Research Report

Scoring Essays on an iPad Versus a Desktop Computer: An Exploratory Study

ETS RR–22-08

Guangming Ling
Jean Williams
Sue O'Brien
Carlos F. Cavalie

December 2022

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Scoring Essays on an iPad Versus a Desktop Computer: An Exploratory Study

Guangming Ling, Jean Williams, Sue O'Brien, & Carlos F. Cavalie

ETS, Princeton, NJ

Recognizing the appealing features of a tablet (e.g., an iPad), including size, mobility, touch screen display, and virtual keyboard, more educational professionals are moving away from larger laptop and desktop computers and turning to the iPad for their daily work, such as reading and writing. Following the results of a recent survey of individuals who serve as ETS raters, more than 40% reported that they would prefer to use an iPad or other type of tablet to score essays. However, iPad-based essay scoring could affect scoring accuracy and scoring time because the smaller screen and other features of an iPad may also affect raters' reading comprehension and score assigning processes. To address this issue, we invited 10 experienced raters to score holistically 40 essays for a graduate admission test using a desktop computer and an iPad following a counterbalanced design. We compared the raters' scores against the criterion scores and analyzed scoring times, scoring behaviors, and raters' answers to a structured interview after the scoring experiment. The results reveal no obvious differences between the two devices in the scoring accuracy or average scoring time per essay, which suggests that scoring on an iPad may not reduce scoring quality or scoring productivity for essays that are holistically scored as compared to scoring the essays on a desktop computer. We also found a few iPad-specific issues that raters reported, including issues associated with the invisible scrolling bar and the extra scrolling needed to reach the score-assignment panel, difficulty navigating between the prompt and the essay response, and oversensitivity of the touch screen.

Keywords Desktop computer; iPad; essay scoring; accuracy; rating time; scrolling

doi:10.1002/ets2.12349

In the last three decades, advancement in computer- and internet-related technology has greatly energized educational testing. In the new Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), technology-related issues are described and discussed for almost all stages of a test, especially during test development, test administration, test scoring, score reporting, and score use and interpretation. Traditionally, constructed responses (such as essays) were delivered and responded to with paper and pencil and were scored on paper by raters. Since the 1990s, Pearson NCS has processed National Assessment of Educational Progress (NAEP) writing tests through online scoring using computers (Bennett, 2003); ETS has also implemented the Online Scoring Network (OSN) to score a variety of tests with constructed-response items. There are several advantages associated with scoring on a computer or online compared to traditional scoring on paper. For example, Bennett suggested that one advantage of online scoring of constructed responses (such as essays or answers to open-ended math problems) is that responses can be easily distributed to raters around the world, as long as the raters can access the secured online system through the internet. Other advantages include improved security and enhanced quality control of the scoring process, especially for identifying and flagging responses that have discrepant scores between raters and therefore may need to be rescored by an adjudicator (Falvey & Coniam, 2010; Raikes et al., 2004). In 2013, a new online-scoring platform, Online Network for Evaluation (ONE), was launched at ETS, and several testing programs have since moved to this new scoring system. The new system mimics most of the scoring features of the OSN (Drasgow et al., 2006, p. 495). Currently the ONE system does not allow the use of devices with a small screen, such as the iPad, to score essays for operational purposes but can implement a small-scale study where the iPad can be used for research purposes.

Given the rapid change in technology and its uses, along with technology's pervasive impact on people's reading habits (e.g., Zickuhr et al., 2012), it seems likely that many raters may prefer to use an iPad or other type of tablet to read and assign scores to constructed responses. A recent internal survey of close to 1,000 ETS raters indicated that about

Corresponding author: G. Ling, E-mail: gling@ets.org

40% said they would like to use an iPad or other type of tablet to score essays. However, rating essays on a tablet may pose a fundamental threat to the scoring quality of constructed responses, introduce construct-irrelevant variances, or raise fairness issues. We intended to provide some preliminary information to address these concerns with a small-scale experiment.

After a brief search of related literature, we could not find any study looking directly at the possible effects on rating quality when moving from scoring using a desktop computer to scoring using a tablet. However, typical behaviors related to scoring essays on an iPad may be compared to doing the same task on a desktop computer. A handful of studies were found to have investigated the comparability of on-screen scoring (or online scoring) and paper scoring. In addition, a few studies investigated reading comprehension based on an electronic screen versus on paper. We summarize these studies in the following pages.

When reading content is presented on a computer screen, the content and format to be presented, the screen size, the font, and the resolution of the screen can directly determine the amount of scrolling required to read and answer questions during a test. Studies have been conducted to investigate these factors and their effects on reading comprehension and reading test scores. Although earlier studies have shown that screen size (e.g., a 60-line screen vs. a 20-line screen) did not affect reading comprehension (Richardson et al., 1989), more recent research has suggested that screen size could affect reading. For example, Bridgeman et al. (2003) found that scores on a multiple-choice reading test tend to be lower when using displays with a lower resolution and smaller screen size than when using displays with a higher resolution and a larger screen size (i.e., with a practically small effect size of .25). Other studies found negative effects associated with scrolling while reading on an electronic screen for test takers (Choi & Tinkler, 2002; Kingston, 2009; Pommerich, 2004; Sanchez & Wiley, 2009).

