12-2015

Paper Based Testing vs. Mobile Device Based Testing in an EFL Environment: What's the Difference?

Austen M. Gordon

Paper Based Testing vs. Mobile Device Based Testing in an EFL Environment:

What's the Difference?

by

Austen Gordon

Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree

Master of Arts in

English: Teaching English as a Second Language

December 2015

Thesis Committee:
James Robinson, Chairperson
Isolde Mueller
Eric Reynolds

Abstract

Mobile devices are becoming increasingly more ubiquitous. This trend is especially true with young people. An instructor's job is to best service their students. If there are possible testing means that are available, it is the responsibility of instructors to know if these mobile devices are as capable of performing assessments as traditional paper and pencil tests. It is the purpose of this research to evaluate if there is a difference in actual performance in Mobile Device Testing (MDT) versus Paper Based Testing (PBT) and if there are any perceived differences. Participants (N=150) of university EFL learners in South Korea were broken into groups, two different EFL tests were given, the majority received PBT first followed by the MDT and the remaining performed the tests in reverse order. Upon completion of both tests, the participants completed a survey evaluating both testing mediums. Analysis of Variance (ANOVA), F-tests and t-tests were used to validate the comparability of the two different EFL tests, check for overall correlation and test direct comparisons of one group versus another. The results found that the tests were comparable in the performance of the participants, there was no overall group that had a variance that could be attributed to the testing medium, students perceived no difference in difficulty based on testing medium, and that students actually preferred the MDT method over the PBT. These results indicate that MDT is a viable alternative to PBT due to the comparability in performance and student motivational factors.

Table of Contents

Chapter Page

List of Tables

Chapter 1: Introduction

The growing pervasiveness of technology is a trend that cannot be ignored. In fact, a recent survey puts smart phone penetration among South Korean men and women aged 18-25 at 97% and 98.5%. respectively (EMarketer, 2013). Trends like this indicate that educators should be integrating technology into the classrooms. However, before educators, and specifically language educators, throw away testing procedures they have been using for over one hundred years, there should be research done on whether there is a significant difference in performance due to a change in testing mediums. This is the purpose of this paper. As Bomhold (2013) states "There is much current research on the use of mobile devices and computing in the literature, although few look specifically at the use of application by undergraduate college students"(p. 425). It can be assumed that even fewer studies do this with EFL students. Thus, there is a gap in current literature where this study can lend a hand.

Technology Trends

Computer Assisted Language Learning (CALL) is not a new field, but it has been continually changing to keep up with the technological advances of the modern day. The computer has now become an influential component of second language learning pedagogy (Lai, 2006, p. 2). Computers are used as a means to create language opportunities outside of the class, to facilitate collaborative learning tasks, and to disseminate information quickly and efficiently to students among a vast array of other functions. It is only natural that this evolution continues. The purpose of this study is to measure a new medium of assessment and its efficacy compared to traditional methods. This new medium that will be investigated is the smart device, specifically phones and tablets that use applications to access the internet.

The pervasiveness of mobile technology is a trend that continues at a rapid rate. As a result, a growing number of students are now in possession of smart devices (smartphones or tablets). By implementing smart devices in the classroom, students are afforded the opportunity to use a device they are intimately familiar with, as opposed to a generic (and often outdated) school-issued computer. Often, the smart device may even be more effective (speed-wise in processing) than a computer since students tend to carry a phone that has been issued within the last few years, whereas a school updates its computer labs far less frequently.

In addition, the formality and hassle that comes with moving a class to a computer lab is avoided. Consequently, the use of smart phones creates a more efficient use of time and thus more learning opportunities for students. Instruction time is at a premium, as any teacher will attest to. Any time that can be preserved is extremely valuable. Not only is the students' time managed more effectively, the use of smart-device quizzes also saves the instructor a great deal of time due to the software doing the task of marking the assessments, leaving the instructor with simply double checking or managing any disputes over scores.

Online Assessments and Course Management

There are a wide array of course management sites that are available to educators that provide the teacher a means of creating an online-classroom that can act as a substitute or supplement to their traditional classrooms. Sites such as MyiCourse, RCampus, Schoology, SchoolRack, Edmodo, Blackboard, and Moodle have become increasingly popular over the years. These sites give the instructor the ability to group students according to their classes, create discussions and forums, and create, assign, collect and assess student work. The functionality of these sites keeps getting bigger every few months. These present more avenues for teachers to

engage students in their native, digital world. Adapting to technology is critical to making education relevant to modern students.

There are also sites designed specifically for the purposes of assessments. Many of the course management websites have the ability for assessments, such as Moodle or Edmodo, but that is not their main focus, just a function. Websites such as Socrative or Quiz Revolution are specifically designed for assessments and nothing else. This study will focus on assessments through Socrative, due to its specialized focus.

Since Socrative is the primary tool of this study that represents the Mobile Device-Testing,it is necessary to have a solid understanding of its functions. Diechman (2014) outlines the following description:

> Socrative 2.0 is an online assessment and student response tool tat can be run on any platform that has a connection to the Internet. An AASL Best Website and Best App for 2013, this system was created by a team that is passionate about education, and that passion is obvious in their work. The system can be used as a student engagement tool and as either a formative or summative assessment mechanism. Laptop or computer users can just log in to join their class though a student website. An app is also available for tablets and smartphones. Two separate apps are needed to use the Socrative 2.0 tool: a Teacher App and a Student App that must be downloaded onto each device. As of this writing, both the online website and the tablet apps are free to all users. (p. 72)

Socrative was chosen for this study because of its ease of use, its cost (it is a free service), and its ability to quickly assess prior knowledge where students can be easily engaged and assess their understanding of a lesson.

Teachers' Attitudes

As important as integrating technology and teaching are, there are still a significant number of teachers that are staunch in the face of change. Many teachers are reluctant to adapt their teaching styles or methodologies. The reasons for this can be the lack of knowledge of the new

technologies available to them, a refusal to learn, being stuck in their ways, or being intimidated by a medium that students are more familiar with than them.

Another big reason that is cited as a reason for staying with traditional teaching methods is a lack of trust in new technologies. Trust that it will be able to do the same task as traditional instruction and trust that the results will be fair. This lack of trust is unfounded. No method is perfect, including traditional methods. Any new method will have some growing pains associated with it, but once those are dealt with and teachers and students alike are trained adequately, methods involving technology may prove to be just as, if not more effective than traditional methods.

Research Questions

Is there a statistically significant difference in performance of young adult EFL learners based on assessment methods–paper and pencil based tests (PBT) vs. mobile device tests (MDT)?

Are there significant differences in the perception of the difficulty and preference of testing methods?

Chapter 2: Review of the Literature

This chapter will give a description of previous literature published regarding the central issues of this research paper. Such issues include the field of Computer Assisted Language Learning (CALL) because of its pioneering into the alternative of a paper and pencil testing, smart phones due to their prevalence in South Korea and their use in this study, assessments in general since that is the central issue of this study, and this chapter will close with a detailed discussion of a previous study which this paper will conceptually replicate but with several key changes.

