The difference in performance between computer and paper administered tests in a stressful environment

Corey Macrander, Regina Manansala, Stephanie Rawson, and JeeYoung Han

Correspondence

Corey Macrander

1001 W. Dayton Street Apt 3E., Madison WI 53715, USA
E-mail:macrander@wisc.edu

Short title: Computer versus paper test performance in stressful environment

Key Points

computer administered testing, paper administered testing, stressful environment, test performance

Abstract

With advancements in technology, many tests such as graduate school entrance examinations are being administered online,while some remain as a paper based examination. Many studies have aimed to determine whether the difference between the formats of the test administered have an effect on students' performance. Many researchers have explored different variables that may affect this performance difference, but not many have studied the effects stress also has on performance. The aim of this study is to determine whether stress plays a significant role in performance on computer and paper administered tests. To do so, a paper and online version of the Graduate Record Examinations' (GRE) verbal reasoning section were administered to each participant while simultaneously being asked to squeeze a dynamometer every 30 seconds for ten minutes. A galvanic skin response receptor, a pulse oximeter, and a dynamometer were used to observe the possibility of increased stress levels. While the stress induced environment may not have been as stressful as expected, the data gathered suggests that there is no difference in performance between computer and paper tests. Future tests should aim to create a more stressful environment.

Introduction

With advancements in technology, an increasing amount of tests are being administered online such as state drivers' license exams, military training exams and entrance exams in post-secondary education. (Russo, 2002; Trotter, 2001) In addition, students are being asked to complete more of their coursework online. This is a problem for many students that are easily distracted from their work by the Internet, as it allows for one-click access to more entertaining material than their homework. Computer programs have begun to surface that allow students to monitor how they are spending their time on the Internet, even allowing them to block themselves from websites after a certain amount of time. However, are there other factors underlying this issue with working online? Is it possible that working on a computer itself can inhibit productivity and performance?

Many researchers have questioned if the form of a test administered, whether it be taking an exam on paper or on a computer, affects students' performance. A study by Bunderson et al in 1989 found that out of 23 educational studies, 11 studies showed no difference in performance, 9 favored paper-based tests, and 3 favored computer-based tests. In researching these questions, there are many variables that need to be taken into account, such as computer and content familiarity. Wallace and Clariana (2000) observed that those more familiar with the content and computers, performed better than those who were not as familiar. This allowed them to conclude that both content and computer familiarity were both important factors to consider when studying the differences between computer based and paper based tests. Academic attainment is another variable needed to take into consideration when performing this study because those with higher levels of academic attainment often have a higher exposure to computers than those with lower levels academic attainment. Studies have shown that this factor is indeed considered important when those with higher levels of academic attainment performed better on computers than those with lower levels of academic attainment (Watson 2001; Wallace and Clariana 2002). Gender and ethnicity were also other factors to consider in this study. A study completed on first year medical students at the Illinois College of Medicine at Urbana-Champaign found no gender differences in test administered on-line (Kies et al. 2006). Ethnicity was also found to not be a significant factor when computer familiarity was taken into account for tests administered by computer. However, in their research, several of the experiments did not measure

performance under increased stress levels. We believe that this could be a possible variable when considering the differences between tests administered by computer and by paper. We would like to explore a possible correlation between online coursework and increased stress levels, since it has yet to be observed. **We hypothesize that under stress induced environments, students will have more distractions and higher stress levels while working on a computer. Therefore, we also hypothesize that students will perform better on paper than on the computer.**

We base this hypothesis off of recent research conducted on first year medical students at the University of Heidelberg. These students were given an option between taking a test off a computer or off paper. Nearly 63% of the students choose to take the paper based examination. Of the reasons listed for picking the paper based test, nearly 42% listed that the paper based test allowed them to take notes on the test, which the computer based test does not allow you to do. Also, 11% of the students listed fear of P-C errors or technical difficulties and the loudness of the keyboards as a distracting factor for reasons for preferring the paper based test. This is supporting evidence for our hypothesis because these students would find a mandatory computer based test to be more stressful and would possibly do worse than if they took the test on paper (Hochlehnert et al. 2011). Also, Mourant, Lakshmanan, and Chantadisai (1981) showed that students became more fatigued when reading from a computer screen when compared to the same text on paper.