For scoring purposes, as Bennett (2003) summarized, there seems to be no difference in average essay scores, inter-rater agreement, or raters' passing rates when handwritten essays are scored on paper or on a screen after being scanned (Zhang et al., 2003, for several titles of Advanced Placement tests). Similar findings were also seen in the NAEP writing test (Johnson, 1993); the Graduate Management Admission Test writing test (Powers et al., 1997); *THE PRAXIS SERIES*® assessments (Powers & Farnum, 1997); and, more recently, the Writing Paper of Hong Kong Certificate of Education (Coniam, 2009; Falvey & Coniam, 2010). In these studies, handwritten responses were scanned and put on the computer, which is very different from having someone respond directly on the computer. Studies have shown that when writing an essay on computers, test takers tend to write more, which, at a minimum, can certainly impact the time for scoring each essay.

Two other studies examined differences in reading comprehension related to small-screen devices (iPad and mobile phone). Singh et al. (2011) compared the reading comprehension of 50 participants reading a Facebook privacy policy on a desktop computer screen and on a mobile screen (e.g., an iPhone screen), with reading comprehension measured by a cloze test following the reading session. They found that 39% of participants who read the content on a desktop screen answered at least 60% of items correctly on the postcloze test, while only 19% of those who read on a mobile screen did so. Ling (2015) compared reading comprehension for 403 eighth graders on a test assembled from NAEP-released items. He found that using a desktop computer (17-inch screen) or an iPad (10-inch screen) did not affect students' reading comprehension scores.

In summary, some previous studies revealed the impact of various devices on reading, and other studies suggested that on-screen scoring has a level of quality and reliability comparable to on-paper scoring. However, the generalizability of these findings to the context of scoring responses on a tablet is questionable. It is reasonable to argue that scoring using a tablet may make raters more susceptible to fatigue and thus cause them to produce fewer ratings or lead to a lower level of rating quality. For example, reading an essay with 400 words displayed on a 17-inch screen is likely to require little or no scrolling, while reading the same essay displayed on a smaller screen may require some scrolling because it is likely that the rater will need to make necessary adjustments to format and font for the smaller screen so that the reading interface is clear enough. The increase in scrolling with smaller screens may make raters more susceptible to feeling fatigued than they would be using a computer screen. On the other hand, a reasonable counterargument may be that using a tablet would result in less fatigue because raters could choose to sit on a more comfortable chair or sofa rather than having to sit at a desk to score using a desktop computer. Finally, a question remains whether using a tablet would pose a threat to the safety and security of the scoring system, as the ultra-portability of a tablet may make scoring possible anywhere or anytime, instead of scoring in a confined, quiet, and, most importantly, secure environment. This study was aimed at addressing concerns related to fatigue and scoring quality, which could weaken the validity of score uses and inferences. The third security-related concern would require investigations employing different methods and procedures than the current study did.

In this study, we examined whether raters' scoring behaviors, including scoring quality and productivity, were associated with using an iPad or a desktop computer. We used the average scoring time per essay to quantify rating productivity. A short scoring time per essay on average means faster and more productive scoring, whereas a longer scoring time per essay suggests the opposite. The rating quality was quantified by two statistics: the proportion of agreement with the criterion scores (to be described later) and the difference between the criterion score and the rater-assigned score (see the section "Method").

Research Questions

Three specific research questions were explored:

1. Do raters have any difficulties or challenges when scoring on an iPad?
2. Do raters spend a longer time on average scoring an essay on an iPad than on a desktop computer?
3. Do raters make more scoring errors (or have a lower scoring accuracy) when scoring on an iPad than on a desktop computer?

Method

Raters

A recruiting e-mail was distributed to the ETS essay raters working for testing programs at the college level or beyond, which resulted in 10 raters who voluntarily participated in this study. These 10 raters had either a master's or a doctoral degree. Four raters were male, and six were female. Nine of the 10 raters identified as White, and one identified as African American. Eight raters were or had been essay raters for the graduate admission test, with four of them having more than 2 years of essay scoring experience with the test. Two raters had experience scoring essays from other types of college-level tests. All 10 raters were familiar with the ETS scoring process and the use of the ONE platform to score essays. Five raters confirmed that they normally used desktop computers for scoring and other types of daily work; four used laptop computers for daily scoring and other types of work; and one used an iPad in addition to desktop and laptop computers in her daily work, but an iPad was not used for scoring. Only one rater reported using the iPad very often, while the others reported never ($n = 2$), rarely ($n = 3$), or sometimes ($n = 4$) using an iPad. Eight of 10 raters had experience with an iPad and knew how to use it to read, navigate, and score.

