Assessing the Usability of Smartwatches for Academic Cheating during Exams

Stephanie Wong, Lillian Yang, Bernhard Riecke, Emily Cramer and Carman Neustaedter
School of Interactive Arts and Technology, Simon Fraser University
250 – 13450 102nd Avenue, Surrey, BC, Canada, V3T 0A3
swa163@sfu.ca, lillian.mj.yang@gmail.com, ber1@sfu.ca, ecramer@sfu.ca, and carman@sfu.ca

ABSTRACT

Smartwatches are growing in usage, yet they come with the additional challenge of regulating their usage during the taking of academic tests. However, it is unclear how effective they are at actually allowing students to cheat. We conducted an experiment that examines the use of smartwatches for cheating on Multiple-Choice Questions (MCQ) and Short Answers (SA) with either Pictures/Text shown on the watch to aid students. Our results indicate that smartwatches are neither efficient nor have a high usability rating for cheating. However, students are able to score higher on Multiple-Choice Questions compared to Short Answers. We use the cheating paradigm as an example to understand the perceived usability and appropriation of smartwatches in an academic setting. We provide suggestions that help to deter cheating in an academic setting. Our study contributes to the research on academic integrity and the growing demand of wearable technologies.

Keywords

cheating, academic integrity, usability, wearables, smartwatches, Usability Metric for User Experience (UMUX).

ACM Classification Keywords

H.5.3 [*Computer-supported cooperative work*]: Group and Organization Interfaces

INTRODUCTION

Cheating on tests is a recurring problem in both schools and the workplace. According to Park *et al*. [25], students cheat when there is a lack of time, when students perceive the test as impacting on their future employment opportunities and when students lack motivation to study. Similarly, professionals such as pilots, police officers, firefighters and even taxi drivers are reported to cheat for performance bonuses, job security and promotions [10] [28]. People have the desire to be ahead and are required to pass an assessment with high stakes either for entrance exams or for maintaining their licenses and certification examinations [10]. A number of studies have shown that students who engage in academic dishonesty later display unethical practices in their workplace, including business students [15] [18], nursing students [18] [25] and accountants [5]. With advances in technology (e.g., Google Glass, Bluetooth pens, Spy cameras, smartwatches and embedded sensors), the methods to cheat and gain access to answers have become more sophisticated and can easily go undetected in universities and in professional settings [30] [33]. Bachore [4] states with virtual environments where assessments are conducted online without proctors physically monitoring, there will be more opportunities and attempts to cheat with a lesser likelihood of being caught. A study by Migicovsky *et al*. [21] demonstrated how dishonest students can collaboratively cheat in real-time using the smartwatch prototype, ConTest, with other combined resources (a cloud-based service, a smartphone, and a client application). The benefits of smartwatches for students among other wearables are minimal interactions for user input to gain easy, fast and discreet access to stored text and images, real-time data from the internet and to applications saved in the mobile phones [36]. Smartwatches look like a normal wristwatch in terms of its appearance and size, which makes it easier for students to use and hide it under their sleeves during cheating [21].

The issue of appropriate detection and prevention of cheating enabled with advanced technologies such as smartwatches in academic environments are still not clearly understood or tested thoroughly by HCI researchers and educational institutes [10]. A quick, preventive measure to deter cheating implemented by many institutions such as SAT college, Queen Mary, University of London and Kyoto University is to prohibit smart devices from exam halls [37]. While, there is a diverse range of wearables from which students can cheat, we focus on smartwatches as they reach a broad consumer audience and there is a need to gather a better understanding of the perceived usability and

MobileHCI '17, September 04-07, 2017, Vienna, Austria
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5075-4/17/09…$15.00
http://dx.doi.org/10.1145/3098279.3098568

appropriation of smartwatches in an academic setting. Past HCI research has emphasized on enhancing interaction and improvement for everyday users [22], but we still do not know how effective, efficient and usable smartwatches actually are for allowing students to cheat in during classroom exams. If students from varied disciplines (with no prior knowledge) were to gain access to answers through their smartwatches, could they perform well in an exam? And for what kind of questions would a smartwatch be most suitable? Thus, we use this cheating' paradigm to gather an understanding of the perceived usability and appropriation of smartwatches in an academic setting. The better we understand the specific affordances and capabilities of smartwatches for cheating, the better we can translate recommendations for deterring everyday cheating in the academic setting. To investigate this, we examined the assessment and response format of an anatomy test, as these assessments typically involve Multiple Choice Questions (MCQs) and questions with images and text [32]. We used an Apple smartwatch for this study because it supports common interactions (tilting the arm to glance, tapping for selection, and using the palm of one hand to de-activate the display) and is among the most popular smartwatches being sold [22]. In the following sections of the paper, we present related literature, describe the methods, and then discuss how the results will influence future classroom settings and any conceivable concerns that may arise.

RELATED WORK

Academic Dishonesty

Over the past few decades, research into academic dishonesty has been conducted mainly in quantitative design, and self-reported surveys with students indicating their own cheating behaviours [3]. The various cheating behaviors that students have reported in the surveys includes the use of unauthorized materials in exams or assignments; fabricating information, references or results, plagiarism and fraud (e.g., having other students commit a dishonest academic act such as arranging with other students to give or receive answers using cheat sheets or past term papers) [19] [27]. There are many reasons why students cheat including lack of time, belief that academic performance strongly affects their employment opportunities and a lack of motivation to study [25], however Wideman [35] conducted semi-structured interviews with eleven nursing students and interpreted that students tend to neutralize their cheating behavior as caring and frequently engaged in collaborative cheating to complete assignments. Assessment format also contributes to the cheating behavior of students. Unlike constructed responses, where students have to think to write short answers, essays or problem-solving questions (e.g., mathematics proofs), Multiple Choice Questions (MCQs) typically only require students to recall a single word or fact [11] [12]. Thus, because it is easier to cheat on MCQs than any other test format, the prominence MCQs in standardized education enables academic cheating at even high levels of education [11] [12].

