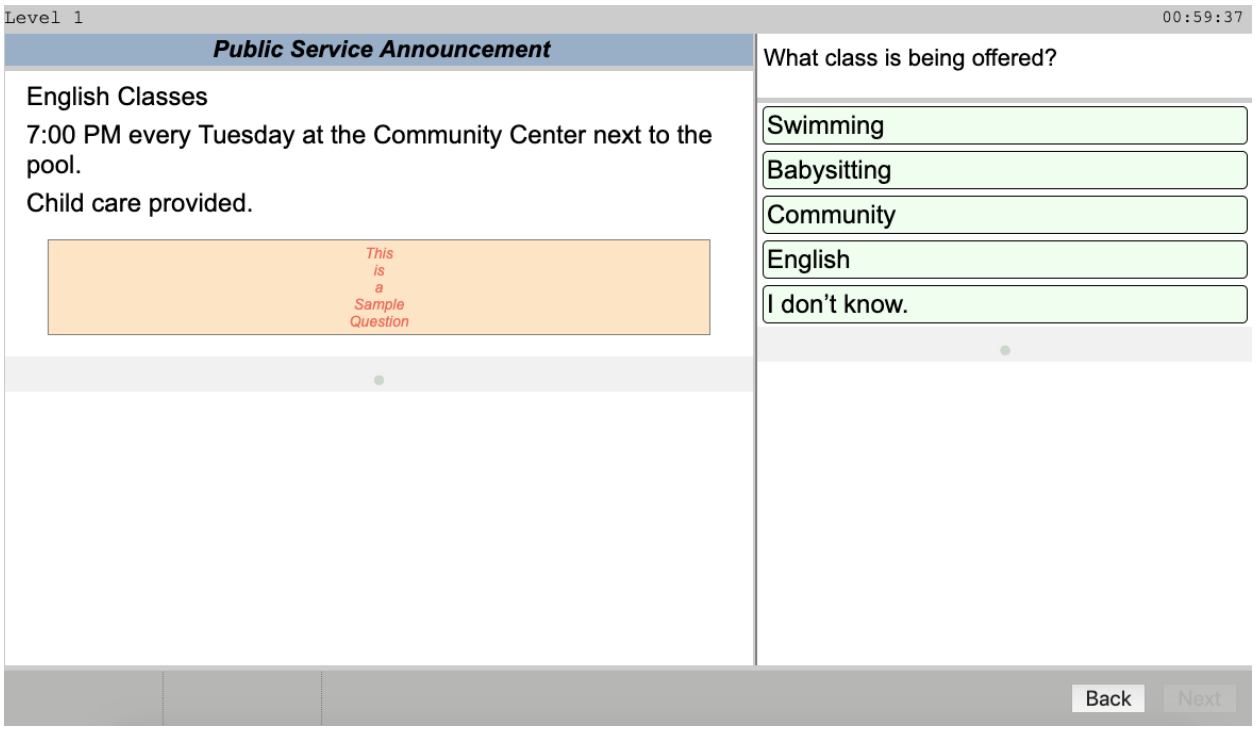
Method

Instruments

For this study, two forms of an internet-based exam for both reading and listening were created. They were hosted on the BYU Center for Language Studies website. This website has a minimal design where test takers can see the question and the multiple-choice options. It is accessed by going to a URL, so users can open it on their computer or their phone. Each skill area had two separate forms (A and B) that were equal in difficulty, reliability, and length. We know the forms are the same difficulty because they were created using pairs of validated existing items. Additionally, all four tests had questions at the intermediate, advanced, and superior level. The items were timed, with more time given to the higher-level questions. During each test administration session, participants took one form of a listening test followed by one form of a reading test. The maximum time available for both tests was 66 minutes, including a few minutes for the transition between test forms and devices.

The appearance of the tests was similar, but not identical, on computer screens and mobile screens. Figures 1 and 2 are images of sample reading and listening test questions on both computers and mobile devices. The computer screen is wider than the mobile screen, and the *next* button is larger.