Essay and Criterion Scores

A total of 40 essays were selected from operational responses to a graduate admission test. All of the essays were based on the same prompt extracted from the test. Each essay was selected in a way that it had a clear and unambiguous score (McClellan, 2010) at a particular score level between 1 and 6. This means the scoring decisions on these essays were relatively easier than borderline essays, which are more likely to be scored with disagreement between raters. The scores of the selected essays were reviewed and confirmed by a panel of three expert raters prior to the scoring experiment and were then used as the criterion scores (or validity scores) to evaluate the rating accuracy in this study.

Half of the essays were relatively long, having more than 400 words. The other half were shorter, with each having fewer than 300 words. The 40 essays were divided into two sets (labeled as the first and second sets; see Table 1). Each set consisted of 20 essays, including 10 short essays and 10 long essays. Each set of essays had the same distribution of criterion scores and the same number of essays at each score level (see Table 1) and had the same mean and standard deviation as well ($M = 3.62$, $SD = 1.43$). In each set, there were more essays in the middle of the scale and fewer essays with an extreme score. During the scoring sessions, the essays were presented in a fixed order, regardless of rater or device.

Table 1 Criterion Score Distribution of Essays in Each Set

	Score level					
No. of essays	1	2	3	4	5	6
First set	2	3	5	5	3	2
Second set	2	3	5	5	3	2

Instruments

Scoring Platform

The essays were delivered to the raters through the ETS ONE system. All the raters had used the system before this study, and they were all familiar with the system and interfaces. The ONE system in this study was set as an interactive training session rather than an operational scoring session to avoid confusion with ongoing operational scoring. The scoring experiment was arranged in a cognitive lab at the Princeton, New Jersey, office of ETS. The total time for the scoring was about 2 hours, plus about a half hour for the introduction, the prior background survey, and the post hoc interview. The system interface was piloted on an iPad prior to the scoring experiment to ensure its comparability with the desktop computer or laptop computer.

Devices

Each rater scored 20 essays on an iPad (with a 10-inch touch screen but no external keyboard or external pointing device) and another 20 essays on a desktop computer (with a 20-inch screen, a physical keyboard, and a physical mouse). Both devices had a comparable processing speed, with no obvious lapse or delays for regular navigation using either device.

Background Questionnaire

A background questionnaire was designed to collect raters' gender; ethnicity; education level; prior scoring experience; and experience with an iPad, a desktop computer, and a laptop computer (see Appendix A).

List of Behaviors to Observe

Scoring an essay on a desktop computer typically involves at least four sets of behaviors: (a) logging in to the system platform and navigating to the page of the designated scoring session; (b) selecting and activating the first essay on the desktop or laptop screen using a physical mouse or other type of pointing device (a touch pad or a pointing stick, often seen as a red point on a Lenovo laptop keyboard); (c) reading the essay, recalling the scoring rubrics, and deciding the score level for the essay, with necessary scrolling up and down to read the essay and cross-check the scoring rubrics or sample essays; (d) scrolling to the scoring panel and selecting and submitting the score; and (e) repeating Steps b–d for the following essays and exiting the system after completion of the scoring tasks. It should be noted that no typing (e.g., putting in a score) is needed as part of the scoring process. It is possible, though it rarely happens, that raters may need to communicate with scoring leaders using instant messaging, but more often, such communications are completed via telephone in practice.

Scoring on an iPad using the same system typically involves these steps, but with two major differences: (a) The screen to view and score the essay is much smaller on an iPad than on a desktop or laptop screen and (b) the navigation, selecting, browsing, reading, and score submission are completed by using a finger or multiple fingers to tap the iPad screen. It is also likely that raters can hold the iPad at various positions that they feel are more comfortable, while those using a desktop computer need to stick with a monitor screen that is set in a fixed location and position. With this information, we decided to focus on behaviors related to these hypothetical differences during the observation and post hoc interview.

A list of eight specific behaviors was prepared for observation purposes as raters scored using each device. The behaviors included whether the rater changed any display or reading preferences after logging in or moving to the scoring page (such as changing font size or changing to full screen) and whether the same preferences change were made on both devices,

Table 2 Scoring Design

Rater group	First set (20 essays)	Second set (20 essays)
A	Using an iPad	Using a desktop
B	Using a desktop	Using an iPad

how the rater moved around the screen using each device, how the rater scrolled using each device, whether the rater made any movements or gestures that indicated unfamiliarity, and whether the rater asked questions related to using a particular device (see Appendix B).

Interview Questions

A short survey was designed to collect raters' opinions of and reactions to the scoring using different devices. The questions included whether raters preferred using a particular type of device for daily reading, work-related reading, and essay scoring; whether raters noticed differences in the lengths of essays when using different devices to score; and whether scoring longer essays using the iPad posed more difficulties or challenges (see Appendix C).

Procedures

Each scoring session had only one rater and one to two observers. It started with a brief introduction to the rater about the purpose and steps of the study and was followed by a set of background questions. The rater was then given the computer or iPad to start the scoring. A 5-minute break was taken once the first session was completed. During this break, the interviewer prepared for the next session. Within each scoring session, the interviewer observed the scoring process and logged scoring behaviors following the predetermined list of behaviors to observe (see Appendix B). Once the second session was completed, an interview session was immediately conducted, posing the predetermined interview questions (see Appendix C).