Smartwatch Interactions

Smartwatches such as Apple Watch and Samsung Gear have the potential to provide unobtrusive and discreet access to phone message notifications, applications, and incoming calls [22]. With minimal user input and micro interactions, such as touching the screen, pressing the side buttons/dial or using gestures, users can be hands-free and attain their required information in seconds [2] [36]. An example is provided by Akkil *et al.* [2], where they use glance awareness and gaze gestures (looking left, right and up) for selection of items on a smartwatch. Their experiment, conducted with twelve participants, revealed that the gaze-based interaction was practical for simple tasks and haptics was the preferred feedback modality ([2]. In the area of optimizing text entry and improving a users' performance to achieving a high entry speed, Oney *et al.* [24], Chen *et al.* [9], and Komninos & Dunlop [16] prototyped a QWERTY keyboard and explored zooming, swiping and next-word predictions to enable faster text entry. Perrault *et al.* [26] investigated eyes-free tactile interaction using a watch strap to overcome problems with visual occlusion and the 'fat finger problem' with twelve participants in their pilot study and eight participants in the user study. Their study revealed auditory solutions were more accurate than gesture techniques which had a lower success rate with participants. Bernaerts *et al.* [6] targeted office workers and uses pre-defined (knock, lock, and return) gestures on a Samsung Galaxy Gear S smartwatch application to grant access to physical rooms. Their application used audio, vibration and graphics for output responses. Bieber *et al.* [7] also used gestures to allow workers to receive new instructions without having to stop their current tasks or read the next chapter of the manual without having to touch the display of the smartwatch. Nebeling *et al.* [23] used speech commands with seven participants to collaborate with 29 crowd workers each, to dictate tasks, respond to questions, and receive notifications of major edits on their smartwatches.

Despite being valuable and necessary studies, there is limited research that examines the perceived usability and appropriation of smartwatches in an academic setting [8]. Romero *et al.* [29] studied how smartwatches notifications can help to improve e-learning environments and developed an application to help students self-regulate, avoid procrastination and submit timely assessment in a massive open online course. Commercial products such as 24kupi smartwatch have been designed to help students cheat by auto-scrolling through a student's notes [1]. Migicovsky *et* *al*. [21] demonstrates how dishonest students can collaboratively cheat in real-time using the smartwatch prototype, ConTest, with other combined resources (a cloud-based service, a smartphone, and a client application). From afar, their smartwatch application appears like a normal digital watch with date and time, but the answers are encoded in groups of missing pixels. Without much interaction, students can also vote for a particular answer by double-clicking the watch buttons [21]. Migicovsky *et al*. [21] research establishes that cheating is possible using smartwatches, but it focuses on only one type of assessment format, i.e. MCQ, and does not address the research question: how effective, efficient and usable are smartwatches to cheat in typical classroom assessments?

METHOD

The goal of our study was to assess students' abilities to cheat on exams using a smartwatch, where we are interested in comparing picture-based and text-based answers, as well as MCQs and SA questions. Given the past evidence that shows that MCQ assessments are in common use and easy to cheat on [11] [12], we hypothesize that students would have more correct answers on Multiple-Choice Questions (MCQ) than Short Answers (SAs) in the assessment **(H1).** Secondly, Kortum [17] states that a small display screen will impact people's mental and physical effort to interact with them, i.e. some tasks will take longer time for users to complete. We predict that the mental effort for a student to recognize and match the answers using images would take longer when using the smartwatch. Therefore, we hypothesize that students would complete the text questions in a shorter time period than pictorial questions **(H2).** As for the physical effort, we predict that SA will require more reading and scrolling because it consist of multiple sentences, whereas MCQ consist only of a set of letters. Therefore, we hypothesize that students will result in having a higher frequency of interaction with the smartwatch when completing the SA than MCQ **(H3).** Also, as most commercial smartwatch applications do not support pinching to enlarge images, such as those for Apple Watch, we hypothesize that students would have a higher frequency of interaction with the smartwatch when searching for picture answers than text answers **(H4).** Lastly, as smartwatches may help students to cheat without having to study, we hypothesize that students will experience an overall high usability for a text response format as compared to a picture response format **(H5).**

Participants

We conducted a controlled experiment with a total of 16 participants (5 males and 11 females). The average age of male participants was 21.4 years (*SD* = 2.01), the average age of female participants was 21.9 years (*SD* = 2.71). The participants were recruited through snowball sampling from the course called Information Design, taught at Simon Fraser University. All participants were undergraduate students (2 international students and 14 residents), studying in the field of design, web, and business. All participants were regular users of desktop, laptop, smartphone, and tablets. Only 2 participants had prior experience with smartwatches as they owned Apple Smartwatch and Gear Fit 2. Participants included 2 left-handed students and 14 right-handed students. In addition, there were 10 participants who disclosed that they had cheated in the past, while 6 participants claimed they had not cheated. Participants were remunerated with university course credit for their contribution.

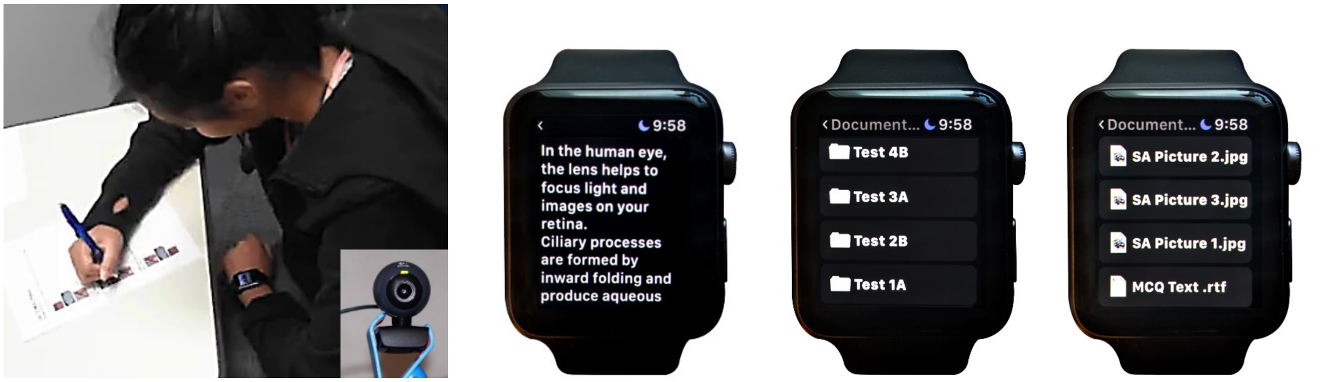
Reliability and Validity

To control external validity and content validity, we took samples from the student population, so that the results can be a close representation of the population performing an academic assessment. For the internal validity and reliability, we modeled the exam questions closely to a real assessment such as the study of anatomy, which is fundamental in medical exam questions and requires the understanding of the physical nature of parts of the human body and their spatial relationships [13] [20]. Therefore, in this study, we constructed the assessment content to include text and pictures from the anatomy subject, which can also be generalized to other tests such as architecture or website wireframes. Also, as per Cizek's [10] suggestion on reporting cheating behaviours, we used scenarios (separate for the male and female participants) to ensure that we are ethically compliant and we did not bias the participants. These scenarios motivated participants to think and act as another person, but enabled them to report more truthfully on their own behavior to cheat [10]. An instance is:

Craig has a test in his Psychology class in the morning and an Anatomy test in the evening. His grade in this class matter as they are tied to his scholarship for the next term. Craig is having a hard time understanding some of the material, and he complains to his mentor in his social organization about it. His friend comments that the library keeps copies of old tests in a file, and they are available for studying. Craig looks in the file and finds that his class's test is in the file and they are the same for every year. He makes copies of it in his smartwatch, and uses it to achieve the maximum number of correct questions in the test, without being caught.

Setup and Stimuli

We set the experiment to look like a classroom with only one participant and one exam moderator per trial (Figure 1a) in a university. The stimuli were the questions in the assessment provided in a paper-based format. We positioned a web-cam on the top-left side of the participant to enable a clear view of the participant's interaction with the smartwatch (Figure 1a). We recorded the students from



In the human eye, the lens helps to focus light and images on your retina. Ciliary processes are formed by inward folding and produce aqueous

(a) (b) (c) (d)

Figure 1. (a) Classroom Simulator with webcam, (b) Apple Smartwatch, (c) Document Pro (d) Answer Cheat Sheet

the web-cam using the Debut Capture software installed on the laptop. The experiment included an Apple Smartwatch (width: 42mm by height: 42mm) with (312 x 390) screen resolution. The Apple Smartwatch consisted of the response, which are the anatomy answers modelled as student's cheat-sheet (as shown in Figure 1b).

The application 'Document Pro was installed and set as the default app for glancing for information on the smartwatch (Figure 1c). For each assessment question (the stimuli), students were required to locate their stimuli's test folder on the smartwatch and then find the associated response (Figure 1d, shows 2 text and 4 picture files). Students could not preview the contents of the stimuli. The stop-watch was used to assist the exam moderator in monitoring consistently the same time interval for each student and to record the student's task completion time. Further, the participant's informed consent, demographics form (age, use of technology, cheating history), cheating scenarios and a post-questionnaire (measuring the overall usability of the smartwatch along with open-ended questions asking about participants' experiences) was collected in a paper- based format.

Design and Tasks

The experiment is a 2x2 factorial within-subject design as shown in Table 1. There are two independent variables in Table 1: the assessment format at two level (MCQ and SA) and the response format at two levels (Picture and Text). For the dependent variables, we measured the task success scores, task completion time, interaction frequency on a ratio scale, and for the user satisfaction (rated on a 11-point Likert scale), we used the ordinal scale.

Table 1: 2x2 Within Subjects Experimental Design

Assessment Format	Response Format	
	Picture (P)	Text (T)
MCQs	P_{MCQ1}, P_{MCQ2}	T_{MCQ1}, T_{MCQ2}
SAs	P_{SA1}, P_{SA2}	T_{SA1}, T_{SA2}

Table 2 shows an example of a complete assessment Test 1A, with a total of 12 questions divided into 4 components (a combination of $\mathbf{P}_{MCQ}$, $\mathbf{T}_{MCQ}$, $\mathbf{P}_{SA}$, $\mathbf{T}_{SA}$). The first row consists of example questions of the assessment, and the second row shows the corresponding response which are the answers in Picture or Text on the smartwatch. Participants are required to do each component individually. For instance, if a participant did component $\mathbf{P}_{MCQ1}$ first, he/she will check for its corresponding answer ***MCQ Picture*** on the smartwatch. Once completed, he/she can do the 3 questions of component $\mathbf{T}_{\mathbf{MCQ1}}$ with the corresponding response ***MCQ Text*** on the smartwatch. To counterbalance for any systematic errors, the order of these components was randomized within the 4 conditions (MCQ, SA, Picture, Text), thence making it a complete 8 order experimental design ($P_{\mathbf{MCQ2}}$, $T_{\mathbf{MCQ2}}$, $P_{\mathbf{SA1}}$, $T_{\mathbf{SA1}}$) as shown in (Table 3). To mitigate for any reactivity or experimenter effects that may bias our results, we randomly assigned participants to all conditions and provided a short training to all participant on how to use the smartwatch.

Table 3: Order of the components

Group	1	2	3	4
1	P_{MCQ1}	T_{MCQ1}	P_{SA2}	T_{SA2}
2	T_{MCQ1}	P_{MCQ1}	T_{SA2}	P_{SA2}
3	P_{MCQ2}	T_{MCQ2}	P_{SA1}	T_{SA1}
4	T_{MCQ2}	P_{MCQ2}	T_{SA1}	P_{SA1}
5	P_{SA1}	T_{SA1}	P_{MCQ2}	T_{MCQ2}
6	T_{SA1}	P_{SA1}	T_{MCQ2}	P_{MCQ2}
7	P_{SA2}	T_{SA2}	P_{MCQ1}	T_{MCQ1}
8	T_{SA2}	P_{SA2}	T_{MCQ1}	P_{MCQ1}

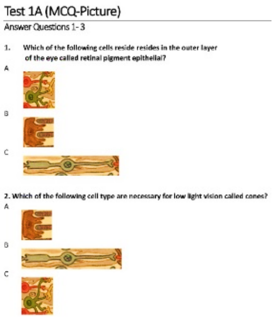
Pilot Study

Prior to the actual assessment, we conducted a pilot study with additional participants to ensure a natural exam setting. The pilot study helped to set a baseline for the estimated task success rate, task time completion, time-interval (30 seconds/minute for when the exam moderator will look at the student) and to determine if prior learning

Table 2: Example of Assessment Components

Stimuli Assessment Format

P_{MCQ1}



Test 1A (MCQ-Picture)

Answer Questions 1- 3

1. Which of the following cells reside resides in the outer layer of the eye called retinal pigment epithelial?

A

B

C

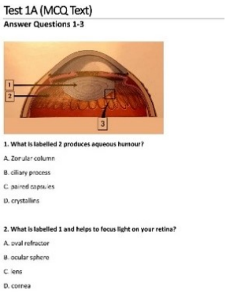
2. Which of the following cell type are necessary for low light vision called cones?