Prior to the test administration, we conducted a pilot study of the mobile version of the test with 16 students at the ELC. During this pilot, half of the students took a sample listening test and half took a sample reading test that used the administration website and had similar items. After completing the sample tests, we asked students for their feedback about the user experience. Some pointed out that the buttons were difficult to locate on the screen and many expressed that they had to change their phone settings to desktop mode in order to view all the content on the screen. Based on this feedback, we changed the proctoring instructions to address potential difficulties. The students in the pilot study had both Android phones and iPhones, but none of them had difficulty accessing the tests.



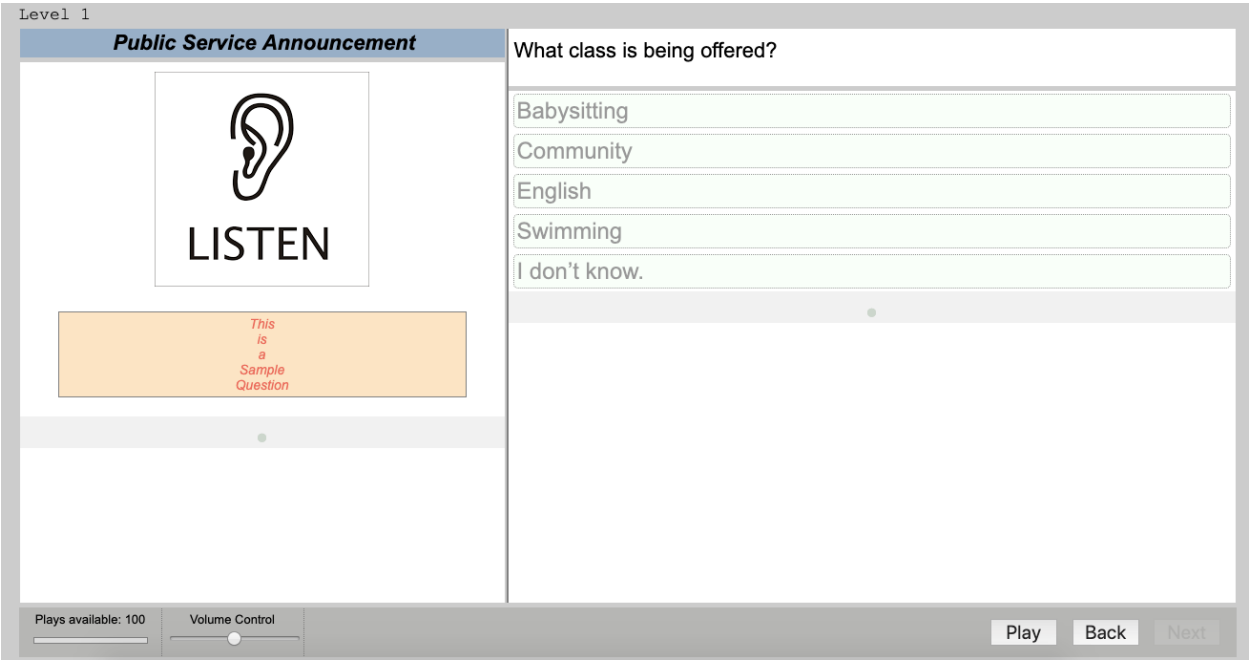


Figure 1. The appearance of the computer screen during the reading and listening tests.