The raters who were not familiar with the iPad received a brief training session immediately before their first scoring session. The training focused on how to use the touch function of the iPad to enter the scoring system, scroll up and down, set up the reading environment, and assign a score. The raters were allowed to move on to the scoring experiment once they told the interviewer that they felt comfortable with using the iPad.

The 10 raters were randomly divided into two groups, with 5 raters in each group. In Group A, each rater was asked to score 20 essays in 1 hour using an iPad; this was followed by a short interview (10–20 minutes). Raters were then asked to score the second set of 20 essays for another hour using a desktop computer; this was again followed by a 10- to 20-minute interview. The raters in Group B were asked to score the first set of essays using a desktop computer and then to score the second set of essays with an iPad (see Table 2).

Data and Analysis

Each essay had three sets of scores: the scores assigned by the raters in Group A using one of the two devices, the scores assigned by the raters in Group B using the other device, and the criterion scores. The scoring time for each essay was recorded and used to examine any differences in the average scoring time per essay between the two devices.

Qualitative analysis and descriptive analysis were performed to summarize raters' interview results in order to address the first research question. The average scoring time per essay using each device was computed, and these were compared to answer the second research question.

For the third research question, interrater agreement indices, including percentages of exact agreement rates (with a discrepancy of 0) and adjacent agreement (with a discrepancy up to 1 point), Pearson product–moment correlation, and quadratic-weighted kappa (QWK), were computed for ratings assigned by each rater on a particular device against the criterion scores. In addition, descriptive analysis was applied using score errors. A score deviation was computed using the rating assigned by a particular rater minus the criterion score for a particular response. The score deviation was then squared and added up within each device for each rater and then divided by the number of ratings within each device,

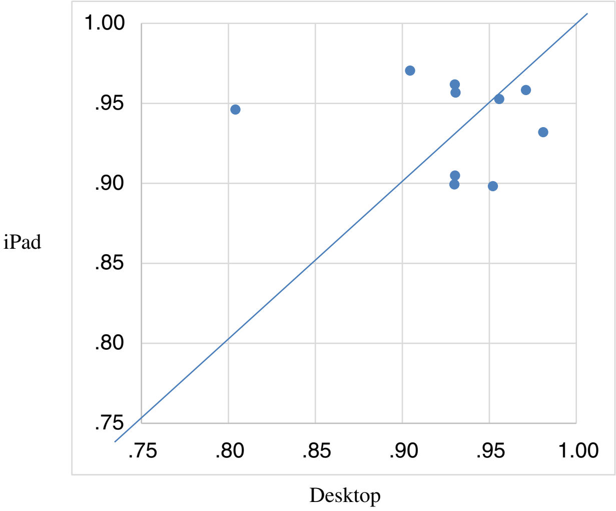

Figure 1 The plot of correlations between rater-assigned score and criterion score by device.

which was treated as the device-specific mean square scoring error[1] (MSE) for a particular rater. Then the mode-related MSE difference for this rater was computed using the MSE for iPad minus that for a desktop computer.

Results

Scoring Results

When scored on a desktop computer, the mean score ($M = 3.55$) was slightly lower than the criterion scores, but also with less variation ($SD = 1.34$). The criterion scores mean and SD were 3.62 and 1.43, respectively, for both desktop- and iPad-based essays. When scored on an iPad, the mean was slightly higher on average ($M = 3.63$) but slightly less varied ($SD = 1.40$) than the criterion score. The iPad-based scores were slightly higher and more varied than the desktop-based scores. However, the mean score differences were all trivial and negligible, less than 10% of the pooled standard deviation.

The 10 raters had an overall exact agreement proportion of .72 with the criterion scores and an overall adjacent agreement of .27, which totals up to .99. If both exact and adjacent agreements are considered acceptable, the overall rating accuracy in this study appears to be adequately high. Variations exist among the raters on these proportions—between .62 and .81 on the exact agreement and between .38 and .19 on the adjacent agreement—but they are quite similar when the two are combined, between .95 and 1.00. The ratings based on an iPad and those on a desktop computer had comparable exact agreement rates (.71 and .72, respectively) and adjacent agreement rates (.28 and .27, respectively).

The correlation between the ratings assigned in this study and the criterion scores were between .87 and .96 among raters, with an average of .92. The ratings assigned using an iPad correlated with the criterion scores between .90 and .97 among raters, with an average of .93 overall, and the ratings assigned on a desktop computer correlated with the criterion scores between .80 and .98, with an average of .92. The two sets of correlations are plotted against each other in Figure 1, where each dot represents one rater, the x-axis value represents the correlation coefficient on a desktop, and the y-axis value represents that on an iPad. As the plot shows, for all raters except one, the two devices resulted in similar correlations above .90. One rater appears to be an outlier (leftist dot in Figure 1), with a correlation coefficient on a desktop of .80, which was much lower than that on an iPad (.95) by the same rater or that on a desktop computer by any one of the other raters.