A

B

C

T_{MCQ1}



Test 1A (MCQ Text)

Answer Questions 1-3

1. What is labelled 2 produces aqueous humour?

A. Zorular column

B. ciliary process

C. paired capsules

D. crystallins

2. What is labelled 1 and helps to focus light on your retina?

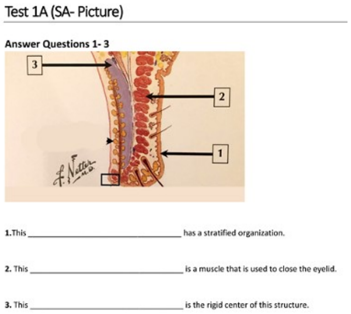
A. oval refractor

B. ocular sphere

C. lens

D. cornea

P_{SA2}


3
2
1
F. Netter M.D.

Test 1A (SA- Picture)

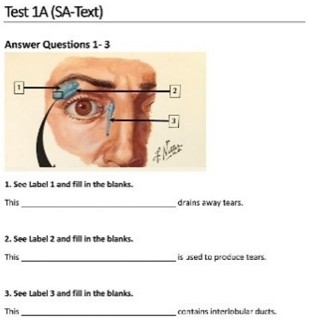
Answer Questions 1- 3

1.This ______________________________ has a stratified organization.

2. This ______________________________ is a muscle that is used to close the eyelid.

3. This ______________________________ is the rigid center of this structure.

T_{SA1}


1
2
3
F. Netter M.D.

Test 1A (SA-Text)

Answer Questions 1- 3

1. See Label 1 and fill in the blanks.

This ______________________________ drains away tears.

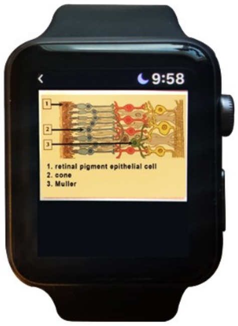
2. See Label 2 and fill in the blanks.

This ______________________________ is used to produce tears.

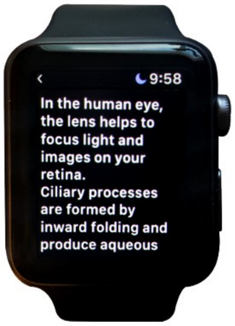
3. See Label 3 and fill in the blanks.

This ______________________________ contains interlobular ducts.

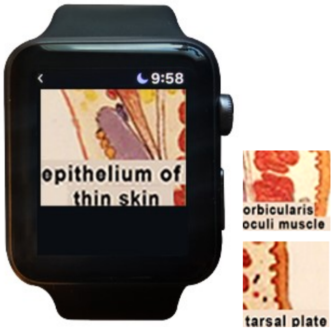
Response Information Mode

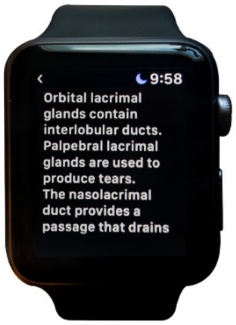


1. retinal pigment epithelial cell
2. cone
3. Muller



In the human eye, the lens helps to focus light and images on your retina. Ciliary processes are formed by inward folding and produce aqueous





Orbital lacrimal glands contain interlobular ducts. Palpebral lacrimal glands are used to produce tears. The nasolacrimal duct provides a passage that drains

material was required. While we had initially integrated the option for the learning material, but students could complete the task without it, so we removed it.

Procedure

After welcoming a participant and confirming their consent, we began to record their audio and video. Depending if they were right or left-handed, we adjusted the participant's seating position and explained the main procedure and purpose of the study to them. We provided each participant with a cheating scenario, helped them to wear the smartwatch and provide them with a short training session on the common interactions with the smartwatch, such as pressing the side button/tilting the arm to activate the display, tapping for selection/scrolling the content, and using the palm of one hand to place on the smartwatch face to de-activate the display. In the training, they were shown if the application crashed and how they could navigate back to the location of their response (answer cheat-sheet). This took 5 minutes.

Once ready, the exam moderator stayed in the room and provided the participant with the first component of the assessment and pressed the stop-watch button to start the trial. When the participant completed their task on answering the component, they were asked to inform the exam moderator to stop the stop-watch. The exam moderator noted the task completion time and then provided the participant with the second component. The same procedure was repeated for the remaining two components with a few seconds' break between the tasks. After participants completed the experiment, they were given a post-questionnaire that included 4 questions based on the Usability Metric for User Experience (UMUX), and an additional five open ended questions. We chose to use the UMUX, as it is a standardized, validated questionnaire and a close and concise variant of the System Usability Scale (SUS) for measuring the subjective responses of product's or system's perceived usability [14] [31]. The UMUX is a four-item Likert scale that measures the user experience based on usability components: effectiveness, satisfaction and overall efficiency [14]. The Participants were asked to provide their subjective ratings on the UMUX scale, from (0-Strongly Disagree) to (7- Strongly Agree). The total experiment took 45 minutes to complete. The UMUX 4 questions were varied for this study, these include:

1*. I think that I would like to use the smartwatch as a cheat-sheet again.*

2*. I found the smartwatch unnecessarily complex to use.*

3. I found it easy to read the text content on the watch.

4. I think that I would need the support of a technical person to be able to use the smartwatch.

For the picture content, the same questions were used, but ***3rd*** statement was replaced by: *I found it easy to read the picture content on the watch.*

The five open-ended questions added after the UMUX included:

1. Did you find the smartwatch more effective for cheating with Multiple Choice Questions or Short Answers and why?

2. Did you find the text or picture on the smartwatch more helpful for cheating and why?

3. What did you like or dislike about using the smartwatch to cheat?

4. Can you describe a real-life scenario where you think cheating on a smartwatch would be helpful for you or a friend might use it?