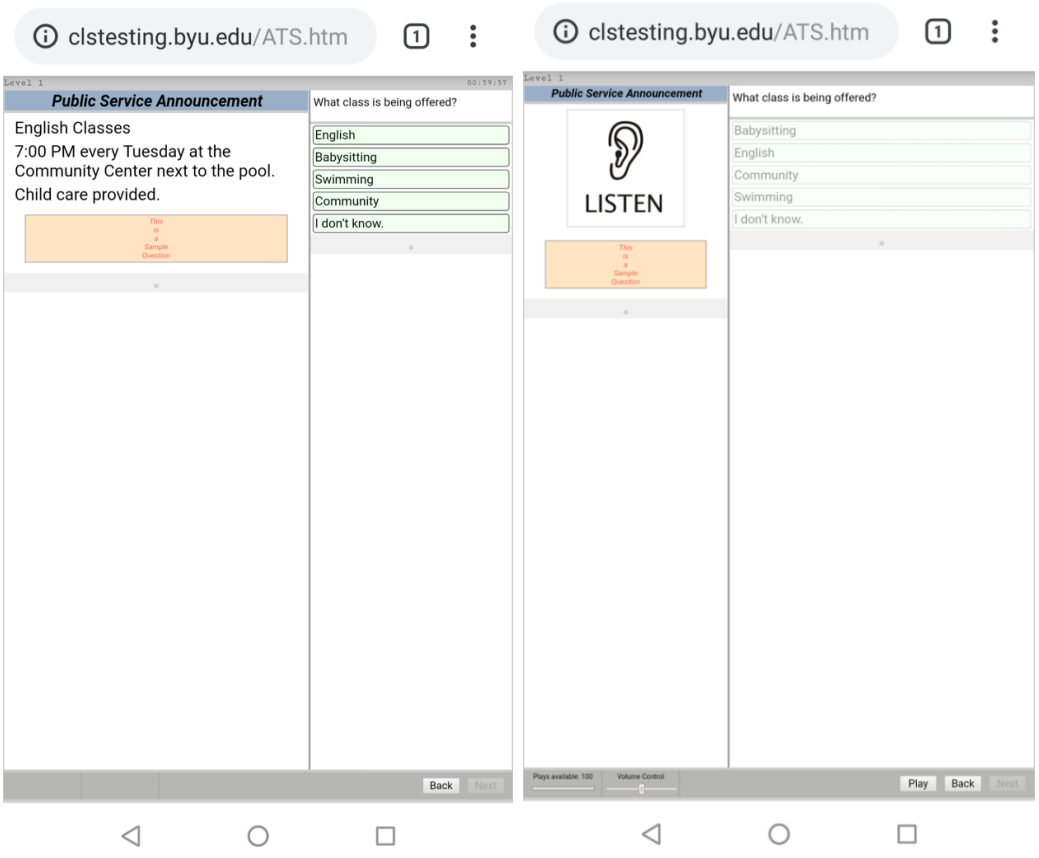


Figure 2. The appearance of the mobile device screen during the reading and listening tests.

Listening instrument. The paired forms of the listening comprehension test both had twelve timed questions. In order to hear the audio, the test taker had to click on a button at the bottom of the screen. However, the time for the question started regardless of whether the audio had been started and when the timer ran out on a question, the test continued to the next item.

Reading instrument. The paired forms of the reading comprehension test both had twenty questions. When the timer ran out on a question, the test continued to the next item. On the longer reading passages, some mobile screens were too small to view the whole thing at once, which required scrolling.

Administration

The test was administered in a Latin square design, where all participants took both versions of the reading and listening tests over two days of testing that occurred 3-5 days apart. One test form was taken on a computer and one was taken on a mobile device. In order to randomize the forms and devices, we assigned each of the students to one of four groups by the class they were enrolled in. Thus each class had an equal number of students in the four groups. This randomization allowed us to have a stratified sample by proficiency level and negated the possibility of a group membership affect related to either class or teacher. The group assignment determined what form they took first and which devices they used. The group descriptions are shown in Table 1. All four groups included both male and female students and a variety of native languages.

Table 1

Group Form and Device Assignments

	Day 1		Day 2	
Group	Test	Device	Test	Device
1	Listen A	Mobile	Listen B	Computer
	Read A	Computer	Read B	Mobile
2	Listen A	Computer	Listen B	Mobile
	Read A	Mobile	Read B	Computer
3	Listen B	Mobile	Listen A	Computer
	Read B	Computer	Read A	Mobile
4	Listen B	Computer	Listen A	Mobile
	Read B	Mobile	Read A	Computer

Prior to the test days, we created cover sheets for each student. The sheets had their names and other identifying information as well as instructions for the test. When students received their sheets, they learned what device they should use for each section of the test. In addition, there were passcodes to access the tests. After the test, these sheets were collected by the test proctors for record-keeping.