The QWK was also comparable, with .92 for the iPad-based rating with the criterion scores and .93 for the desktop-based rating with the criterion scores, and .92 overall (see Table 3). Using the two types of devices resulted in similar times to assign a score. Overall, raters spent 95 seconds on average scoring each essay, with a standard deviation of 1.89. The average time per essay was also comparable, at 95 seconds on the iPad and 94 seconds on a desktop computer (see Table 3).

The MSEs were between .143 and .571 among 10 raters for the ratings assigned on an iPad and between .095 and .762 for ratings assigned on a desktop across raters. The MSE difference (iPad minus desktop) was −.005 on average across raters ($SD = .239$).

Table 3 Scoring Accuracy and Time by Device

Rater group	Exact agreement	Adjacent agreement	Overall agreement	Pearson correlation	QWK	Response time(s)
iPad-based	0.71	0.28	0.99	0.93	0.92	95
Desktop-based	0.72	0.27	0.99	0.92	0.93	94
All ratings	0.72	0.27	0.99	0.92	0.92	95

Note. QWK = quadratic-weighted kappa.

Table 4 Summary of Observed Scoring Behaviors

	No. of raters with device	
Issue with the device	Desktop	iPad
Adjust/change font or background	1	4
Adjust browser window		1
Adjust iPad position		0
Scrolling and navigating issues		3
Pop-up window issue		1
Difficulty with selecting a score		1
Touch screen sensitivity		1
Locating cursor on the iPad screen		2
Viewing one longer essay on one screen		1
Scrolling bar issues		2
Switching between scoring screen and benchmark essays		1

The ratings that were scored in the first set had a lower exact agreement rate of .65 than those in the second set (.79), which was compensated by the higher adjacent agreement rate in the first set than that in the second set of ratings (.33 vs. .21). However, the first set of ratings also took a longer time than the second set on average (98 vs. 90 seconds) and had a smaller standard deviation (*SD* = 1.78 vs. 1.90).

Observed Scoring Behaviors

Because raters were used to scoring using the desktop or laptop computer, we did not receive any questions when they used the desktop computers during the experimental scoring sessions, nor did we observe any navigation movement that could be attributed to unfamiliarity or difficulty. Only one rater changed the font size on the desktop; another made changes to the browser screen size (e.g., minimizing it) on the desktop computer (see Table 4).

All raters used the preset position of the iPad, which stood at a 45° angle facing the rater. No rater attempted to change the position or even asked about it. Only four raters used the pinch and zoom features on the iPad, while the others did not make any changes to the display on the iPad.

Seven raters seemed to work efficiently using the iPad to scroll and navigate. When scrolling on the iPad, three raters initially showed some level of unfamiliarity with iPad-based scrolling, but they soon got used to it. Only one rater seemed to have continuous difficulty when scoring the essays using the iPad. One participant struggled with closing the pop-up window that displayed the rating guide. Two raters displayed some difficulty making a scoring selection using the iPad. One in particular kept her finger on the screen for too long so that it occasionally led to a pop-up menu. Along the same lines, one participant was confused by the lack of reaction from the iPad when she touched it by using her fingernail. Still, she appeared to get used to using the skin of her finger after being given appropriate instruction (see Table 4).

Two raters asked about the location of the scrolling bar on the iPad screen at the beginning of the scoring session. One rater asked if she could navigate away from the essay to the benchmarks and scoring guide on the iPad. This rater also asked whether she could see a complete essay response on one screen on the iPad (with no scrolling required). This rater also asked whether the zoom effects would remain if she zoomed in on one essay and then moved on to the next essay (see Table 4).

Table 5 Summary of Rater Perceptions During the Experiment

Rater perception	No. of raters
Experiment session was comfortable	7
Scoring environment is different from daily scoring work (e.g., size and position of chair and desk, lighting, screen size)	10
Font of the screen display does not affect scoring	10
Scoring on an iPad seems easier (than on a desktop computer)	3
iPad screen layout does not affect scoring	7
iPad screen size does not affect scoring	8
Scoring on an iPad would not affect scoring accuracy	10
Specific iPad-related difficulties/challenges for scoring essay	
Scrolling	3
Touch screen too sensitive	2
Time lapse on an iPad	1
iPad screen too small to display everything on one page	1
Hard-to-identify invisible scrolling bar	3
Extra scrolling may be needed to reach the scoring panel	1
Scoring on an iPad for 4–8 hours would affect the scoring quality and accuracy	2
Scoring on an iPad takes a longer time than scoring on a desktop computer	3
No difficulty	4
iPad-specific features useful for essay scoring	
Convenient to use	9
Easy to navigate	8
Easy to adjust font/picture size (enlarge or minimize)	9
Allows scoring with flexible sitting positions	8
Preference of scoring essays on a desktop/laptop computer	4/4 (before/after)
Preference of scoring essays on an iPad	4/2
No preference between iPad and desktop/laptop computer	2/4

Results Based on Post Hoc Survey/Interview

Table 5 summarizes the results derived from the survey and interviews that were carried out after experimental scoring. Overall, seven raters thought that the experimental scoring session was comfortable and similar to an operational scoring session, while two thought the experimental scoring session was more challenging than operational scoring. One thought that the experimental session was less challenging than a typical scoring day in operation.