Other evidence supporting a more stressful situation in computers comes from the ease of flipping back and forth between questions and text. In a paper based test a student can easily do this, but on a computer based test this is not so trivial (Wallace and Clariana 2002). Haas and Hayes (1986) found that this factor lead to computer based scores to be lower than paper based scores when the text passage required more than one page. Other research conducted concluded that computer based tests may require a higher level of focus for the students, which could add to heightened stress levels (Clariana, 1997; Clariana and Smith 1988). These are all factors which could cause a heightened stress level in students taking a computer based exam off and help lead us to our hypothesis.

To test our hypothesis, students will be administered both an online and paper standardized test on reading comprehension, each lasting ten minutes. Participants will be interrupted periodically during the test and asked to squeeze a dynamometer as hard as they can for 5 seconds. We plan to use the dynamometer as a mental stressor that distracts students, as well as a tool to measure fatigue levels. By including this stressor, we hope to amplify a possible correlation between stress levels and form of test in a shorter period of time. Throughout the ten minutes of testing, we will use a pulse oximeter to measure heart rate to aid in determining changes in stress levels. A galvanic skin response sensor (GSR) will also be used to measure the subjects' sweat levels. Finally, the test scores of the students will determine their performance levels under induced stress conditions.

Materials and Method

Before each experiment, subjects are given and asked to sign a consent form that states the background of the study and its procedures. Students are given two versions of a practice Graduate Record Examinations (GRE) verbal test - one typed into a survey program to be administered on the computer and another printed to be administered on paper. While the students are taking the test, they will be hooked up to a galvanic skin response sensor and a dynamometer which are connected to a computer via a Biopac System and a pulse oximeter. The galvanic skin response sensor (GSR) is wrapped around the pinky and ring fingers of the subjects' dominant hand. This records his or her sweat levels throughout the entirety of the experiment. The pulse oximeter is clipped on to the middle finger of the subjects' dominant hand allowing us to record the changes in blood pressure and heart rate. Electrodes are attached to the subjects' non-dominant hand which they use to squeeze a handheld

dynamometer every thirty seconds throughout the ten minute test. The dynamometer and GSR were calibrated before testing began and each subject was given a couple minutes of rest after calibration.

The order of the tests administered is varied between subjects and the order the test is taken is dependent upon the subjects' preference. Each subject is given ten minutes to complete the computer and printed version of the GRE tests. During the ten minutes, a member of the group instructs the student to squeeze the dynamometer for five seconds every thirty seconds while another member of the group records the subjects' blood pressure every fifteen seconds. Subjects are notified when there are five minutes, two minutes and thirty seconds left in the test. After completing the first test administered, subjects are given a five minute break before beginning the second administered test. The tests are later compiled and scored.

Per request of our reviewer, an extra stressor was added during each test. Every thirty seconds beginning from the fifteen second mark, the subject is given three cards and asked to multiply them together. Face cards had a value of 10, aces had a value of 11 and the rest of the cards were used as is. The subjects' are unaware that their answers have no effect on their the study but is used just to create a more stressful testing environment.

Results

We collected data for n=15 subjects, four of which (n>11) underwent the second trial of tests that included the additional stressor. Variability amongst subjects was kept at a minimum by using all subjects from the same Physiology 435 course, implying similar age group, knowledge-base, and education. Both males and females are represented almost equally (males = 8, females = 7). A one-way ANOVA was done on measured force, test scores, mean arterial blood pressure and number of stress spikes during testing for computer versus paper.