The testing occurred over two days. The first administration session was on a Friday in the English Language Center (ELC) computer lab and was proctored by staff and researchers. Each student was assigned to come take the test at a certain time. When they arrived, they got their cover sheets and used their assigned device to go to the testing website, put in the passcode, and begin the test. After completing the first section of the test, they progressed at their own pace to the second section. The proctors monitored this process by comparing the student device usage with the instructions on the cover sheets. Due to a software update that occurred between pilot testing and the test administration, Apple devices did not allow examinees to input the passcode. Proctors and programmers were able to diagnose the problem and fix the code on the website within 90 minutes, however the bug affected all of the participants who had iPhones in the first two testing sessions of the day.

The second administration sessions were between Tuesday and Thursday of the following week. ELC teachers were assigned to take their students to the computer lab and complete the test on a certain day. Everyone followed the same procedure during the second round of testing with a new set of cover sheets.

As an additional precaution, test proctors marked each participant's cover sheet in one of two ways—a check mark if they followed all the directions exactly or a note if something had changed or malfunctioned. This was useful information during the analysis process.

After the two days of testing, the data were collected to prepare for analysis. When participants began the test, they were asked to provide identifying information on the test website. This information was used to keep track of their scores. The scores on all four test forms were compiled into a spreadsheet.

Participants

This study took place at Brigham Young University's English Language Center (ELC). The ELC is an intensive English program that focuses on preparing students for academic success. Additionally, part of the mission of the ELC is research and students agree to be research participants as part of their role at the school. The 175 participants in this study were all students at the ELC and their proficiency levels ranged from novice high to advanced mid on the ACTFL scale. The age range of participants was 18 to 59 years old with a mean age of 25.43 (sd= 6.2). There were fourteen native languages represented—Arabic, Chinese, Creole, French, Haitian Creole, Japanese, Korean, Malagasy, Portuguese, Russian, Spanish, Swedish, Thai, and Turkman. Table 2 shows the demographic breakdown of the participants by native language and sex. At the ELC, students are required to have a personal mobile device in order to access the school's authentication system. Consequently, all the participants had at least a basic familiarity with mobile technology. The two tests were integrated into the curriculum at the

ELC during the semester in which the research took place. All enrolled students took the exam and received a score report and completion grade for finishing the two tests.

Table 2

Breakdown of the Participants by Sex and Native Language

Native Language	Sex		
	Male	Female	Total
Spanish	58	45	103
Portuguese	12	11	23
Japanese	6	10	16
Chinese	8	6	14
Other	9	10	19
Total	93	82	175

Listening. We started with the 175 students who were all assigned to four groups, but had to eliminate 69 who did not meet the protocol guidelines by taking the tests in the order and on the device they were assigned. Of those 69, 6 were absent one of the days, 54 who used the wrong devices (most due to technical problems though some by choice) and 9 with other problems. This left us with a total of 106 participants (see Table 3) divided among the 4 groups with usable data for the listening test.

Table 3

Listening Participant Demographic Information by Group

	Group 1		Group 2		Group 3		Group 4		Total	
	M	F	M	F	M	F	M	F	M	F
Spanish	5	7	11	9	10	9	7	4	33	29
Portuguese	2	0	2	2	2	4	1	1	7	7
Japanese	1	3	0	1	0	4	2	2	3	10
Chinese	2	1	0	0	2	0	1	0	5	1
Other	1	0	1	2	4	3	0	0	6	5
Total	11	11	14	14	18	20	11	7	54	52

Reading. We started with the 175 students who were all assigned to 4 groups, but had to eliminate 53 who did not meet the protocol guidelines by taking the tests in the order and on the device they were assigned. Of those 53, 6 were absent one of the days, 42 who used the wrong devices (most

due to technical problems though some by choice) and 5 with other problems. This left us with a total of 122 participants (see Table 4) divided among the 4 groups with usable data for the reading test.