5. Have you been in exams, where the smartwatches are prohibited from the exam halls? If yes, do you agree with the policy and why?

Analysis

To test our hypothesis, we conducted a Shapiro Wilk test to assess the normality and a two-way repeated measures ANOVA to assess the effect of assessment format (MCQ vs. SA) and response format (Picture vs. Text) on the student's task success rate, time completion and the frequency of interactions. Shapiro-Wilk test revealed that our ten combinations violated hypothesis of normality, except for Time Completion for P_{MCQ} and T_{SA}. As our independent variables have only two levels, the Mauchly test could not confirm if the sphericity was violated and as ANOVAs are robust to deviations from normality, we proceeded to run parametric statistics. We defined the task success levels of the assessment format on a four-point scoring method by Tullis & Albert [34] (Table 3).

Table 3. Four-points Scoring Method for Task Success Levels

0 = Failure/gave up	The user provided all the wrong answers or gave up before completing the task.
1 = Major problem	The user struggled to find the answers and made a mistake on two questions.
2 = Minor problem	The user made one mistake, but was successful in completing the task.
3 = No problem	The user completed the task successfully without any difficulty.

For the overall usability ratings for Picture vs. Text, we converted the 11 points scale to the UMUX 7 points scale and recoded the odd item as [score – 1], and even items are scored as [7 – score]. To calculate the UMUX score, the sum of the four questions was divided by 24, and then multiplied by 100 (to make it equivalent to the SUS standard range of 0-100) [31]. The UMUX score highly correlates with the SUS scores, therefore systems that are scored above 68 are considered "above average" with a good usability experience, those below 68 are considered "below average", and is an indication that changes need to be incorporated to improve the product or system [31].

RESULTS

Task Success Rate Evaluation for Question Type

To understand which assessment format (MCQ vs. SA) was effective to cheat using the smartwatch, we compared their task success rate. The ANOVA revealed that there was a significant main effect of the assessment format (F (1,15) = 6.98, p = .002, η^2 = .074) on the student's success rate, but there was no main effect of response format (F (1,15) = .95, p = .035, η^2 = .025) and the interaction between the assessment format and response format (F (1,15) = .47, p = .05, η^2 = .005). A follow-up comparison indicated that participants could achieve a higher score rating on the assessment format for MCQ (M = 2.5, S.D. = .84) than SA (M = 1.9, S.D = .125). Students thus appeared to more effectively use the smartwatch to cheat on the MCQs than on the SAs.

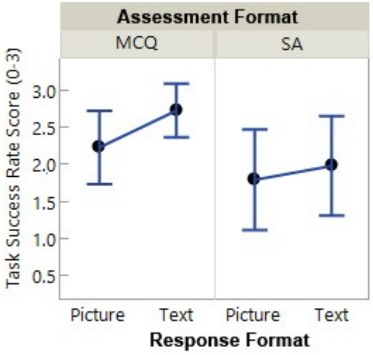


Figure 2. The Task Success Rating's mean and CI of ratings for MCQ vs. SA. Each error bar is constructed using a 95% CI of the mean.

Time Completion for Response Format

To understand which response format (Picture vs. Text) supported more efficient cheating with the smartwatch, we compared the time completion rate to answer the questions for (MCQ vs. SA). The Shapiro Wilk tests indicated that the distributions are not normal for T_{MCQ} (W = .74, p =.0005) and P_{SA} (W = .87, p =.003). The ANOVA revealed that there was no significant difference on the main effects, the assessment format (F (1,15) = 1.92, p = .19, η^2 = .03), response format (F (1,15) = .003, p = .099, η^2 = .0) and the interaction between the assessment format and the response format (F (1,15) = .004, p = .95, η^2 = .000004). Thus, students do not take a shorter period to complete the text questions than pictorial questions (Figure 3).

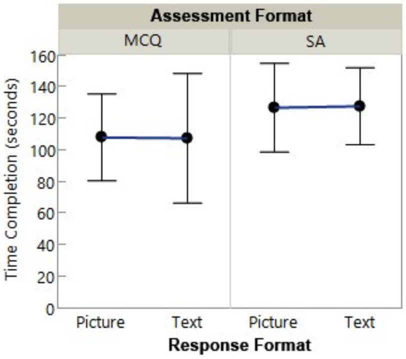


Figure 3. The Time Completion mean and CI of ratings for P vs. T. Each error bar is constructed using a 95% CI of the mean.

Interaction Effort for Assessment and Response Format

To analyze the physical effort to cheat in the assessment format vs. response format, we counted the frequency of participant's interaction (includes taps, swipes, pressing of side button/dial). The Shapiro Wilk tests shows that the distributions were not normal for P_{MCQ} (W = .73, p =.0004), T_{MCQ} (W = .71, p =.0002), P_{SA} (W = .75, p =.0006) and T_{SA} (W = .86, p =.018). The ANOVA revealed that there was no significant main effect of the assessment format (F (1,15) = .87, p = .37, η^2 = .016) and of the response format (F (1,15) = 1.32, p = .32, η^2 = .012) and the interaction between the assessment format and response format (F (1,15) = 1.33, p = .27, η^2 = .013). However, students do have a higher interaction with the smartwatch when completing the P_{SA} (M = 25.06, SD = 22.4) than T_{SA} (M = 17.56, SD = 15.4) as shown in Figure 4.

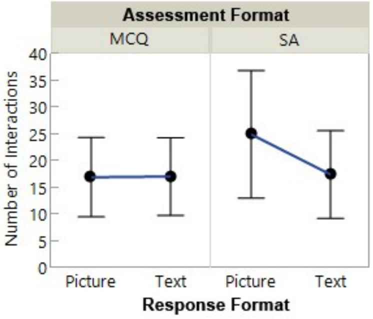


Figure 4. The Frequency of Interaction mean and CI of ratings for MCQ vs. SA and P vs. T. Each error bar is constructed using a 95% CI of the mean.