Test Scores

We saw no statistical difference (for n=15, p-value =0.47307) between computer and paper test scores. Eight subjects scored better on the computer test, and seven subjects scored better on the paper test.

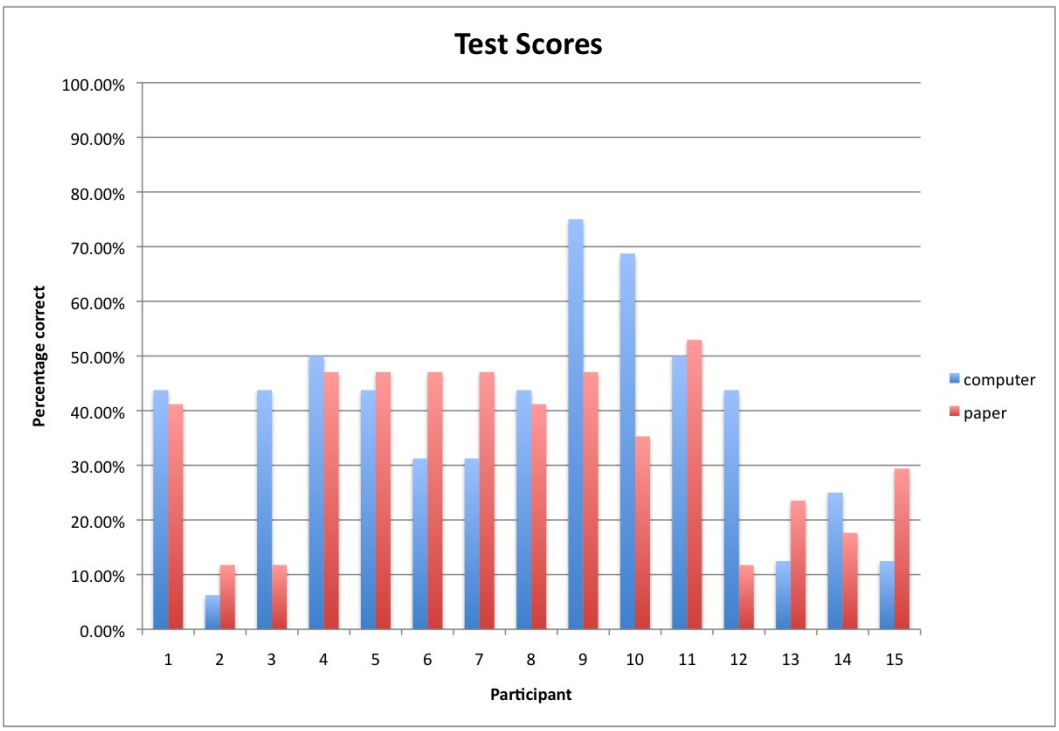

Figure 1. Percentage test scores for each participant, for both computer and paper administered tests.(p-value =0.47307)

Blood Pressure

We saw no statistical difference between average blood pressures during computer and paper administered tests for each individual (p = 0.8420).