Computer Assisted Language Learning (CALL)

CALL is a field of education that has been gaining notoriety as more and more educational professionals are discovering the educational functions of the technological tools that are available to them. Lai (2006) points out that CALL reduces a lot of the stress and anxiety that learners feel when they are in an intimidating L2 environment (p. 2). He also states that computers create fun games that students get to experience while involved in communicative activities (Lai, 2006, p. 2). Lai concludes that these factors serve to promote second language learning. However, painting a picture of CALL as a laid-back game platform is misleading. CALL is much more than that. CALL involves creating a medium with which students can work collaboratively in an L2 environment. Students can have asynchronous conversations in chat rooms or message boards or even via YouTube. Students can even work together on a project despite not being in the same country or having the same L1. CALL involves removing the restrictions of using L2 strictly in class time. The opportunities to use language are consequently opened up to any free time a student may have. As a result, CALL puts a lot of the learning back into the students' hands. The only real limit to CALL is the instructor's imagination.

In addition to the pedagogical implications of CALL, Mintz and Aagard (2012) explore the idea of using technology as a means of persuasion to create attitudinal/ behavioral changes in a meaningful, educational way (p. 496). Their study focused heavily on the effects of persuasive technology on learners with learning disabilities, but they were able to generalize their conclusions and claim that "Persuasive design principles could usefully be applied to extend and enhance emerging technologies in educational settings" (Mintz & Aagaard, 2012, p. 497). Smartphones are just such an emerging technology and this paper will explore their role in an educational setting.

Smart Phones

In the last 20 years, CALL has been predominantly concerned with computers, as the name would suggest. However, in the developed world we live in today, computers are not the only tool at educators' disposal. One of the most ubiquitous forms of technology in the world today is the smart device (smart phone/tablet). Pew Research Center (2014) states that as of 2014, 64% of Americans own a smartphone, a trend that has been increasing rapidly, up 35% from 2011. No longer do students have to carry around a bulky laptop to be able to effectively use technology in the classroom. Therefore, with the advent and pervasiveness of smart technology and corresponding Wi-Fi friendly environments that universities create, this is a new tool at the disposal of educators. In her article examining the educational use of smartphones, Jubien (2013), of University of Alberta infers that "some of the newest objects in education, smartphones and tablet computers, may be profoundly shaping and influencing educational practice, whether we are aware of it or not" (p. 2).

These mobile devices have already made their way into the classroom and have been used as educational tools in one way or another. Dresselhaus and Shrode (2012) conducted a survey of undergraduates at Utah State University and found that in 2010, "70.2 percent of respondents indicated that they would be likely or very likely to use library resources on smartphones if they

owned capable devices and if the library provided easy access to materials" (p. 84) and that in 2011, "54% of them use their mobile devices for academic purposes" (p. 87). While this information may be slightly outdated due to the rapid changing in the smart phone-educational setting, it still shows that this is an area that is worth inspection and where opportunities exist for greater learning.

Integrating smartphones into the classroom is not a brand new concept. When the smartphone technology was still in its infancy (and almost unrecognizable when compared to its modern day counterparts), it was still being used in the classroom and being received favorably. Milrad and Spikol (2007) conducted a study using smart phones as a tool to deliver curricular content to support communications learning at a university. They concluded from their surveys "both students and teachers are open and intrigued while using every day mobile communication and collaboration tools in education" (Milrad & Spikol, 2007, p. 69).

A more recent example of integrating smart phones into classrooms was done at Georgia Gwinnett College in 2010. In this study, students were given access to class content related flashcards, podcasts, and lesson plans before class. The goal was to use mobile devices to execute the Thayer Method of significant preparation before class. The program was so successful that based on its overwhelming positive feedback, it was to be adopted by additional chemistry courses at the university (Paredes, Pennington, Pursell, Sloop, & Soi, 2010, p. 193).

In another comparative study of mobile technology and education, researchers found that "Mobile devices are highly portable, easily distributable, substantially affordable, and have the potential to be pedagogically complementary resources in education" (Kim et al., 2011, p. 465). These are the primary reasons why these researchers decided to use a mobile, technology-based, mobile learning model in two primary schools in Mexico. They were able to find a "strong positive effect of supplementing regular classroom education with such technology" (Kim et al 2011, p. 478).

Another interesting result was that the effects of mobile devices were “less susceptible to teacher perception or prior experience or even school infrastructure” (Kim et al., 2011, p. 482). In other words, the use of technology in the classroom was able to remove teachers’ bias from the learning experience, while still fostering learning.

Gikas and Grant (2013) looked into students’ perceptions of using mobile devices in educational settings. The major themes found were advantages and frustrations. The advantages included accessing information quickly, communication and content collaboration, a variety of ways to learn, and situated learning. However, students encountered frustration with anti-technology instructors, device challenges, and devices being a distraction. Overall, they found that students found courses that used mobile devices more beneficial because of mobile devices’ ability to help the students engage with the content (Gikas & Grant, 2013, p. 21).

Honesty

The issue of assessment performance with using MDT, a form of Computer Based Testing (CBT), is the central issue of this paper, but there are ancillary benefits and uses associated with CBT as well. One such advantage is the honesty level achieved in CBT. According to Booth-Kewley, Larson, and Miyoshi (2007), the computer-based surveys produce a social situation that reduces inhibition in respondents (p. 471). This information proves that there is worth in computer or mobile-based reporting and surveying in addition to the benefits associated with assessments.

Assessments

The vast array of CALL applications is beyond the scope of this study; instead we turn now to the area of investigation within CALL- assessment. As long as students have been attending class, teachers have been evaluating them. A current trend is the emphasis on quantifying student assessments and assigning objective values to justify grading. This is a serious matter for both

teacher and student. Therefore implementing computers (and in the case of this study- mobile computers) as a means of assessment was not done without forethought, there have been numerous studies investigating the efficacy of this method of assessment compared to the traditional pencil and paper method.

As we have moved from the general CALL, to a more specific CBT assessment, we now move to research even closer to the focus of this study. Choi, Sung Kim, and Boo, (2003) performed a study comparing the results of participant responses on PBT and CBT. Their focus was much broader than the intended scope of this study, but one of their main conclusions was that there was "comparability of the subjects' scores across CBT and PBT modes" (Choi et al., 2003, p. 316). It is the goal of this study to create PBT and compare them to the results of the technological cousin of the CBT, the Mobile Device Test (MDT), in an attempt to confirm similar results. Choi et al.'s study utilized a question bank and divided questions between the tests from that bank.

The previous study focused on linguistic features heavily, which will not be thoroughly investigated in this study. The research done in this paper will seek to confirm Choi et al.'s conclusion, but have significantly different methods. One difference will be that their study used two different groups doing the same test in opposite order; this study will involve four different groups doing different tests at different times. This will be done for a few reasons. First, students will not be able to recall test questions as they did in Choi et al.'s study. Second, by having four groups instead of two, there essentially is a replication of this same study within itself in another classroom with a different teacher. If results are as expected, this will help reduce the variable of the instructor. Third, there are more opportunities for within and between-group comparisons. However, the biggest difference to note is that instead of utilizing CBT, this study will employ MDT.

Another similar study to this paper was done by Dosch (2012), where he examined computer-based testing in nurse anesthetists national certification exams. He found that there was no difference in the pass rate between students who wrote a paper-based test (PBT) and those who wrote a computer-based test (CBT), reasoning that students of high ability are indifferent to test modes (p. 63). However, Dosch (2012) does state that when writing computer-based tests "students may perceive CBTs more negatively, elevating anxiety and perhaps decreasing scores" (p. 63), and that "Computer anxiety may decrease the capacity of working memory, thus functioning as extraneous cognitive load" (p. 63). Therefore, there are concerns with CBT. Dosch (2012) reasons, however, that the participants who experienced these negative cognitive effects were the students with little CBT experience. Therefore, one could conclude that these ill-effects could be avoided with proper training.