All raters reported that the scoring environment was different from that of their daily scoring work, with various differences mentioned. The listed differences included the type of chair and desk, the type of mouse used, the size of the screen, the lighting, and other factors, such as background music, which one rater mentioned. However, seven raters reported that these differences did not affect their scoring performance; three mentioned that the intermittently displayed scrolling bar likely distracted them, especially at the beginning of the scoring session using the iPad.

No rater indicated that the font on the iPad screen affected their scoring negatively. Moreover, three raters reported that reading using the iPad seemed easier or more comfortable than using a laptop or desktop computer because it was much easier to adjust the font size on an iPad.

In reporting the layout of the scoring screen, six raters reported that the layout on the iPad was fine and did not affect their scoring; one indicated that she preferred the layout using the iPad. Three raters reported that the layout on the iPad required more scrolling than did desktop or laptop computers; the other seven raters reported that using the iPad resulted in comparable ($n = 5$) or even less scrolling ($n = 2$) than a laptop or desktop computer by their observations. One rater also reported being concerned about the layout on the iPad, which required extra scrolling to see the score-assignment panel.

Eight raters reported that the size of the iPad screen did not have any effect on their scoring (e.g., a quicker onset of tiredness or fatigue). One rater reported that she had to move her head closer to the screen for the reading and score assignment using the iPad than she did when using a desktop computer; another rater reported that although the iPad-based scoring was easier than using a laptop computer, desktop-based scoring seemed even easier and more convenient.

Only two raters thought that using an iPad to score for 4–8 hours, as in operational scoring, would require more breaks than if they were using a laptop or desktop computer, while the others thought that having more breaks would not be necessary when scoring using the iPad. Two raters also reported that they would require more breaks using a desktop computer than an iPad.

T hree raters indicated that they felt they spent more time on average reading and scoring an essay using the iPad than using the desktop computer, while the other raters thought there was no difference in the time spent per essay. No rater believed that using an iPad to score essays would affect their scoring accuracy in comparison to using a desktop or laptop computer. One rater even reported that she felt that using a desktop computer would lead to lower scoring accuracy than when using an iPad. She argued that the fixed screen position and use of mouse clicking on a desktop computer would likely result in more boredom and fatigue.

When asked to list one or two features related to the iPad that raters considered useful for the essay scoring task, the raters mostly indicated that the iPad was convenient to use, easy to navigate, easy to adjust (enlarge or minimize the font sizes), and easy to hold or to place in different positions. Three raters also pointed out the convenience associated with the touch screen.

Regarding specific difficulties associated with scoring using the iPad, four raters reported no difficulty, three reported issues associated with scrolling, two reported issues associated with the touch screen (that it was too sensitive and required some time to get used to), and one rater mentioned that he felt it took longer to respond when using the iPad than when using the desktop computer. Finally, one rater was concerned that the small iPad screen could not display everything on one screen in comparison to a typical desktop computer screen.

When asked which device (the desktop or the iPad, if it were allowed) was more comfortable for scoring essays before the scoring experiment, four raters endorsed the desktop computer, while another four endorsed the iPad and two considered there to be no difference between the two.

Similarly, after the scoring sessions, the four raters who endorsed the computer in scoring prior to the scoring session stayed with their earlier preference, reporting that they would prefer to use a desktop computer to score essays in the future. Among the four who endorsed the iPad prior to the scoring session, two raters stayed with their earlier preference to use the iPad for future scoring. The remaining four raters indicated that they had no preference between the two devices for future scoring work.

Summary and Discussion

With the fast-paced development and innovation in the areas of science and technology, new devices are likely to be continuously adopted in educational assessment and evaluation. While bringing in convenience and features that are not possible with traditional assessment (e.g., response time and adaptive testing), the adoption of new technologies is also likely to challenge traditional methods of evaluating measurement properties, including scoring essays and speaking responses.

For the first research question, we found that raters reported iPad-specific difficulties and challenges for essay scoring purposes, including the invisible scrolling bar, the extra scrolling needed to reach the score-assignment panel, difficulty navigating between the prompt and the essay, and oversensitivity of the touch screen. However, fewer than half of the raters thought that scoring using a desktop or laptop computer was more comfortable and preferable for scoring essays. More raters either preferred the iPad or did not have any preference for either the iPad or the desktop or laptop computer. With regard to the second and third questions, the results suggest that scoring on an iPad does not lead to an obvious increase in scoring time or scoring error over scoring on a desktop computer.