Overall Usability Rating

In contrary to our predictions that participants would experience a higher usability for the text response format as compared to a picture response format), the overall average UMUX score for Picture (M = 26, SD = 14.1) and Text (M = 27, SD = 14.7) were both below the average UMUX usability ratings of68 (Figure 5). This indicated that participant's overall subjective perception on the smartwatch usability to cheat in an assessment was neutral and required further improvement to the system (smartwatch and application).

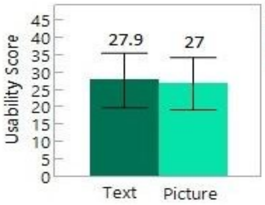


Figure 4. The UMUX mean and CI of ratings for P vs. T. Each error bar is constructed using a 95% CI of the mean.

Table 4. The UMUX

	Text	Picture
Min	8	8
Max	54	58
Range	46	50

Qualitative Reactions

Qualitative feedback from the participants revealed that as the SA required them to remember and spell the words, they made more mistakes in recognizing the answers.

"Short Answer was more effective for cheating, but for the picture answer, I had to go back and touch the screen often and zoom- which would get me caught." – P13, Female

A few participants faced challenges when reading answers from the display screen. As it automatically turned off, they had to continuously interact with it or use gesture to reactivate it.

"The screen kept turning off, and the display light was so bright, it made me nervous of being caught." – P10, Female

The majority of the participants said they preferred to organize and list the answers differently than the experiment as navigating was added an extra step for information search:

"*Navigating would be hard, especially if you don't know what information is going to be on the test."* – P10, Female

Based on the experience in the experiment, participants said that if they hypothetically were to cheat, then they liked how easy it was to cheat especially in larger exams hall with more students. Also, they would recommend the smartwatch to others for cheating.

"I liked that it was easy to hide and turn off very quickly then reactivate with the flick of a wrist. It made cheating very unobvious." – P9, Female

"I liked being able to answer quickly without stress." – P12, Female

Others felt that the button on the smartwatch to access the main menu were not supportive for cheating, as many participants wrote with their right hand, but the smartwatch was on the left.

"I really liked how easy it was to cheat. I somewhat disliked the small display, and having to touch the button on the left side, watch was also on the left) as I wrote with the right and I was in a hurry." – P15, Male

DISCUSSION

The purpose of the study was to address the research question on how effective, efficient, and usable smartwatches are to cheat in a classroom assessment such as an anatomy test. We examined the assessment format (MCQ vs. SA), and response format (Picture vs. Text) to test five hypotheses with students who did not have much experience with smartwatches, in cheating, and who did not have any prior knowledge about anatomy. In this section, we discuss the limitations of the study and its relation to the results.

For this study, we tried to provide some level of ecological validity by creating scenarios and simulating the laboratory setting to be as close and natural to a classroom, our study was limited by factors such as classroom size, the number of proctors and students, the type of course and the level of difficulty of the assessment, and the various technology permitted in a daily classroom setting. These factors may influence a student's behavior and reaction time to cheat, which would further influence the study's results. For instance, in a larger classroom with more students, and few proctors to physically monitor, there would be more opportunity to cheat using the smartwatch as compared to our laboratory setting. Nonetheless, we tried to help participants to mentally enact the situation as in a real-life assessment using scenarios that are based on everyday challenges that students face and can relate too; such as managing time, money, academic progress to achieve future goals and scholarships. Our qualitative feedback revealed that participants even in the un-real consequences, incentive, and set-up, were quite tensed of being caught and were motivated to achieve the correct answers.

Also, we did not limit the participants to any specific time, nor did we encourage them to cheat or try to influence their behavior in any way. We did instruct them that they were timed for each assessment, and informed them that they had the option to use the available answers on the smartwatch. Participants had the choice to guess, cheat or fail the experiment without using the smartwatch. From our observations of the recording, we found that majority of the participants made use of the opportunity to cheat and despite not having prior cheating experience were able to solve both Multiple-Choice Questions (MCQ) and Short Answers (SAs). **H1** was proven to be true i.e. students can effectively use the smartwatch to gain more correct answers on MCQ than SAs in the assessment.

While our use of scenarios did help students to emulate the classroom setup, the use of available answers on the smartwatch can be considered as impractical and a limitation of the study. We do agree that how students receive the answers is speculative (i.e. from the internet or libraries), but we disagree that these scenarios are impractical or have a low probability of occurrence as mentioned by a participant. The scenarios portrayed are a common practice for instructors to re-use exam questions from a previous term, which makes its probable for students to use smartwatches to get help from senior students or online libraries.

Our results also illustrate that students would use smartwatches and other smart devices if this can reduce their exam stress, and have the opportunity to improve their success. This implies that if these students have the motivation and flexibility to cheat with technology, they may even cheat when they become professionals in the workplace [10]. This means that smartwatches which are not designed to be a cheating tool, can be appropriated to cheat, and provide contextual and glanceable information in both text and pictures **(**confirming **H2).** Whereas, in terms of usability of the smartwatch, H3 and H4 were not supported. The data results revealed that students did not have a higher frequency of interaction with the smartwatch when completing the SA than MCQ (disproving **H3**), and students did not have a higher frequency of interaction with the smartwatch when searching for picture answers than text answers (disproving **H4**). However, we observed that the placement of the smartwatch and the task to write with the same hand had an impact on the frequency of interaction. For instance, participants who had worn the smartwatch and wrote the answers with the same hand were more likely to put extra physical effort to interact with it. Using glances discreetly or pressing the button to navigate to the main menu was rendered difficult. Others who were using different hands for each task could interact seamlessly, but they were not satisfied with the side button/dial to interact with the smartwatch.

To maintain academic integrity, one possible solution to deter cheating is to use more constructed assessment format, which requires students to recall information and create their own answers [10]. We have explored one possible constructed response format such as short answers (SA), however based on the experiment we were able to confirm that the smartwatches are sensitive to differences in the efficiency of different response formats **(H4)**. Our data results for **H4** revealed that there is a high interaction with the smartwatch when completing the Short Answers for Picture than text (P_{SA} vs T_{SA}). Thence future work could test other formats that require participants to provide their input and tests their knowledge, skills and frequency of interaction to cheat on the smartwatch. Also, using our data results for **H3** and **H4,** we suggest future researchers to explore how the frequency of interactions is related to the placement of the smartwatch. These insights could help in exploring in-depth how cheating can be deterred for long term, but also developing constructive hands-free interaction for educational purposes.