Similar to Dosch, Escudier, Newton, Cox, Reynolds, and Odell (2011) conducted a study with dental students in a high-stakes test they needed to pass their course, administering the test half in paper and half in computer format as well as measuring the participants' attitudes towards the formats. They found that no significant difference in test results attributable to the test formats (Escudier et al., 2011, p. 446). If there was any advantage it was in the online test (Escudier et al., 2011, p. 446). Additionally, they found that student attitudes were very favorable towards online testing, specifically: "Over 70% of students rated the online test as acceptable, and 90% felt that the online format did not disadvantage them, even in a summative and high stakes examination" (Escudier et al., 2011, p. 447).

These two studies from Dosch (2012) and Escudier et al. (2011) illustrate the importance of preparing the participants properly to use new testing mediums. If inadequately trained or if participants feel uncomfortable around the technology being used, there may be negative

perceptions which can translate into poorer performance, as seen in Dosch's (2012) study. On the other hand, when perceptions are favorable, as in Escudier et al.'s (2011) study, the results can be equally positive. These were lessons kept in mind while designing the research done in this paper, and while preparing the instructors who proctored the research.

Likewise, Ockey (2009) has a more positive stance when it comes to language testing via CBT, as discussed in his research on computer-based testing in assessing a second language. He suggests that "computer technology has made it possible to better control how these tasks (multiple-choice, short-answer, or matching tasks) are delivered to test takers and the processes that test takers must use to complete these tasks" (p. 840). He later concludes that:

> CBT has had a great impact on language assessment practices, including affecting the way language ability is defined and consequently assessed, making possible scoring of constructed responses to writing and speaking prompts more reliable, more practical, and almost instantaneous, and paving the way for the development of more authentic task types than has been realized with traditional paper-and-pencil tests. (Ockey, 2009, p. 845)

However, Ockey (2009) does warn that computer-based testing still needs to improve its security and that current measures of evaluating meaning and feeling in speaking and written discourse are not yet adequate (p. 845).

Another study that investigated the comparison between computer-based testing and paper-pencil testing performed by Chua (2012), found that "The computer-based testing is more reliable in terms of internal and external validity and significantly reduced testing time and developed stronger self-efficacy, intrinsic and social testing motivation in the participants" (p. 1584). There was a positive finding in computer-based testing motivation, although this motivation did not lead to improved performance. The study concluded with the assertion that "CBT can reliably replace the PPT in testing" (Chua, 2012, p. 1581). This type of conclusion shows that there is value in using computer testing, be it a desktop computer or a handheld device, in the classroom. The study also found there

was no significant difference in the performance between the two testing mediums and consequently that there was not a relationship between the increased motivation and actual test performance (Chua, 2012, p. 1584). Chua (2012) and Escudier et al. (2011) are a few examples of which this research paper aims to replicate and see if the same conclusions can be found when implementing language learning and while using smart devices.

Chua and Don (2013) did more than one study on the topic of CBT and PBT, and was able to produce similar results in his second study. By implementing an achievement test, psychological test, and motivational questionnaire in a Solomon-four-group design, Chua found there were no significant differences in performance (Chua, 2013, pp. 1892-1893).

Chua was not the only researcher to find that testing through technological mediums produced positive feelings towards non-traditional methods. Karadeniz (2009) investigated the relationship between achievement and perceptions of students while using web-based and mobile-based assessments. After the first week, there were significant differences in the achievements scores and the students had positive perceptions of the web and mobile-based assessments due to their ease of use, as well as their comprehensive and instant feedback (Karadeniz, 2009, pp. 988-989). In fact, the paper-based tests were the least favored testing medium (Karadeniz, 2009, p. 989).

Taking into account the language learning aspect of Computer-based testing vs. Paper-based testing with a comparative study in Korea, Jung (2014) measured differences in satisfaction between testing methods. The findings were that “learners tend to consider new ways of studying English by adopting new technologies. That is, learners are more likely to adopt innovative methods than traditional ones as they become more IT-savvy” (Jung, 2014, p. 112). More specific to this study, Jung (2014) even found that ubiquitous learning (learning though smart phones) is increasing in popularity

for English education among Koreans (p. 113). One of the goals of this paper is to measure attitudes in testing mediums, so findings like Jung's can be verified.

A similar type of study was carried out by Yurdabakan and Uzunkavak (2012) in Turkey. Their study involved primary school students and measuring their attitudes towards Computer-based testing. Students generally held positive attitudes towards CBT and they even went further to distinguish that there was no significant difference between boys' and girls' attitudes (Yurdabakan & Uzunkavak, 2012, p. 183). Although this study did not implement language learning or smartphones, its findings establish a baseline that CBT are gender-bias free and should not produce different results across gender.

Jamil, Tariq, and Shami (2012) took into account teachers' perceptions of using computer-based assessments. They found that the vast majority of teachers surveyed were 'highly inclined' to use CBT to enable them to assess large groups of students in less time (Jamil et al., 2012, p. 374). Surveys also revealed that CBT examinations facilitate improved student comprehension (Jamil et al., 2012, p. 374).

Bennet (2012) wrote a paper that illustrates the benefits of adopting a more modern testing medium (like MDT). Among the reasons put forward for this are that paper-based testing was designed for a paper based world, and as society moves further and further away from paper-based jobs and tasks, computer/device-based testing is closer approximation of what students will see in the real world (Bennet, 2012, pp. 6-8).

Furthermore, computer-based testing offers improved measurements and precision (Educational Testing Service, 2011). The precision of a computer-based test can be adaptive, meaning that the questions given are not necessarily fixed, and can therefore adapt based on the students' previous responses. This gives the possibility for the computer to adjust the test-taker's performance

estimate and provide a question that can match their level. This type of test can allow instructors to still gauge a student's comprehension, without having poor performers feel overwhelmed and depressed by poor scores. Finally, convenience is a major factor in the trend towards computer/ mobile device testing precision (Educational Testing Service, 2011). Self-proctoring through time limits on devices, distribution and collection of test papers is avoided, and teacher proctoring is limited due to random question and answer order are just a few of the benefits of computer/ mobile device testing. The benefits are not limited to the instructor, the students receive immediate scoring and can have their performance mapped through the educational tools used.