Table 4

Reading Participant Demographic Information by Group

	Group 1		Group 2		Group 3		Group 4		Total	
	M	F	M	F	M	F	M	F	M	F
Spanish	8	12	12	8	9	5	12	8	41	33
Portuguese	5	0	2	1	0	3	1	3	8	7
Japanese	1	3	1	1	0	0	2	2	4	6
Chinese	3	2	0	0	0	0	2	4	5	6
Other	1	1	0	3	3	2	1	1	5	7
Total	18	18	15	13	12	10	18	18	63	59

Repeated Measures ANOVA

To answer our research questions, we used a repeated measures ANOVA test with four balanced groups of participants. The groups were created by numbering 1 through 4 repeated down a list of all the participants. As shown in Table 1, each group was assigned to take a certain form of the test first and use a certain type of device. This was done to prevent an ordering effect in which the second testing session will result in a higher test score even though the trait being assessed doesn't change (Wilson, 1987). Because of this effect, we administered the tests in a different order to each group of participants. If there is an interaction effect between group number and device type, it could mean that order plays a role in score. For the repeated measures ANOVA analysis, the between-group variable was group membership and the within-subject variable was the device type (mobile or computer).

Results

Scoring

All of the test items were multiple choice, so they were graded automatically. In order to analyze the data, we used Rasch measurement to calculate a person's ability estimate in logits that was converted to a scale in which each logit had a value of 10 with the mean was centered at 50. Next, we listed each student by name and their score on all four test sections along with the type of device they used for each

section. Before proceeding with the statistical analyses, we used the annotated cover sheets to identify students who had had any problems using their mobile devices and updating the information to reflect what actually occurred during test administration and only included those who met the protocol criteria.

Listening Scores

The mean score on the computer version of the test was 51.60 logits (sd= 9.90) and the mean score on the mobile version of the test was 50.35 logits (sd= 11.31). The difference between the mobile and computer tests was not statistically significant with $F(1,102)=.714$, $p=.40$. Furthermore, the interaction between group and device was not significant with $F(3,102)=.755$, $p=.52$. Table 5 shows the descriptive statistics for the listening scores and Figure 3 presents the comparison between computer and mobile means for each group.

Table 5

Descriptive Statistics for Listening Tests

		Computer				Mobile			
Group	N	Mean	Std. Deviation	Lower bound	Upper Bound	Mean	Std. Deviation	Lower bound	Upper Bound
1	22	51.18	7.26	48.15	54.22	48.65	9.42	44.71	52.58
2	28	49.09	10.73	45.12	53.07	50.43	13.22	45.54	55.33
3	38	52.90	10.08	49.70	56.11	49.94	10.97	46.45	53.42
4	18	53.28	10.92	48.23	58.32	53.15	11.37	47.90	58.40
Total	106	51.60	9.90	49.72	53.49	50.35	11.31	48.19	52.50

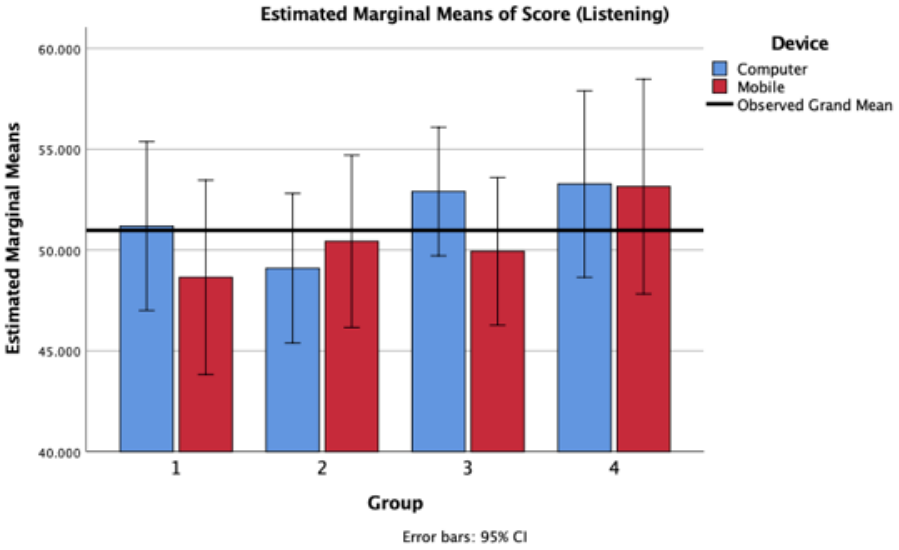


Figure 3. Estimated marginal means of listening scores by group.