Even though the study included only a very limited sample of raters, the results are encouraging and show no obvious reduction in scoring quality due to the use of an iPad. Although only two raters had used an iPad regularly, no raters felt that using an iPad to score essays would affect scoring accuracy. These self-reported results corroborate the results based on scoring accuracy and scoring time.

The fact that there was no observable difference in scoring accuracy or scoring time per essay between the iPad and desktop computer suggests that using an iPad to score essays is not likely to affect operational scoring quality, even with some raters being unfamiliar with the iPad. Scoring quality as indicated by the agreement between the raters' scores in this experiment and the criterion scores was comparable to that reported in other studies based on operational data of graduate admission tests. For example, Bridgeman et al. (2009) reported that the exact agreement between raters was 72%

for *GRE*® test essays, which was very similar to the rate we found in the current study. However, we recommend a training session or even multiple training sessions for raters who have no experience or limited experience using an iPad so that raters will not be affected by the issues raised in this study or, at least, will be affected to a lesser degree.

The findings of this study suggest that for situations where reading, navigating, selecting, and tapping are required to complete a task (e.g., other tasks similar to rating an essay), using an iPad does not appear to be associated with an obvious disadvantage in terms of lower accuracy or longer time per essay on average. Such findings are encouraging for scoring of other types of essay tasks or even speaking tasks, as long as the scoring involves only navigation, such as scrolling, selecting, and tapping, for reading and/or listening responses on an iPad. However, further investigations are desirable with a larger sample over a longer period (e.g., 8 hours).

This study has several limitations. The raters were not randomly selected, and the number of raters was very small, which does not allow for a formal statistical test of the no-difference hypothesis between the two devices. Although the raters could be considered experienced raters of college-level essays, their backgrounds associated with using iPads or other types of devices appear to vary to some extent. Furthermore, the 2-hour scoring session was much shorter than an 8-hour regular shift in operational scoring. The 20 essay ratings per device were only a fraction of the number of essays a typical rater would score in a day, which may have limited the capacity to demonstrate the possible differences between the two types of devices used to score essays in an operational setting. Furthermore, the essays included in this study were relatively easier to score compared to those scored in an operational setting. That being said, these types of essays (each with a clear, unambiguous score at a particular level) are often used in training and certifying raters in practice to make predictions and inferences about raters' performance in an operational setting (e.g., McClellan, 2010). Also, a variety of tablets have functionality that differs from the iPad's. Caution needs to be used before generalizing the results to all types of tablets. Finally, the small sample size did not provide enough statistical power to conduct formal model-based analysis (e.g., the general linear model) to test the mode effects and the order effects on scoring accuracy and scoring time, which seems desirable in future replications. Moreover, as noted earlier, the current study's findings cannot address security-related concerns about iPad-based essay scoring, so further investigations are necessary.

Conclusion

Despite these limitations, the results support a tentative conclusion that scoring on an iPad does not present an obvious disadvantage over scoring on a desktop computer with regard to scoring accuracy and scoring time.

Note

[1] Mean square error was computed by using rated score minus criterion score, taking a square, and then taking an average across all ratings within a rater and a device.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (Research Memorandum No. RM-03-05). ETS.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*(3), 191–205. https://doi.org/10.1207/S15324818AME1603_2

Bridgeman, B., Trapani, C., & Attali, Y. (2009, April). *Considering fairness and validity in evaluating automated scoring* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Diego, CA, United States.

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper and computer-based assessment in a K–12 setting* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, United States.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, *15*(3), 243–263. https://doi.org/10.1080/13803610902972940

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–516). Praeger.

Falvey, P., & Coniam, D. (2010). A qualitative study of the response of raters towards onscreen and paper-based marking. *Melbourne Papers in Language Testing*, *15*(1), 1–26.

Johnson, E. G. (1993). *The results of the NAEP 1993 field test for the 1994 National Assessment of Educational Progress*. ETS.

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education*, *22*(1), 22–37. https://doi.org/10.1080/08957340802558326

Ling, G. (2015, April). *Does it matter if one takes a test using an iPad or a desktop computer?* [paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL, United States.

McClellan, C. A. (2010, February). Constructed-response scoring—Doing it right. *R&D Connections*, *13*. https://origin-www.ets.org/Media/Research/pdf/RD_Connections13.pdf

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment*, *2*(6), 1–44.

Powers, D., & Farnum, M. (1997). *Effects of mode of presentation on essay scores* (Research Memorandum No. RM-97-08). ETS.

Powers, D., Farnum, M., Grant, M., & Kubota, M. (1997). *A pilot test of online essay scoring* (Research Memorandum No. RM-97-07). ETS.

Raikes, N., Greatorex, J., & Shaw, S. (2004, June). *From paper to screen: Some issues on the way* [Paper presentation]. International Association of Educational Assessment Conference, Manchester, England.

Richardson, J., Dillon, A., & McKnight, C. (1989). The effect of window size on reading and manipulating electronic text. In E. Megaw (Ed.), *Contemporary ergonomics* (pp. 474–479). Taylor and Francis.