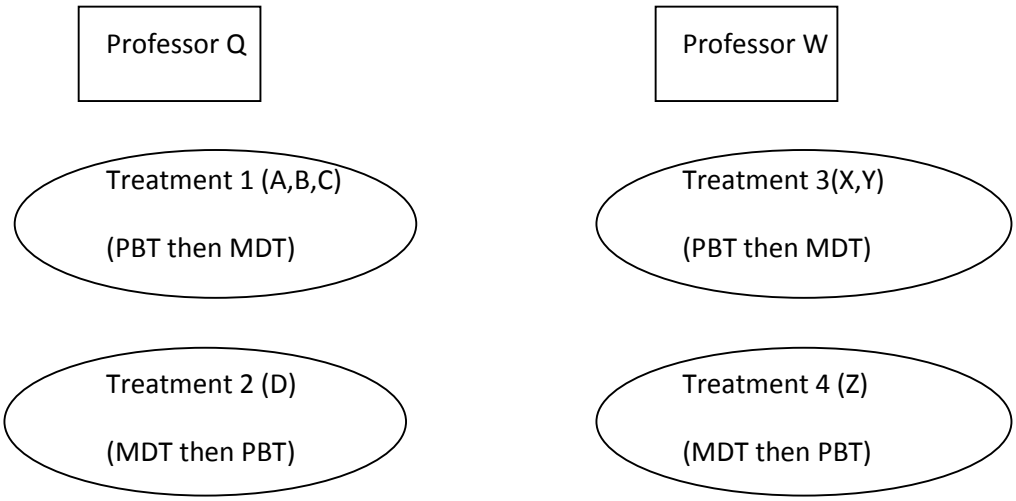
Researchers are not the only ones taking notice of the potential of this assessment platform; the U.S. Department of Education has even recognized this area for years. In the report on computer-based testing published by the National Center on Educational Outcomes (University of Minnesota, 2010), many advantages of CBT and initiatives that the U.S. government have made were outlined. Notably, it states that since the early 2000s, "CBT seems to have advantages over paper and pencil testing" (University of Minnesota, 2010, p. 1). One of the U.S. government's big initiatives, the Race to the Top Assessment Program, was designed to encourage the development of CBT. The theme regarding CBT were overall positive, the only major concern expressed was the risk of using a technology that could be inaccessible. However, the recent proliferation of smartphones had not yet begun its meteoric rise at the time of that paper's publication, nor smartphones' ability to serve as a computer-substitute.

Chapter 3: Methods

This chapter will describe the methods in which the data were collected and analyzed. A pool of 150 freshman students from a South Korean University were the participants. They were from seven different classes. The classes are all the first year required language classes. All participants were given two tests–a paper based test and a mobile device based test. The tests were comprised of 20 questions, all multiple choice, based on the language lessons they received, and counted towards their grades. The reason for two tests and seven groups was to allow for analysis within groups and between teachers. The goal of this grouping was to gather more meaningful data. Upon collection of the data, ANOVA, t-tests, and F-tests were used to measure the variance within and between groups to test the null hypothesis.

Participants

The participants that were used in this study were 150 Korean university students from a University in South Korea. The students were low level English speakers with limited production capabilities and little training in English. The students were in a general education English class (a mandatory language requirement class). Their ages were from 19 to 22, the vast majority were 19. The participants were from seven different classes, but had one of two professors, thus ensuring comparable class format and teaching input. The classes of Professor Qwere Groups A, B, C, and D, while the classes of Professor Wwere referred to as Groups X, Y, and Z. The participants of each professor were divided into two assessment groups, with some of the professor's groups receiving the Mobile Device based test (MBT) first and the paper-based language test (PBT) the following week. Conversely, the remaining treatment had the PBT first followed by the MDTthe following week. Here is a visual representation:



Materials

Question bank. Two tests were developed. They consisted of 20 multiple-choice questions that were designed to measure the participants' comprehension of the material from their general education English class. The subjects covered in the 2 weeks of study were restaurants and shopping. The questions were designed to maintain comparable difficulty. These questions can be seen in Appendix A and B, the copies of tests.

Paper test-PBT. A traditional pencil and paper test was given to each participant (Treatment 1 & 3 first, Treatment 2 & 4 second). The test time allotted was 15 minutes, anecdotally both teachers reported that the vast majority finished well before the allotted time limit. The test was proctored by a professor of same Korean university (professor Q or W), and the tests were marked by me. The questions were all low-interference questions that required little or no subjectivity, thus ensuring no loss of reliability and avoiding the need for additional raters. (See Appendix A and B for full copies.)

Online test (via Socrative App)–MDT. The test using smart devices (MDT) was given to each participant (Treatments 2 & 4 first, Treatments 1 & 3 second). To deliver the test onto the smart devices that students employ in class, Socrative was chosen as the quiz delivery system (www.socrative.com). It is a widely used, free web service that allows for students to do quizzes or even answer quick questions in a manner of polling by using their smart phones. The students were trained on how to navigate the website/app and were given training a week prior to the MDT, however, minimal training was needed since no registration is required to use Socrative; students just need an instructor generated room number. The students were given 15 minutes for the MDT. Just like the PBT, it was reported by the instructors that the overwhelming majority of students finished the quiz with several minutes to spare. The quiz scores were automatically calculated by the web-software. This test has equivalent low-interference and high reliability, like the PBT. (See Appendix A and B for full copy.)

Chapter 4: Design

The groups were completely based on the students who attended the participating professors' classes. Luckily, all students agreed to participate. The questions used in Appendices A and B were plucked right from the books/handouts the students were studying. The primary research methods used were F-tests, t-tests, and ANOVA. As Mackey and Gass (2005) state when describing "comparisons with more than 2 groups ANOVA may be appropriate" (p. 274). An F- test, which can be defined as "a ratio of the amount of variation between the groups to the amount of variation within the groups" (p. 274) can be used to determine variance for t-tests. These t-tests were used to give the final determination of "if the means of two groups are significantly different from one another" (p. 272). There are many groups of data that were taken throughout the course of this study, but the majority of the analysis done focused around comparing two groups, for example MDT vs. PBT, or Test 1 vs. Test 2. Thus, the comparative analyses of f and t-tests were used frequently.

Procedure

I. Participant Training.

Participants trained on the Socrative software in class with their instructors. Instructors were able to verify student participation in real time to ensure all students were properly registered. One week prior to any assessments the students received, the participants were given a brief tutorial to familiarize themselves with the program; submitting answers, moving to the next question, and how to refresh in the event of a freeze.

II. Direct Instruction.

Students received instruction from their professors in language areas determined by their syllabi. The lessons dealt with common vocabulary and grammatical structures in

English based on survival-type situations. The particular lessons were based on the topics of restaurants and shopping. The content of what was being tested was reflective of the content of what had been taught in the class.

III. Group Assignment.

The groups were simply the classes of each professor. Put more simply, each uniqueclass that a professor taught was a separate group. Each of the instructors had at least one group in each test order. This way, comparisons were able to be drawn from within-group and between-group results, ensuring greater validity and increasing the likelihood of statistical significance.

IV. Assessment 1.

The first set of assessments began one month into the semester. The questions were based on what the students studied the week prior. Participants received either the PBT or the MDT based on whatever group they were assigned to. Instructors proctored to ensure no cheating or technology problems. Test times were comparable, as the instructor was responsible to monitor and proctor the time for both MDT and PBT. Upon completion, the MDT was automatically submitted and scored, while the PBT was collected by the proctoring instructor.

V. Assessment 2.

The second set of assessments came two weeks after the first. The questions again were based on what the participants learned a week prior to the assessment. All relevant processes were equivalent to the first set of assessments.

VI. Statistical Analysis.

Analysis was then conducted on the comparisons between PBT and MDT within each professor's students and between the two different professors' student groups. ANOVA, F-tests, and t-tests were used to evaluate the data. All relevant measures were reported, including N, means, standard deviations, critical values, t-values, and p-values. These were used to evaluate the null hypothesis of no significant difference existing between means of assessment.

Analysis

The aim of this study is to compare results between the PBT and the MDT to see if any significant differences in performance arose. To measure this, the key data element used was the raw scores from the different groups. ANOVA repeated measures test was used to measure any differences within-group from first test to second test. ANOVA was used to measure the variance between and within groups. F-tests and t-tests were used to evaluate the overall differences between tests, regardless of testing medium, in order to ensure comparability among tests.