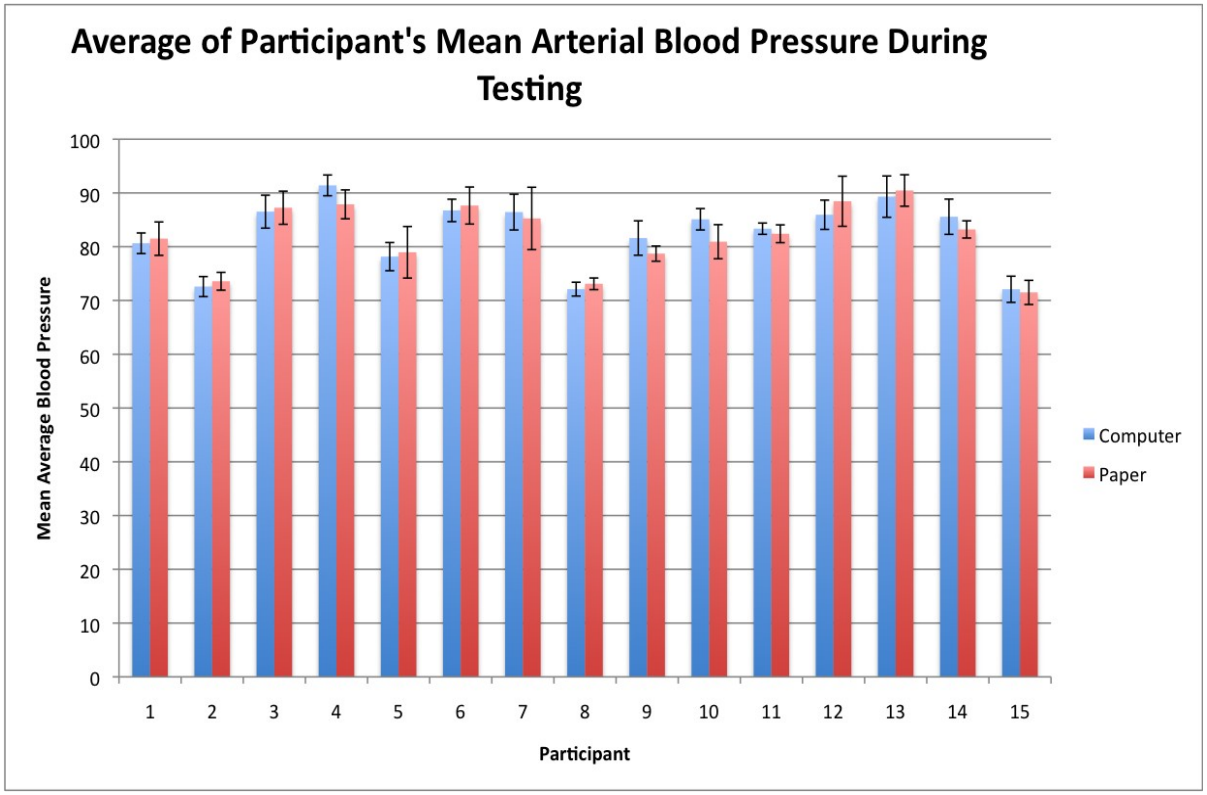

Figure 2. Average of mean arterial blood pressure during testing for each participant, both computer and paper administered test. p-value=0.84201, error bars of one standard deviation.

Sweat Responses

During a persons resting period, a mean value of sweat conductance was taken from any visual spikes in the GSR graph. This was established as the threshold conductance. Any spikes greater than this value during the examination were considered to be stress induced from the actual examination. The number of spikes for each subject during the two different test periods were counted and are listed in Figure 3.