Reading Scores

The computer version of the reading test had a mean score of 49.25 logits (sd = 9.71) while the mobile version had a mean of 48.03 logits (sd = 10.36). Table 6 shows the descriptive statistics for the reading tests. The difference between mobile and computer was not statistically significant with $F(1, 118)=1.264$, $p=.263$. Figure 4 shows the estimated marginal means for each group on the reading test. Furthermore, the interaction between group and device was not significant with $F(3,118)=.503$, $p=.681$.

Table 6

Descriptive Statistics for Reading Tests

		Computer				Mobile			
Group	N	Mean	Std. Deviation	Lower Bound	Upper Bound	Mean	Std. Deviation	Lower Bound	Upper Bound
1	36	50.07	6.57	47.93	52.22	49.96	13.80	45.45	54.47
2	28	48.15	7.33	45.44	50.86	46.18	6.32	43.84	48.53
3	22	48.32	7.35	45.24	51.39	48.53	10.55	44.12	52.94
4	36	49.86	14.36	45.17	54.56	47.23	8.69	44.39	50.07
Total	122	49.25	9.71	47.53	50.98	48.03	10.36	46.19	49.87

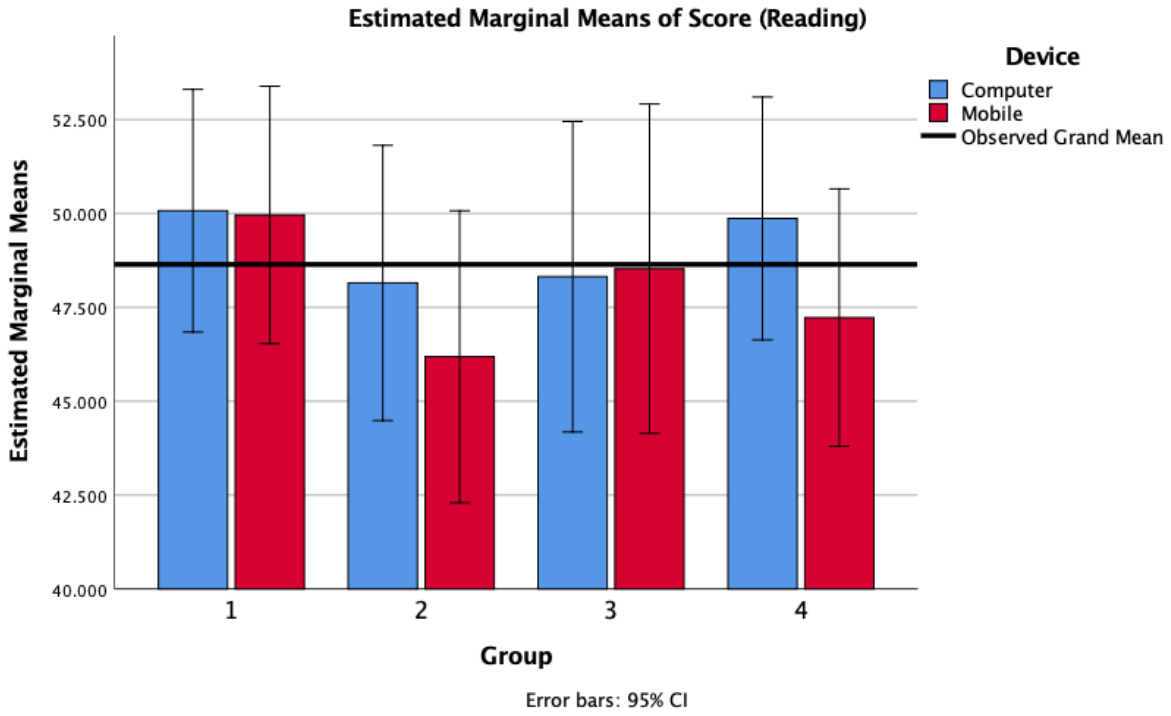


Figure 4. Estimated marginal means of reading scores by group.