The expected result of the main research question was that the p-value measurement of the differences in performances of all groups between PBT and MDT will be more than .05, thus confirming the initial hypothesis of no significant statistical difference in performance exists.

The results were separated into sections in order to better comprehend the observations made from the data. The following will present data organized into the following sets: Aggregate data, which examines all data observed from all groups in both tests, Test 1 and Test 2, which look at differences within the first test measured across both testing mediums, Test 1 vs. Test 2, which compares the results of both tests to ensure comparable difficulty, PBT vs. MDT (in both Test 1 and

2), examine results on each specific assessment and compare results across mediums, and finally Student Surveys reveal the participants' feelings towards the testing mediums.

Chapter 5: Results

Aggregate

First examined was a broad look at all the data. All the data here was compared within group and between to see if any correlation or relationships exist. Looking first at all the samples from all the groups in Table 1, it can be observed the p-value (.44) is substantially larger than the .05 cutoff. This is a very strong indication that there are not any reasons to conclude that the means differ; no significant relationship exists based on testing mediums when comparing all samples from all groups.

Table 1: Aggregate (Anova: Single Factor)

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
A-test 1	19	335	17.63157895	6.134502924		
B- test 1	19	352	18.52631579	5.374269006		
C- test 1	13	218	16.76923077	18.35897436		
D- test-1	21	359	17.0952381	14.29047619		
X-test 1	16	262	16.375	15.18333333		
Y-test 1	21	350	16.66666667	9.333333333		
Z- test 1	26	441	16.96153846	6.438461538		
A-test 2	17	280	16.47058824	11.13970588		
B-test 2	19	327	17.21052632	5.953216374		
C-test 2	17	267	15.70588235	18.09558824		
D-Test 2	22	369	16.77272727	18.37445887		
X-test 2	10	177	17.7	3.122222222		
Y-Test 2	22	375	17.04545455	8.140692641		
Z-test 2	17	314	18.47058824	6.514705882		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	136.2281364	13	10.47908742	1.00573779	0.446195446	1.760269
Within Groups	2552.729393	245	10.41930364			
Total	2688.957529	258				

Test 1

Moving towards the more specific, the ANOVA one way test on the data compiled from only the first test shows that the p-value (.45) is again very high, which also indicates the null hypothesis should be accepted; no significant relationship among variable (PBT vs. MDT) exist in test 1.

Table 2: Test 1 (Anova: Single Factor)

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
A-test 1	19	335	17.63157895	6.134502924		
B- test 1	19	352	18.52631579	5.374269006		
C- test 1	13	218	16.76923077	18.35897436		
D- test-1	21	359	17.0952381	14.29047619		
X-test 1	16	262	16.375	15.18333333		
Y-test 1	21	350	16.66666667	9.333333333		
Z- test 1	26	441	16.96153846	6.438461538		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	57.76149883	6	9.626916472	0.956227166	0.457688732	2.170155
Within Groups	1288.653316	128	10.06760403			
Total	1346.414815	134				

Test 2

The same analysis applied to test 2 yielded similar results, seen in Table 3. The p-value (.31) was much higher than the .05 threshold yet again, indicating that within the samples from test 2, the null hypothesis should be accepted.

Table 3: Test 2 (Anova: Single Factor)

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
A-test 2	17	280	16.47059	11.13970588		
B-test 2	19	327	17.21053	5.953216374		
C-test 2	17	267	15.70588	18.09558824		
D-Test 2	22	369	16.77273	18.37445887		
X-test 2	10	177	17.7	3.122222222		
Y-Test 2	22	375	17.04545	8.140692641		
Z-test 2	17	314	18.47059	6.514705882		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	76.91586	6	12.81931	1.186526094	0.318278256	2.17699966
Within Groups	1264.076	117	10.80407			
Total	1340.992	123				

Test 1 vs. Test 2

The comparison between the performances between the two tests is important because it can verify that the two unique tests were similar enough in difficulty that comparisons between them are acceptable. Two tests are needed to verify the data (as with the remaining analysis): an F test to check equality of variance and a t-test to check the difference in means and for correlations.

Table 4 performs an F-test that shows that F< F Critical one-tail, so a t-test using unequal variances can be used and the first indicator of the null hypothesis being true is given. Table 5 uses the t-test and since neither t Stat <-t Critical two-tail, nor is t-Stat not > t Critical two-tail, the null hypothesis should be accepted; No significant differences or correlations exist between tests (regardless of testing medium).

Consequently, any results found in the following comparisons of performances on tests based on groupings of testing mediums (PBT vs. MDT) can be attributed to the testing mediums, and not the tests themselves.

Table 4: Test 1 vs. Test 2-F-test (F-test: Two-sample for Variances)

	All of Test 2	*All of Test 1*
Mean	17.02255639	17.16296296
Variance	11.11312372	10.04787175
Observations	133	135
df	132	134
F	1.106017671	
P(F<=f) one-tail	0.280848499	
F Critical one-tail	1.331233328	

Table 5: Test 1 vs. Test 2-t-test (t-test: Two-sample Assuming Unequal Variances)

	All of Test 1	*All of Test 2*
Mean	17.16296296	17.02255639
Variance	10.04787175	11.11312372
Observations	135	133
Hypothesized Mean Difference	0	
df	265	
t Stat	0.353246716	
P(T<=t) one-tail	0.362092164	
t Critical one-tail	1.650623976	
P(T<=t) two-tail	0.724184327	
t Critical two-tail	1.968956281	

Test 1-PBT vs. MDT

The same tests that were applied to test 1 vs. test 2 were applied to all participants who took the first test grouped into the PBT and the MDT. The results from the F-test found in Table 6 found that F< F Critical one-tail in the F-test, therefore the null should be accepted and the t-test could be used with unequal variance. The subsequent t-test from Table 7 revealed the t Stat >-t Critical two-tail, and t-Stat < t Critical two-tail, as a result the null hypothesis should be accepted; No significant differences or correlations exist between testing mediums in test 1.

Table 6: PBT Test 1 vs. MDT Test 1-F-test (F-test: Two-sample for Variances)

	PBT1	*MDT1*
Mean	17.23863636	17.02128
Variance	10.32170846	9.716929
Observations	88	47
df	87	46
F	1.062239799	
P(F<=f) one-tail	0.418398881	
F Critical one-tail	1.558906831	

Table 7: PBT Test 1 vs. MDT Test 1-t-test (t-test: Two-sample Assuming Unequal Variances)

	PBT1	*MDT1*
Mean	17.23863636	17.02128
Variance	10.32170846	9.716929
Observations	88	47
Hypothesized Mean Difference	0	
df	97	
t Stat	0.381841383	
P(T<=t) one-tail	0.351707123	
t Critical one-tail	1.66071461	
P(T<=t) two-tail	0.703414246	
t Critical two-tail	1.984723186	

Test 2-PBT vs. MDT

Near identical results were found when the second test was examined comparing results from the PBT to the MDT. The similar results were interesting because not only was the test different (test 1 vs. test 2), but also more participants were using the opposite medium. In test 1, the majority of participants used PBT (N=88) compared to MDT (N=47), but in test 2 the numbers are almost reversed (N= 48 and N=85 respectively). Here are the results, as seen in Table 8, F< F Critical one-tail in the F-test, therefore null should be accepted and t-test can be used with unequal variance. This

t-test in Table 9 found the t Stat >-t Critical two-tail, and t-Stat < t Critical two-tail, as a result the null hypothesis should be accepted; No significant differences or correlations exist between testing mediums in test 2.