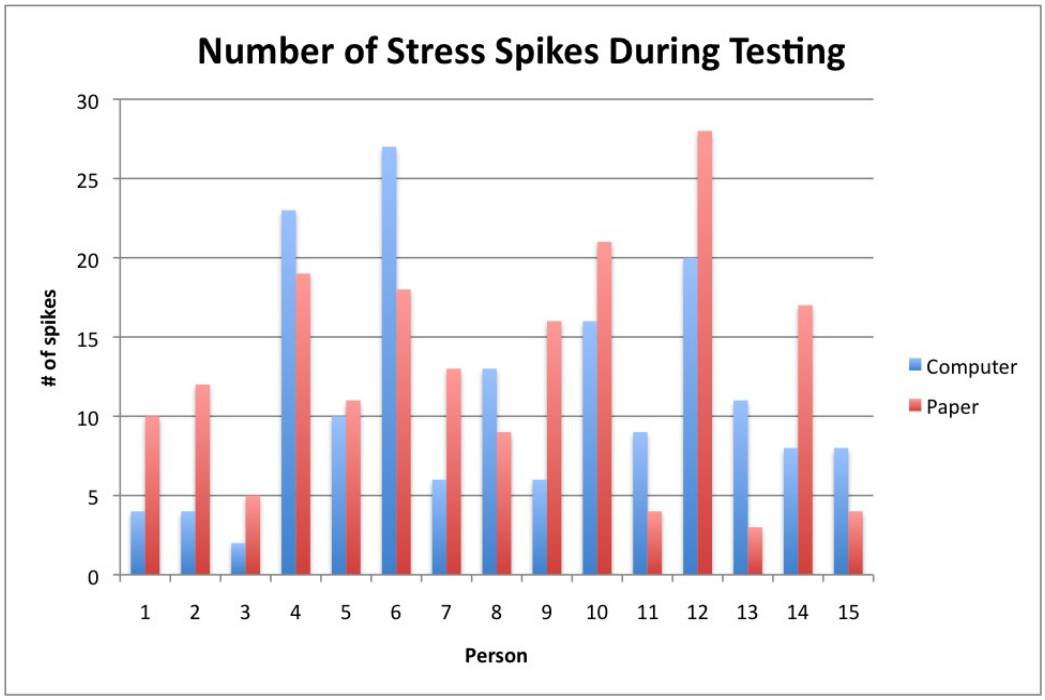

Figure 3. Number of stress spikes counted above the calculated threshold. p-value = 0.56959.

Stress and Fatigue Levels

The participants were asked to squeeze the dynamometer for five seconds, every thirty seconds, for ten minutes. The average force from each five second squeeze was calculated, and the average of all average forces for each subject are graphed (Figure 4) Eight out of the fifteen participants generated greater force during paper administered exams (n=2,3,4,5,8,11,13,15).