In terms of usability rating, we did not assess the usability rating for each component of the assessment and did not explore in-depth qualitative feedback on the participants use of the response formats. This includes the challenges in the social layout of the picture and text format and the possibility of discovering an answer for a previous question. Future work could take this limitation into consideration and test the ratings on individual components, and elicit more comprehensive feedback. This could help to understand the reason for a below average usability rating and improvements of smartwatches for educational purposes.

Future work could also consider including a broader range of stakeholders such as instructors, left-handed participants, cheaters vs non- cheaters and students from an anatomy course. Having instructors would help to increase the validity of the other assessments formats and the number of questions (e.g. what combination of SAs vs MCQs would be suffice to test a student's knowledge). While, the comparison of different set of groups could help us determine what kind of participant's background and experience could affect the results. For instance, in our recordings, we observed that the detection of cheating committed by the left-handed participants was more difficult to capture. In our setup we made adjustments to accurately replicate and capture left-hand participants, but we realized that in a daily academic scenario (such as the virtual distant learning), left-handed student have the advantage to cheat more flexibly as the camera position does not capture all angles and there are no proctors to physically monitor [4]. With our limited number of participants, we could not make a comparison of the different set of participants.

Our last limitation is the generalizability of our study. Our laboratory setting is limited to real life setups with one proctor and one participant and to assessment that are similar to an everyday classroom assessment. While, it is important to focus on high stakes examination, we chose to examine an everyday assessment as it emulates an actual classroom setup, where the students are tested for questions that include both picture and text, and where smart devices are not prohibited to be used. These real life setups can include make-up exams (assessments that instructors give students to make up for their grades), virtual classrooms, and professional assessments for pilots, police officers, or firefighters [28]. A makeup assessment is similar to our laboratory setup, however in the virtual classrooms there are no proctors and the camera only records the student's screen and their headshot. Therefore, students have more flexibility to interact with the smartwatch without being caught. The same applies for assessments for professional (such as pilots, police officers), where they may not have strict rules for being penalized for interacting with the mobile phones or smartwatches when they perform their assessment [10]. With advances in wearable technology and the challenges of different classroom setups (virtual vs physical); cheating will become more prevalent and difficult to detect. This will also impact the learning environment of students and work ethics of professionals; thus, future work would need to consider these limitations as a basis to understand cheating behavior and further research how to maintain academic integrity with other wearable devices.

CONCLUSION

Our study contributes to the research on academic integrity and the growing demand of wearable technologies. We conducted a laboratory experiment and focused on one wearable (i.e. smartwatches) to test how students can appropriate them to cheat in an anatomy assessment. We examined the effectiveness, efficiency and usability of the smartwatch based on our five hypotheses. We discovered that smartwatches even with low usability rating and efficiency to interact with the two-response format (Picture vs Text), students with no prior cheating experience or knowledge of assessment can cheat effectively. Our study highlights that students and professionals would be motivated to cheat and use the opportunities if it reduces stress and helps to gain grades or monetary benefits. Therefore, our study emphasized the need to explore how cheating can be deterred using: (a) more constructive response formats, (b) further investigation on the interactions for Short Answers (SA) for the two response formats (Picture and Text) and (c) the frequency of interaction in relation to the position of the smartwatch. We also suggested including a broader number of stakeholders (e.g. cheaters vs non-cheaters, left-handed vs right-handed, instructors) and testing with other wearables.

REFERENCES

1. 24kupi. 2017. Buy Cheating Watches for Easy Studying, exams and tests. Retrieved January 26, 2017 from https://www.24kupi.com/webshop-en/?___store=default&___from_store=default