Table 8: PBT Test 2 vs. MDT Test 1-F-test (F-test: Two-sample for Variances)

	PBT2	*MDT2*
Mean	17.4583333	16.77647059
Variance	13.7003546	9.628011204
Observations	48	85
df	47	84
F	1.42296829	
P(F<=f) one-tail	0.0796012	
F Critical one-tail	1.51044302	

Table 9: PBT Test 2 vs. MDT Test 1-t-test (t-test (t-test: Two-sample Assuming Unequal Variances)

	PBT2	*MDT2*
Mean	17.4583333	16.77647059
Variance	13.7003546	9.628011204
Observations	48	85
Hypothesized Mean Difference	0	
df	84	
t Stat	1.07988297	
P(T<=t) one-tail	0.1416426	
t Critical one-tail	1.66319668	
P(T<=t) two-tail	0.28328519	
t Critical two-tail	1.98860967	

PBT vs. MDT- Aggregate

Table 10: PBT vs. MDT-F-test (F-test: Two-sample for Variances)

	PBT-All	*MDT-All*
Mean	17.31617647	16.86363636
Variance	11.43262527	9.599583622
Observations	136	132
df	135	131
F	1.190950121	
P(F<=f) one-tail	0.15775849	
F Critical one-tail	1.33197915	

Table 11: PBT vs. MDT-t-test (t-test: Two-sample Assuming Unequal Variances)

	PBT-All	*MDT-All*
Mean	17.31617647	16.86363636
Variance	11.43262527	9.599583622
Observations	136	132
Hypothesized Mean Difference	0	
df	265	
t Stat	1.142881763	
P(T<=t) one-tail	0.127059578	
t Critical one-tail	1.650623976	
P(T<=t) two-tail	0.254119156	
t Critical two-tail	1.968956281	

When looking at the aggregate scores from both mediums, the data is similar to that when looking at Tables 1-9. The F value is lower than the F Critical 1-tail thus the null is supported and the t-test can use unequal variance. Table 11 shows a t value that is less than the t Critical 2-tail and

greater than the negative t Critical two-tail. This is further proof that there is no significant difference in the performance between PBT and MDT.

Other noteworthy observations from these tables are the mean and the t-Stat. The F-test shows that the mean difference in performance is about a half point higher on the PBT (17.3 vs. 16.8). This shows that the PBT achieved a marginally higher score, about 2.5% larger. Also, the t-stat falls into the range that corresponds with the t Critical two-tail, but it is a larger t-value than in previous tables. This can be partly attributed to the inequity in the grouping of tests- PBT used mostly test 1 and the majority of the MDT were done with test 2.

Student Surveys

The survey results from Table 10 show that more than two-thirds of the students viewed PBT positively and three-fourths of the students viewed the MDT positively. The biggest result found was in questions 3; revealing that an overwhelming majority prefer the MDT over the PBT (even with a 'no preference' option given). Lastly, the vast majority did not perceive a difference in the testing mediums. There were 139 responses taken.

Table 12: Survey Results: Group All

Q1	**What was your satisfaction with the paper test?**					
Response Choice:	**1**	**2**	**3**	**4**	**5**	
Actual responses	0	8	37	47	47	
%	0%	6%	27%	34%	34%	
Q2	**What was your satisaction with the socrative test?**					
Response Choice:	**1**	**2**	**3**	**4**	**5**	
Actual responses	1	6	27	45	60	
%	1%	4%	19%	32%	43%	
Q3	**Which test did you prefer?**					
Reponse Choices	**Paper**	**Socrative**	**No Preference**			
Actual Responses	41	90	8			
%	29%	65%	6%			
Q4	**Do you feel the testing format changed the difficulty of the tests?**					
Response Choices	**Yes**	**No**				
Actual Responses	16	123				
%	12%	88%				

Chapter 5: Discussion

All data sets produced the same results: no significant difference exists between testing methods. Table 1 shows no significant difference between any groups with both testing mediums over both tests. Table 2 shows no significant difference between all tests taken with test 1 with both testing mediums. Table 3 shows no significant difference between all tests taken with test 2 with both testing mediums. Tables 4 and 5 show no significant difference between performances between tests 1 and 2. Tables 6 and 7 show no significant difference in performance in test 1, when comparing PBT against MDT. Finally, tables 8 and 9 show no significant difference in performance in test 2, when comparing PBT against MDT.

These results show significant evidence that there is no meaningful difference when the testing medium is changed. The analysis done has taken into account cross-test difficulty, and Tables 4 and 5 show that the test are in fact comparable with means of 17.16 when N=133 and 17.02 when N=135, respectively for tests 1 and 2.

These results then confirm that there is no quantifiable reason not to use smart phones as part of MDT as an assessment tool in an educational capacity. These findings were gathered in an EFL language-learning environment, which shows that MDT can effectively be used as a language tool. However, there are no major reasons to believe that MDT needs to be limited to the language arena. Many other subjects would be adequate candidates to utilize this method.

Relating these results back to the research question: Is there a statistically significant difference in performance of young adult EFL learners based on assessment methods- paper and pencil-based tests (PBT) vs. mobile device tests (MDT)? The data would indicate that there is not. None of the data sets provided any data to prove otherwise. Aggregate, test vs. test, format vs.

format, all comparisons run showed that the two testing mediums- PBT and MDT, had comparable results.

Survey Results

The participant surveys seen in table 10 indicate two key takeaways from the test taker's point of view: Socrative (the smart technology based test) was preferred over the paper-based test and that there was no perceived difference in difficulty due to the testing medium.

With regard to the research questions: Are there significant differences in the perception of the difficulty and preference of testing methods? The first question is easy to answer from the data- no, there is no perceived difference in difficulty. Eighty-eight percent of respondents answered that the format of the test did not change the difficulty. This answer is clear. Korean EFL students do not find using a web assessment tool, like Socrative, to be any more difficult than a normal paper and pencil test. As for the second question, there appears to be clear support for an answer as well–yes, Socrative is preferred over a traditional pencil and paper test. An overwhelming 65% preferred Socrative, compared with only 29% preferring PBT, with 6% having no preference.

Both perception-based questions have clear data to support an answer for them without any major ambiguity. Some possible causes for these results will now be discussed.

Survey-testing Method Preference

The positive feelings for using the phone application over the paper-based test was not terribly surprising. Following the equivalency hypothesis of this paper, I assumed there would be a much closer result, but 65% of respondents preferred the MDT. Other research suggests that students do prefer using newer forms of technology over traditional ones and generally hold more positive opinions of newer technology compared with paper and pencil tests (Escudier et al, 2011;

Gikas & Grant, 2013; Karadeniz, 2009; Kim et al, 2011; Milrad & Spikol, 2007; Paredes et al., 2010; Yurdabakan & Uzunkavak, 2012).

One possible cause for the favorable attitudes towards the Socrative tests is simply due to the fact that younger people are more open to technology. The youth have generally been the early adapters throughout the development of technology in all its forms. The participants of this study may well have been no different. Almost all the participants in this research were 19 years old. This generation grew up with technology. As Nam (2013) explains, by grade 6 most students in Korea have a smartphone. This type of familiarity with technology means that they would be less intimidated by a new testing medium, more willing to try it, encounter fewer problems in dealing with it, and be more likely to know about or have used this application before. If the population of this research were much older, the survey results may have turned out significantly different.