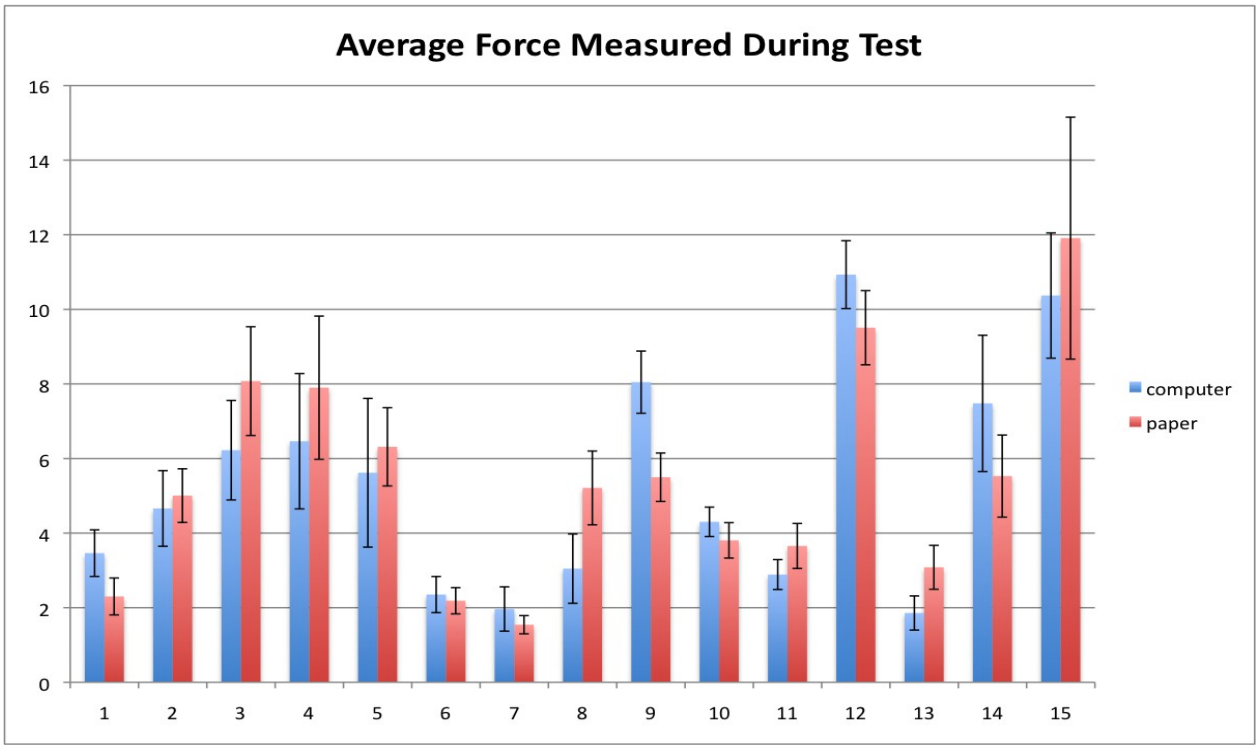

Figure 4. Overall average forces measured for each subject during both tests. p-value =0.90864, error bars of one standard deviation.

Compiled Results

We compiled all our results, including which test was administered first. This was done to see if the type of test administered first was the sole determinant on differences between test performance. This also allowed us to compare how other variables affected each other.

Participant	Paper or Computer First	Test Score Higher on:	Sweat More on:	Blood Pressure?	Force greater on:
1	Paper	Computer	Paper	Paper	Computer
2	Paper	Paper	Paper	Paper	Paper
3	Paper	Computer	Paper	Paper	Paper
4	Computer	Computer	Computer	Computer	Paper
5	Computer	Paper	Paper	Paper	Paper
6	Computer	Paper	Computer	Paper	Computer
7	Computer	Paper	Paper	Computer	Computer
8	Computer	Computer	Computer	Paper	Paper
9	Computer	Computer	Paper	Computer	Computer
10	Paper	Computer	Paper	Computer	Computer
11	Paper	Paper	Computer	Computer	Paper
12	Paper	Computer	Paper	Paper	Computer
13	Computer	Paper	Computer	Paper	Paper
14	Computer	Computer	Paper	Computer	Computer
15	Computer	Paper	Computer	Computer	Paper

Figure 5. Compilation of all results.

Discussion

It was initially hypothesized that the participants would do better on paper administered tests because it poses less stressors than computers. However, based on our results, the type of test administered does not affect participants' performance. Eight out of the fifteen subjects performed better on the computer administered test, with no statistical difference observed. In part, this may have been due to the limitations of our experiment set-up. The same exact test could not be given for the computer and paper tests because subjects would then be given the same article to read twice, therefore eliminating the assessment of the difference in performance on each format. In addition, tests with similar content were chosen pertaining to what the target participants would find stimulating; however,

not everyone's preferences could be accommodated. Because of this, participants scores may have merely been a reflection of which article they preferred reading, not of which test format they preferred.

Another concern in the experiment was a possible correlation between test performance and order of test administration. Due to unfamiliarity with the procedures and content of our tests, the first test administered was anticipated to generate lower test scores and higher stress levels. However, this was not observed, showing that this correlation may be insignificant.

There was also no significant difference in blood pressure during computer and paper tests. This does not support the hypothesis which inferred blood pressure would be higher during which test produced more stress. This could be because there was a failure to create an environment stressful enough to produce differences in blood pressure. To improve on this, the final four participants (12-15) were given the additional stressor of having to multiply three playing cards dealt to them every thirty seconds. Participants were visibly more stressed because of this, and all four participants even verbally commented to us on how stressful the test was. Still, this did not change the significance of the blood pressure. Because of this, we believe that blood pressure was not a good indicator of stress levels.

There was also no statistical difference between the number of stress spikes between computer and paper administered tests. However, it should be noted that ten out of the fifteen participants did worse on the exam that they had more stress spikes measured, thus suggesting that the number of stress spikes did influence their performance.