Discussion

There were no significant differences between reading and listening scores. We examined the data for an ordering effect, and found that the order did not have a significant impact on the score. These results support the possibility of using mobile devices for assessments proposed by other researchers (e.g. Garcia Laborda et al., 2016; Arthur Jr. et al., 2014). Our score results mirror the results found in the low-stakes mobile assessment by Gordon (2015) because there was no score difference between mobile and another medium, but the results differ in the observed impact on student attitude.

The large number of participants who had to be excluded from the analysis due to was a limitation of this study. Although we conducted a pilot test, there were unexpected problems delivering the actual test. One factor impacting the number of eligible participants on the first testing day was an error in the administration website that caused problems on iPhones. In some other cases, the mobile browser froze and did not function properly. In addition, many participants forgot to bring the required

tools, demonstrating a weakness of the *bring your own device* method of assessment. All of these issues probably reflect an authentic real world test experience more than the pilot test did.

While the scores were not significantly different, the user experiences had some distinct contrasts. As mentioned previously, many users were unable to complete the assigned tasks. Even among the students who were successful, there were complaints about the mobile test.

During the mobile test administration, many participants were frustrated. Some of the most common complaints were about reading on the phone screen and using the smaller navigation buttons. Of the participants who expressed annoyance with the mobile devices, several chose to disregard the assignment and took the test on a computer instead. A few students did not complete the mobile test at all. Participants who did not follow the assigned device requirements or who left questions unanswered were excluded from the data analysis.

Some proctors observed a difference in the experience of administering the mobile and computer tests. Cheating was not a problem that arose with either test type, but general frustration was common with the mobile tests. Trying to help frustrated examinees and resolve their concerns about the mobile test created extra work that was not a problem with the computer test.

Conclusion

Based on the results of this study, there was no significant difference between the scores for reading and listening comprehension tests taken on computers and on mobile devices. With the increase in use of mobile devices in classroom settings, stakeholders including test creators, administrators, and teachers should be aware of the benefits and drawbacks of assessment on these devices. While more research is needed to support our findings, these initial results are informative and beneficial. The following suggestions for research may address limitations we encountered and provide direction for other related research.

Future Research

Future studies should address some of the limitations of this study. One problem we faced was that while the testing website worked on all the devices used, it was not optimized for mobile. Test takers had to use their phone screen as a miniature computer in desktop mode rather than a mobile-specific version of the website. Replicating this study with a mobile app optimized for the devices may provide different results than the web-based delivery used in this study.

Furthermore, all the participants in this study were familiar with mobile technology because it was very ubiquitous at the ELC. In fact, two-factor authentication at the university required a mobile device. Most students in the program came with a mobile device or were able to purchase an affordable one upon arrival in the United States. It would be informative to replicate this study with participants in a different setting where mobile device use is less common. While mobile devices are quickly becoming more common than computers in countries around the world, future research should address the impact of familiarity with a mobile device on scores for mobile-based assessments.

While the tests had construct validity and were proven to be reliable, this study did not have complete face validity. Some participants noticed the inauthenticity of using a mobile device to take the test when traditional computers were visibly available for use.

There were many different brands and sizes of phones represented in this study. This is very authentic but there may be value in conducting a similar study and controlling for device type. Additionally, examining the impact of screen size and operating system on test scores may inform decisions regarding an institution's purchase of tablets or other mobile devices.