Another possible cause for the positive impression that the MDT left could be the instant feedback offered by the application. When using Socrative to perform a quiz, the instructor can select an option that gives the student the result of the question they just completed as soon as they submit their response. Since the majority performed well, with a score over 80% most students were getting positive feedback as they were performing the quiz (each correct submitted answer would yield a positive result). This would be an encouraging sign to students as they were performing and could account for some of the positive attitudes given. Additionally, the instant feedback that the application put into contrast with the one week that students had to wait for their instructor to physically grade their paper tests could also account for some of the difference in test medium preferences.

Survey-testing Method's Effect on Difficulty

The survey netted another surprising result- the overwhelming selection of "No" to the question: "Do you feel the testing format changed the difficulty of the tests?" Eighty-four percent of responses were "No". I had expected this result, but not by such a vast margin. I will now explore some possible causes for this outcome.

The easiest cause to explain would be if the results from the MDT were higher than the PBT, however, the opposite was the reality. The mean result for all PBT was 17.3, about a half point (or 2.5%) higher than the mean performance on PBT- 16.8. So this clearly cannot be the reason and is in fact a further testament to their preference for a medium that they actually performed (ever so slightly) worse in.

There are other possibilities that are not as quantifiable. One such factor that could attribute to this survey result could be that the test questions were too easy. If the test questions were too easy, then both tests would be perceived that way meaning no difference in difficulty. In addition, the test questions were all in multiple choice format. This questioning style is often regarded by students as an easier type of test. If other question types were used (short answer, fill in the blank, etc…), there could have been a very different perception of Socrative. These other question types were not used because there are compatibility issues and data issues with how much processing speed and internet data usage is needed for other question types currently. These issues can slow down phones and cause errors.

The results indicate that on average students scored 17.16 and 17.02 on tests 1 and 2, respectively, which are high scores, but there were enough errors to provide variances in the participants' results and provide data for comparison. Had the tests been too difficult, it would have run the risk of having participants drop out thus giving a statistically insignificant sample. Moreover,

the instructors of these participants were involved in the test creation and they both agreed that the tests reflected their students' English level and capabilities.

Chapter 6: Conclusion

The F-tests and p-tests used in comparing the difficulty of two tests show that they were approximately equivalent, which validates the other ANOVA tests done on all data samples to show that there was no statistical difference in performance when changing testing mediums. In addition, a survey of 139 participants revealed that they held an overall positive perception of MDT, prefer MDT over PBT, and did not perceive any difference in difficulty between the two tests. The implications of these findings will be discussed here.

The most obvious takeaway from this study is that paper and pencil testing does not need to be held on a pedestal and regarded as the only option for testing. This is especially true with the proliferation of technology and the ever-growing pool of resources that are at an instructor's disposal. That is not to say that PBT should be abandoned. This study was conducted in one of the most wired countries in the world and has a near 100% smart phone penetration rate. This is not the case globally. Testing should fit the needs of the students. Therefore, only in circumstances where a student population is similarly affluent enough to have access to the technology needed for MDT, should it be considered as a viable alternative.

The positive feelings and preferences of MDT that participants had is another reason that MDT should be further explored as an alternative to PBT. The fact that students are open to it, means there will be little resistance given if a teacher chooses to use a non-traditional testing medium, like Socrative. It is debated whether the positivity affects performance. This study found that it did not. Regardless, the positive feelings are difficult to interpret in a negative way. If assuaged correctly, those positive feelings could be turned into increased motivation, which in turn could create a positive effect of performance. That is a great number of "ifs" though.

Moreover, paper and pencil tests have nowhere to go. The nature of PBT is that they are exactly that- paper and pencil. They could use some pictures and charts, perhaps some color if the school had the resources. But that is it. MDT on the other hand, is a constantly developing medium. It is relatively unlimited in what it can do. Currently it has the ability to embed pictures and videos, as well as use personalization and adaptive questions. All this is from web services that are still in their relative infancy. The list of functions that can be done with mobile devices will undoubtedly grow and be significantly longer in the next 5-10 years.

Limitations

The generalizability of this study is limited by several factors; the participants of the study were all smartphone-literate. Had they not been, the MDT would have encountered problems and the surveys would have been less positive. Secondly, several students did not do both tests due to absences. The effect of this was likely small, but had they performed drastically different on one test compared to another, results on both statistical analysis and surveys may have moved slightly. Thirdly, all the test questions were multiple choice. This was favorable for analysis, but perhaps not the most indicative of what many teachers use in their quizzes on a day-to-day basis. Other question types such as fill-in-the-blanks, open ended questions and matching are available but were found to have issues on certain devices that are commonplace in the student population. Next, the population of this study was limited. A population of 139 is not inconsequential, but greater numbers are always useful for achieving significance. The imbalance of the groupings was another issue. There were five groups doing test 1 then 2, compared to two groups doing tests 1 then 2. This was not by design, but how the instructors of the classes executed the study. Looking forward, future research on the comparison of PBT vs. MBT should use large samples of balanced groupings, could contain different question types, and although the comparison of the two tests was effective based on the analysis,

comparative testing with the same test over two different mediums has been done in much research and could be used in the future, as could statistical analysis more advanced that ANOVA, F-tests, and t-tests.

References

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, *1*(*1*), 1-24.

Bomhold, C. R. (2013). Educational use of smart phone technology: A survey of mobile phone application use by undergraduate university students. *Program: Electronic Library & Information Systems*, *47*(4), 424-436.

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior*, *23,* 463-477.

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.

Chua, Y. P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior*, *28*(5), 1580-1586.

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior*, *29*(5), 1889-1895.

Deichman, J. D. (2014). Socrative 2.0. *Knowledge Quest*, *43*(2), 72-73.

Dosch, M, (2012). Practice in computer-based testing improves scores on the national certification examination for nurse anesthetists. *AANA Journal*, *80*(4), 60-66.

Dresselhaus, A., & Shrode, F. (2012), Mobile technologies & academics: Do students use mobile technologies in their academic lives and are librarians ready to meet this challenge? *Information Technology & Libraries*, *31*(2), 882-101.

Educational Testing Service. (2011). *Practical considerations in computer-based testing.* Princeton, NJ: Davey. Retrieved from https://www.ets.org/Media/Research/pdf/CBT-2011.pdf.

Emarketer. (2013). Smartphone adoption near-complete among younger South Koreans. Retrieved from http://www.emarketer.com/Article/Smartphone-Adoption-Near-Complete-Among-Younger-South-Koreans/1010314.

Escudier, M., Newton, T., Cox, M., Reynolds, P., & Odell, E. (2011). University students' attainment and perceptions of computer delivered assessment: A comparison between computer-based and traditional tests in a 'high-stakes' examination. *Journal of Computer Assisted Learning*, *27*(5), 440-447.

Gikas, J., & Grant, M. M. (2013). Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media. *The Internet and Higher Education*, *19*, 18-26. doi:10.1016/j.iheduc.2013.06.002.

Jamil, M., Tariq, R. H., & Shami, P. A. (2012). Computer-based vs paper-based examinations: Perceptions of university teachers. *Turkish Online Journal of Educational Technology-TOJET*, *11*(4), 371-381.

Jubien, P. (2013). Shape shifting smart phones: Riding the waves in education. *Canadian Journal of Learning and Technology*, *39*(2), 1-16.

Jung, H. (2014). Ubiquitious learning: Determining impacting learners' satisfaction and performance with smartphones. *Language Learning and Technology*, *18*(3), 97-119.