Sanchez, C. A., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity and comprehending complex text. *Human Factors*, *51*(5), 730–738. https://doi.org/10.1177/0018720809352788

Singh, R. I., Sumeeth, M., & Miller, J. (2011). Evaluating the readability of privacy policies in mobile environments. *International Journal of Mobile Human Computer Interaction*, *3*(1), 55–78. https://doi.org/10.4018/jmhci.2011010104

Zhang, Y. L., Powers, D. E., Wright, W., & Morgan, R. (2003). *Applying the online scoring network (OSN) to* Advanced Placement® *program* (AP®) *tests* (Research Report No. RR-03-12). ETS. https://doi.org/10.1002/j.2333-8504.2003.tb01904.x

Zickuhr, K., Rainie, L., Purcell, K., Madden, M., & Brenner, J. (2012). *Younger Americans' reading and library habits*. Pew Research Center. http://libraries.pewinternet.org/2012/10/23/younger-americans-reading-and-library-habits/

Appendix A

Background Questionnaire

1. GENDER: □ Male □ Female

2. ETHNICITY: (check all that apply)

□ American Indian or Alaskan Native
□ Asian, Asian American, or Pacific Islander
□ Black or African American
□ Mexican, Mexican American, or Chicano
□ Puerto Rican
□ Other Hispanic or Latin American
□ White (Non-Hispanic)
□ Other ethnicity not listed above_______________________________

3. What is your highest level of education?

□ Completed some high school
□ GED
□ High school
□ Completed some college
□ Associate's degree
□ Bachelor's degree
□ Bachelor's plus credits
□ Master's degree
□ Master's plus credits
□ Doctoral degree
□ Other____

4. Please describe your essay scoring experience (GRE essay related versus other tests).
5. Which type of device do you normally use for your daily work?

   □ Desktop □ Laptop □ iPad

6. Which type of device do you normally use for scoring essays?

   □ Desktop □ Laptop □ iPad

7. Which type of device do you prefer to use to score essays?

   □ Desktop □ Laptop □ iPad

8. How frequently do you use an iPad?

   □ Frequently □ Sometimes □ Rarely □ Never

Appendix B

List of Behaviors to Observe

1. Does the rater change any display/reading preference once logged in or moved to the scoring page?
   a. Font size larger, smaller
   b. Full-screen display or not
   c. Does this happen in both devices?
   d. Are the choices the same between the two devices? (Ask this question if the display or reading preferences were different between the two devices, and clarify the answers.)

2. How does the rater move around the screen?

   iPad: Desktop:

3. How does the rater employ scrolling?

   iPad: Desktop:

4. Does the rater make any unfamiliarity-related movement, gestures, and verbal or facial expressions? If so, describe them

   iPad: Desktop:

5. What questions, if any, does the rater ask during the scoring process?

   iPad: Desktop:

6. How does the rater position the iPad when scoring (e.g., holding it, placing it on the table)?

   iPad:

7. What types of navigating and typing behaviors does the rater exhibit?

   iPad: Desktop:

8. Note any other observations here

   iPad: Desktop:

Appendix C

Interview Questions

1. Are the settings in today's scoring session different from those in your home office?
   If they are, please give some examples of how they are different.
   Did the settings in today's session affect your scoring today? If so, how?

2. Overall, how did today's scoring using the iPad/desktop computer compare to an operational scoring session?

A. Today's scoring session was comfortable and similar to an operational scoring session.
B. Today's scoring session was more challenging than an operational scoring session.
C. Today's scoring session was less challenging than an operational scoring session.

Device-Specific Challenges or Difficulties for Scoring

3. Did the font on the iPad screen affect your reading and assignment of score levels? If so, how?
4. Did the layout on the iPad screen affect your reading and assignment of score levels? If so, how?
5. Did the size of the iPad screen have any effect on your energy level (e.g., increased fatigue) while reading the essays? If so, please describe.
6. Did you find that you had to scroll the screen more often using one device over the other? If so, which one?
7. Did you feel that you would need more breaks during the scoring session using one device over the other? If so, which one?
8. On average, did you spend a longer time scoring an essay using one device over the other? If so, which one?
9. Do you think using one device over the other would result in a greater scoring accuracy? If so, which one?
10. Please describe any difficulty or challenges that you experienced specifically related to the use of the iPad during today's scoring.
11. Which do you feel more comfortable using to score essays: the desktop computer or the iPad?
12. Based on your experience today, do you prefer using the iPad or the desktop computer to score essays?
13. Please list one or two of the most useful features/functions of the iPad for essay scoring.
14. Please list one or two features/functions of the iPad that were problematic or not useful for the essay scoring.

Suggested citation:

Ling, G., Williams, J., O'Brien, S., & Cavalie, C. F. (2022). *Scoring essays on an iPad versus a desktop computer: An exploratory study* (Research Report No. RR-22-08). ETS. https://doi.org/10.1002/ets2.12349

Action Editor: Rebecca Zwick and Sooyeon Kim

Reviewers: Brent Bridgeman and Alexis Lopez

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/