The dynamometer was also intended as a stressor and as a way to measure fatigue levels; we reasoned that by monitoring grip strength, this would provide a reasonable measurement of particpant energy level during the course of the examination. Throughout each administered test, the force of each squeeze was expected to decline as an indication of fatigue; we also envisioned this physical interruption as an additional stressor during the reading exam. However, we observed that the force remained relatively constant, indicating that the subjects might not have been as fatigued or frustrated as intended. In addition, we observed that each individual did not necessarily generate greater force during the first test administered. One observation that is interesting to note is that eleven out of the fifteen participants generated less force on the dynanometer during the exam that they had more frequent stress spikes. Further testing is required to fully explore this possibility.

Conclusion

Overall, the experiment was designed to test whether stressful conditions affected stress levels during computer and paper based tests, thus ultimately affecting performance based on the type of administered test. A stressful environment was created by administering a challenging test that was too long for the time allotted and by constantly interrupting the participants by squeezing the dynamometer. The test administered, however, had no significance to the participants and therefore could not encapsulate the stressful environment normally found within an actual test. This may be a major flaw in the experimental design and, in further tests, an attempt to make the test significant to the participant would be made. Without actually creating a stressful enough environment, a difference between performance on tests administered through a computer or paper was not observed. An additional stressor of having to multiply three playing cards together was also added to the second trial to further increase the stress of the environment. This stressor did garner a greater stress response from subjects, evidenced by the fact that each subject verbally expressed their stress to us during the test. Further investigation on this topic would include the playing card stressor, as it clearly provided a more stressful environment for the subjects.

Unfortunately, we were unable to collect further information using the playing card method due to time constraints, lack of participants, and shortage of group members to administer the trial. This,

however, does not mean that there is no correlation between performances. A larger group of participants would need to be tested in order to further investigate this correlation.

Acknowledgements

This experiment was completed with the help of Dr. Andrew Lokuta, the Physiology faculty and the rest of the Physiology 435 staff at the University of Wisconsin-Madison

References

Bunderson C V, Inouye D K and Olsen J B (1989) The four generations of computerized educational measurement in R L Linn (ed) *Educational measurement* American Council on Education, Washington DC, 367-407.

Clariana R B (1997) Considering learning style in computer-assisted learning *British Journal of Educational Technology* **28** (1) 66-68

Clariana R B and Smith L J (1988) *Learning style shifts in computer-assisted instruction* Presented at the annual meeting of the International Association for Computer In Educucation (IACE), New Orleans, LA, April 1988 (ERIC Document Reproduction Service: ED 295 796)
Haas C and Hayes J R (1986) What did I just say? Reading problems in writing with the machine *Research in the Teaching of English* **20** (1) 22-35

Hochlehnert A, Konstantin B, Moeltner A, Juenger J (2011) Does Medical Students' Preference of Test Format (Computer-based vs. Paper-based) have an Influence on Performance? *BMC Medical Education* **11**: 89

Kies S M, Williams B D, Freund G C (2006) Gender plays no role in student ability to perform on computer-based examinations *BMC Medical Education* **6**: 57

Mourant R R, Lakshmanan R, and Chantadisai R (1981) Visual fatigue and cathode ray tube display terminals *Human Factors* **23** (5) 529-540.

Russo A (2002) Mixing technology and testing *The School Administrator* (online), 2002_04. Available at: http://www.aasa.org/publications/sa/2002_04/russo.htm.

Trotter A (2001) Testing firms see future market in online assessment *Education Week on the Web* 20 (4) 6.

Wallace P E and Clariana R B (2000) Achievement predictors for a computer-applications Module delivered via the world-wide web *Journal of Information Systems Education* **11** 13-18. http://gise.org/JISE/Vol11/v11n1-2p13-18.pdf.

Wallace P E and Clariana R B (2002) Paper-based versus computer-based assessment: key Factors associated with the test mode effect *British Journal of Educational Technology* **33** 593-602

Watson B (2001) Key factors affecting conceptual gains from CAL *British Journal of Educational Technology* **32** 587-593.