Karadeniz, S. (2009). The impacts of paper, web and mobile based assessments on students' achievements and perceptions. *Scientific Research and Essay*, *4*(10), 984-991.

Kim, P., Hagashi, T., Carillo, L., Gonzales, I., Makany, T., Lee, B., & Gàrate, A. (2011). Socioeconomic strata, mobile technology, and education: A comparative analysis. *Educational Technology Research & Development, 59(*4), 465-486.

Lai, C. (2006). The advantages and disadvantages of computer technology in second language acquisition. *National Journal for Publishing and Mentoring Doctoral Student Research*, *3*(1), 44-50.

Mackey, A., & Gass, S. (2005). Second language research methodology and design. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Milrad, M., & Spikol, D. (2007). Anytime, anywhere learning supported by smart phones: Experiences and results from the MUSIS project. *Journal of Educational Technology & Society*, *10*(4), 62-70.

Mintz, J., & Aagaard, M. (2012). The application of persuasive technology to educational settings. *Educational Technology Research & Development, 60*(3), 483-499.

Nam, I. (2013). A rising addiction among youths: Smartphones. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/SB10001424127887324263404578615162292157222.

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

Paredes, J., Pennington, R., Pursell, D., Sloop, J., & Tsoi, M. (2010). Engaging science students with handheld technology and applications by re-visiting the Thayer method of teaching and learning. *Georgia Journal of Science, 68*(4), 186-195.

Pew Research Center. (2014). *Mobile technology fact sheets* [Data file]. Retrieved from http://www.pewinternet.org/fact-sheets/mobile-technology-fact-sheet/.

University of Minnesota, National Center on Education Outcomes. (2010). *Computer-based testing: Practices and considerations, Synthesis Report 78.* Minneapolis, MN: M. Thurlow, S. S. Lazarus, D. Albus, & J. Hodgson.

Yurdabakan, I., & Uzunkavak, C. (2012). Primary school students' attitudes towards computer based testing and assessment in Turkey. *Turkish Online Journal of Distance Education*, *13*(3), 177-188.

Appendix A: Socrative by Master Connect: Restaurant Quiz

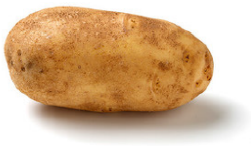
socrative
by MasteryConnect

Name: _____ Date: _____

Quiz name: Restaurants Quiz

1. What is this ?

(A) potato
(B) salad
(C) onion
(D) carrot



2. What is this ?

(A) Strawberry
(B) Apple
(C) Blueberry
(D) Pea

3. What is this?

(A) Banana
(B) Carrot
(C) Onion
(D) Apple

4. What is this?

(A) Juice
(B) Soft drink
(C) Beer
(D) Tea

5. What is this?

(A) Ice cream
(B) Pizza
(C) Potato
(D) Cake

6. Which of the following is a DRINK?

(A) Strawberry
(B) Beer

C Brownie
D Corn

7. 7) Which of the following is a FRUIT?

A Banana
B Juice
C Cake
D Pork

8. Which of the following is a dessert?

A Pizza
B Pie
C Chicken
D Pasta

9. Which of the follow is a MEAT?

A Carrot
B beef
C Apple
D Orange

10. Which of the following is a VEGETABLE ?

A Cherry
B Pasta
C Lemon
D Broccoli

11. What kind of food can you buy at a Diner/Café?

A Pizza
B Alcohol
C Pie
D Nachos

12. What kind of food can you buy at an Italian Restaurant?

A Lasagna
B Chicken
C Fries
D Tacos

13. What kind of food can you buy at a Pub/Bar?

A Lasagna
B Fajitas
C Alcohol
D Pasta

14. What kind of food can you buy at a Mexican Restaurant?

(A) Pizza
(B) Chicken
(C) Sandwich
(D) Fajitas

---

15. What flavor is a lemon?

(A) Sweet
(B) Salty
(C) Sour
(D) Spicy

---

16. What flavor is medicine?

(A) Sweet
(B) Bitter
(C) Savory
(D) Salty

---

17. What flavor is sugar?

(A) Sweet
(B) Spicy
(C) Tender
(D) Chewy

---

18. Which sentence is correct?

(A) Can I hamburger?
(B) Can I get hamburger?
(C) Can I get the hamburger?
(D) Hamburger can I get?

---

19. Which sentence is correct?

(A) I'd like to have the spaghetti
(B) I'd like have spaghetti
(C) I like spaghetti to have
(D) I like to spaghetti

---

20. When do people usually eat dessert?

(A) before dinner
(B) During / at dinner
(C) After dinner
(D) Before, during, and after dinner

Appendix B: Socrative by Master Connect: Shopping Quiz

Name: _____ Date: _____

Quiz name: Shopping Quiz

1. What is the opposite of Noisy/ Loud?

(A) small
(B) ineffective
(C) quiet
(D) strong

2. What is the opposite of Hard?

(A) useful
(B) small
(C) new
(D) soft

3. What is the opposite of New?

(A) old
(B) broken
(C) ineffective
(D) weak

4. What is the opposite of loose?

(A) baggy
(B) big
(C) comfortable
(D) tight

5. Complete the sentence: The sun is _____ _____ the clouds.

(A) bright than
(B) brighter than
(C) more bright
(D) more brighter

6. Complete the sentence: The shoes are _____ _____ the gloves.

(A) Expensive than
(B) Expensiver than
(C) More expensive
(D) More expensiver

7. The jeans are $10.00, the dress is $15.00 Which sentence is correct?

(A) The jeans are more expensive than the dress
(B) The jeans are cheaper than the dress

C The dress is cheaper than the jeans
D The dress is the cheapest

8. Which sentence is NOT correct

A The hat is better than the jacket
B The blouse is more bad than the pants
C The skirt is worse than the scarf
D The t-shirt is better than the necktie

9. The shoes are nine dollars and ninety five cents What is the price of the shoes?

A $9.95
B $9.55
C $9.59
D d) $9.99

10. The suit is twenty four dollars and thirty eight cents What is the price of the shoes?

A $20.48
B $24.13
C $24.38
D $20.38

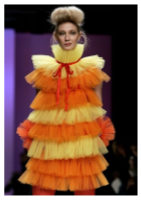
11. The phone is _____ (see the picture)

A Expensive
B comfortable
C big
D bright

12. The pants are _____ (see the picture)

A Cheap
B Short
C Dirty
D Loose

13. The dress is _____ (see the picture)



A plain
B ugly
C long
D heavy

14. The T-shirt is _____ (see the picture)

(A) dirty
(B) formal
(C) dark
(D) beautiful

15. What color is NOT in Korea's flag (태극기)

(A) blue
(B) red
(C) green
(D) black

16. What color is the book we study?

(A) blue
(B) red
(C) green
(D) black

17. Q: _____ A: The shoes are five dollars. What is the question?

(A) How much are the shoes?
(B) How much is the shoes?
(C) How much the shoes is?
(D) How much the shoes are?

18. What kind of clothes go on your FEET?

(A) Pants
(B) shoes
(C) belt
(D) watch

19. What kind of clothes do you wear on your NECK?

(A) socks
(B) jacket
(C) hat
(D) scarf

Q: How much is the pencil case?
A:_____

20. What is the answer?

(A) Ten dollars is the pencil case
(B) Ten dollars are the pencil case
(C) The pencil case are ten dollars
(D) The pencil case is ten dollars