Lastly, assessing speaking, writing, and non-selected response tasks should be studied on mobile devices. There are inherent differences in assessing productive skills and receptive skills. Researchers could examine the effect of touch-screen keyboards or even autocorrect on typing timed responses. For speaking, there may be complications in the recording of responses to prompts.

References

Al-Emran, M., Mezhuyev, V., & Kamaludin, A. (2018). Technology Acceptance Model in M-learning context: A systematic review. *Computers & Education*, *125*, 389-412.

Alderson, J. C. (2009). Test review: Test of English as a foreign language™: Internet-based test (TOEFL iBT®). *Language Testing*, *26*(4), 621-631.

Arthur Jr, W., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment*, *22*(2), 113-123.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing,* 19(4), 453-476.

Boo, J. (1997). Computerized versus paper-and-pencil assessment of educational development: score comparability and examinee preferences. *Unpublished dissertation*, University of Iowa.

Breland, H., Lee, Y. W., & Muraki, E. (2004). Comparability of TOEFL CBT writing prompts: Response mode analyses. *ETS Research Report Series*, *2004*(1), i-39.

Bugbee, Jr., A. C. and Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education*, *23*, 87-101.

Burston, J. (2014). MALL: The pedagogical challenges. *Computer Assisted Language Learning*, *27*(4), 344-357.

Castillo-Manzano, J. I., Castro-Nuño, M., López-Valpuesta, L., Sanz-Díaz, M. T., & Yñiguez, R. (2017). To take or not to take the laptop or tablet to classes, that is the question. *Computers in Human Behavior*, 68, 326-333.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254-272.

Chen, G., Cheng, W., Chang, T. W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter?. *Journal of Computers in Education*, *1*(2-3), 213-225.

Cheng, H. F. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*(4), 544-553.

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

Chou, P. N., Chang, C. C., & Lin, C. H. (2017). BYOD or not: A comparison of two assessment strategies for student learning. *Computers in Human Behavior*, 74, 63-71.

Coiro, J. (2003). Exploring literacy on the internet: Reading comprehension on the internet: Expanding our understanding of reading comprehension to encompass new literacies. *The Reading Teacher*, *56*(5), 458-464.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English-language listening test. *ReCALL*, *18*(2), 193-211.

Cox, T. (2018). Standardized Testing in Reading. *TESOL Encyclopedia for English Language Teaching*. 1-6.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

Dearnley, C., Haigh, J., & Fairhall, J. (2008). Using mobile technologies for assessment and learning in practice settings: a case study. *Nurse education in practice, 8*(3), 197-204.

Garcia Laborda, J., Magal Royo, T., & Bakieva, M. (2016). Looking towards the Future of Language Assessment: Usability of Tablet PCs in Language Testing. Online Submission, *22*(1), 114-123.

Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *The Modern Language Journal*, *93*, 719-740.

Gordon, A. (2015). Paper based testing vs. mobile device based testing in an EFL environment: what's the difference? *Culminating Projects in English.* Paper 38.

Haigh, T. (2006). Remembering the office of the future: The origins of word processing and office automation. *IEEE Annals of the History of Computing*, *28*(4), 6-31.

Jones, L. C. (2008). Listening comprehension technology: Building the bridge from analog to digital. *Calico Journal*, *25*(3), 400-419.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other proficiency assessments. *Language Learning & Technology, 5*(2), 60-83.

Larson, J. (1999). Considerations for testing reading proficiency via computer-adaptive testing. Issues in computer-adaptive testing of reading proficiency. Cambridge: University of Cambridge Press, 71–90.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61-68.

Saadian, H., & Bagheri, M. S. (2014). The relationship between grammar and vocabulary knowledge and Iranian EFL learner's writing performance (TOEFL PBT essay). *International Journal of Language Learning and Applied Linguistics Word,* 7(1), 108-123.

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House*, *72*(2), 97-105.

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

Singer, L. M., & Alexander, P. A. (2017). Reading on paper and digitally: What the past decades of empirical research reveal. *Review of Educational Research*, *87*(6), 1007-1041.

Wilson, K. M. (1987). Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language. *ETS Research Report Series*, *1987*(1), i-68.