2. Deepak Akkil, Jari Kangas, Jussi Rantala, Poika Isokoski, Oleg Spakov, and Roope Raisamo. 2015. Glance Awareness and Gaze Interaction in Smartwatches. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '15), 1271–1276. https://doi.org/10.1145/2702613.2732816
3. Peter Ashworth, Philip Bannister, Pauline Thorne, and Students on the Qualitative Research Methods Course Unit. 1997. Guilty in whose eyes? University students' perceptions of cheating and plagiarism in academic work and assessment. *Studies in Higher Education* 22, 2: 187–203. https://doi.org/10.1080/03075079712331381034
4. Mebratu Mulatu Bachore. 2014. Academic Dishonesty/ Corruption in the Period of Technology: *Its implication for Quality of Education*. *American Journal of Educational Research, American Journal of Educational Research* 2, 11: 1060–1064. https://doi.org/10.12691/education-2-11-9
5. J. A. Ballantine, P. McCourt Larres, and M. Mulgrew. 2014. Determinants of academic cheating behavior: The future for accountancy in Ireland. *Accounting Forum* 38, 1: 55–66. https://doi.org/10.1016/j.accfor.2013.08.002
6. Christopher Bearman, Susannah B. F. Paletz, Judith Orasanu, and Matthew J. W. Thomas. 2010. The Breakdown of Coordinated Decision Making in Distributed Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, 2: 173–188. https://doi.org/10.1177/0018720810372104
7. Gerald Bieber, Thomas Kirste, and Bodo Urban. 2012. Ambient Interaction by Smart Watches. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments* (PETRA '12), 39:1–39:6. https://doi.org/10.1145/2413097.2413147
8. Marta E. Cecchinato, Anna L. Cox, and Jon Bird. 2015. Smartwatches: The Good, the Bad and the Ugly? In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '15), 2133–2138. https://doi.org/10.1145/2702613.2732837
9. Xiang "Anthony" Chen, Tovi Grossman, and George Fitzmaurice. 2014. Swipeboard: A Text Entry Technique for Ultra-small Interfaces That Supports Novice to Expert Transitions. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (UIST '14), 615–620. https://doi.org/10.1145/2642918.2647354
10. Gregory J. Cizek. 1999. *Cheating on Tests: How To Do It, Detect It, and Prevent It*. Routledge, Mahwah, N.J.
11. Gareth Denyer and Dale Hancock. 2012. Multiple choice questions to combat plagiarism and encourage conceptual learning. *Proceedings of The Australian Conference on Science and Mathematics Education (formerly UniServe Science Conference)* 0, 0. Retrieved December 13, 2016 from http://openjournals.library.usyd.edu.au/index.php/IISME/article/view/6380
12. Steven M. Downing. 2002. Threats to the Validity of Locally Developed Multiple-Choice Tests in Medical Education: Construct-Irrelevant Variance and Construct Underrepresentation. *Advances in Health Sciences Education* 7, 3: 235–241. https://doi.org/10.1023/A:1021112514626
13. Jean H. D. Fasel. 1998. Teaching of gross anatomy to medical undergraduates: general practice as a guideline? A synopsis. *Journal of Anatomy* 192, Pt 2: 305–306. https://doi.org/10.1046/j.1469-7580.1998.19220305.x
14. Kraig Finstad. 2010. The Usability Metric for User Experience. *Interacting with Computers* 22, 5: 323–327. https://doi.org/10.1016/j.intcom.2010.04.004
15. Megan-Jane Johnstone. 2016. Academic Dishonesty and Unethical Behaviour in the Workplace. *Australian Nursing & Midwifery Journal* 23, 11: 33–33.
16. A. Komninos and M. Dunlop. 2014. Text Input on a Smart Watch. *IEEE Pervasive Computing* 13, 4: 50–58. https://doi.org/10.1109/MPRV.2014.77
17. Philip Kortum. 2008. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Morgan Kaufmann.
18. Rebekah D. LaDuke. 2013. Academic Dishonesty Today, Unethical Practices Tomorrow? *Journal of Professional Nursing* 29, 6: 402–406. https://doi.org/10.1016/j.profnurs.2012.10.009
19. McCabe DL. 2009. Academic dishonesty in nursing schools: an empirical investigation. *Journal of Nursing Education* 48, 11: 614–623. https://doi.org/10.3928/01484834-20090716-07
20. Robert S. Mccuskey, Stephen W. Carmichael, and Darrell G. Kirch. *Article The Importance of Anatomy in Health Professions Education and the Shortage of Qualified Educators*.
21. Alex Migicovsky, Zakir Durumeric, Jeff Ringenberg, and J. Alex Halderman. 2014. Outsmarting Proctors with Smartwatches: A Case Study on Wearable Computing Security. In *Financial Cryptography and Data Security*, Nicolas Christin and Reihaneh Safavi-Naini (eds.). Springer Berlin Heidelberg, 89–96. https://doi.org/10.1007/978-3-662-45472-5_7
22. Vivian Genaro Motti and Kelly Caine. 2016. Smart Wearables or Dumb Wearables?: Understanding How Context Impacts the UX in Wrist Worn Interaction. In *Proceedings of the 34th ACM International*

Conference on the Design of Communication (SIGDOC '16), 10:1–10:10. https://doi.org/10.1145/2987592.2987606

23. Michael Nebeling, Alexandra To, Anhong Guo, Adrian A. de Freitas, Jaime Teevan, Steven P. Dow, and Jeffrey P. Bigham. 2016. WearWrite: Crowd-Assisted Writing from Smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 3834–3846. https://doi.org/10.1145/2858036.2858169
24. Stephen Oney, Chris Harrison, Amy Ogan, and Jason Wiese. 2013. ZoomBoard: A Diminutive Qwerty Soft Keyboard Using Iterative Zooming for Ultra-small Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 2799–2802. https://doi.org/10.1145/2470654.2481387
25. Eun-Jun Park, Seungmi Park, and In-Sun Jang. 2013. Academic cheating among nursing students. *Nurse Education Today* 33, 4: 346–352. https://doi.org/10.1016/j.nedt.2012.12.015
26. Simon T. Perrault, Eric Lecolinet, James Eagan, and Yves Guiard. 2013. Watchit: Simple Gestures and Eyes-free Interaction for Wristwatches and Bracelets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 1451–1460. https://doi.org/10.1145/2470654.2466192
27. Cornelius B. Pratt and Gerald W. McLaughlin. 1989. An Analysis of Predictors of College Students' Ethical Inclinations. *Research in Higher Education* 30, 2: 195–219.
28. Jakob Rodgers. 2012. AFA discovered cheating by comparing online, final exams. *Colorado Springs Gazette*. Retrieved February 6, 2017 from http://gazette.com/afa-discovered-cheating-by-comparing-online-final-exams/article/139894
29. Cristóbal Romero, Rebeca Cerezo, Jose Antonio Espino, and Manuel Bermudez. 2016. Using Android Wear for Avoiding Procrastination Behaviours in MOOCs. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale* (L@S '16), 193–196. https://doi.org/10.1145/2876034.2893412
30. Sapa-AFP. 2012. Cops expose exam cheating ring in Bangladesh. *The M&G Online*. Retrieved December 1, 2016 from http://mg.co.za/article/2012-10-23-pupils-in-bangladesh-should-get-full-marks-for-innovative-way-of-cheating-during-exams/
31. Jeff Sauro. 2011. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*. CreateSpace Independent Publishing Platform.
32. Hassan Sami Shaibah and Cees P. M. van der Vleuten. 2013. The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy. *Anatomical Sciences Education* 6, 3: 149–156. https://doi.org/10.1002/ase.1325
33. Judy Sheard and Martin Dick. 2012. Directions and Dimensions in Managing Cheating and Plagiarism of IT Students. In *Proceedings of the Fourteenth Australasian Computing Education Conference - Volume 123* (ACE '12), 177–186. Retrieved November 30, 2016 from http://dl.acm.org/citation.cfm?id=2483716.2483737
34. Thomas Tullis and William Albert. 2010. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann.
35. Maureen Wideman. 2011. Caring or collusion? Academic dishonesty in a school of nursing. *The Canadian Journal of Higher Education; Toronto* 41, 2: 28–43.
36. Robert Xiao, Gierad Laput, and Chris Harrison. 2014. Expanding the Input Expressivity of Smartwatches with Mechanical Pan, Twist, Tilt and Click. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14), 193–196. https://doi.org/10.1145/2556288.2557017
37. 2016. Phone and Electronic Device Policy. *SAT Suite of Assessments*. Retrieved December 1, 2016 from https://collegereadiness.collegeboard.org/sat/taking-the-test/phone-electronic-device-policy